PHOTOGRAPH THIS SHEET

AD-A111 477

DTIC ACCESSION NUMBER

II

LEVEL

1

INVENTORY

SCIENCE, TECHNOLOGY, AND THE MODERN NAVY THIRTIETH ANNIVERSARY 1946-1976 ONR-37

DOCUMENT IDENTIFICATION

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

DISTRIBUTION STATEMENT

ACCESSION FOR	
NTIS	GRA&I
DTIC	TAB
UNANNOUNCED	
JUSTIFICATION	
BY	
DISTRIBUTION /	
AVAILABILITY CODES	
DIST	AVAIL AND/OR SPECIAL
A	

DISTRIBUTION STAMP

DTIC
ELECTE
FEB 26 1982
B

DATE ACCESSIONED

DTIC COPY INSPECTED 2

82 02 11 001

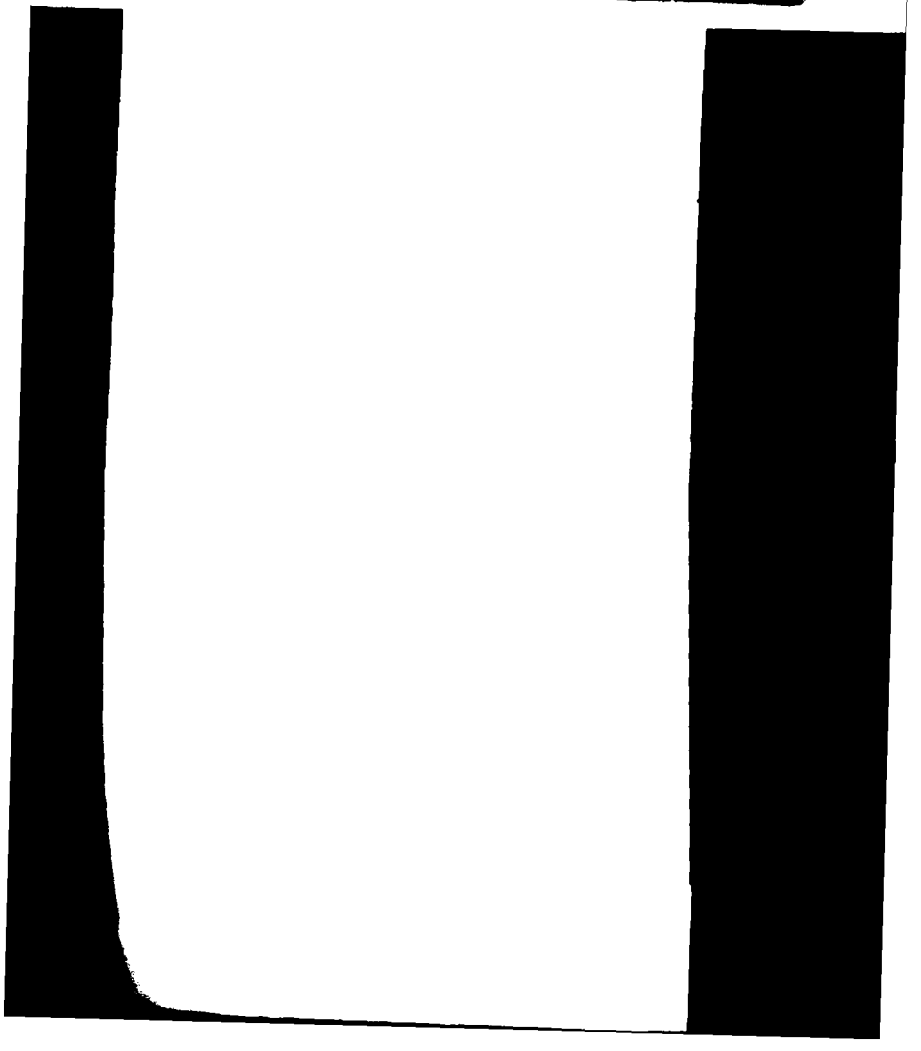
DATE RECEIVED IN DTIC

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-DDA-2

DTIC FORM OCT 79 70A

DOCUMENT PROCESSING SHEET

AD A111477



ONR-37

SCIENCE, TECHNOLOGY, AND THE MODERN NAVY

Thirtieth Anniversary

1946 — 1976

Edward I. Salkovitz, Editor

1976

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

Department of the Navy
OFFICE OF NAVAL RESEARCH
Arlington, Virginia

CONTENTS

Page

Foreword vii
Rear Admiral R. K. Geiger, USN, Chief of Naval Research

Preface ix
Edward I. Salkovitz, Office of Naval Research

Introductory Remarks xi
The Honorable Melvin Price, Chairman, House Armed Services Committee

PHYSICAL SCIENCES

SOLAR-TERRESTRIAL PHYSICS 2
H. Friedman, Naval Research Laboratory

ATOMIC AND MOLECULAR STANDARDS OF TIME AND FREQUENCY 18
N.F. Ramsey, Harvard University

DEVELOPMENT OF SURFACE ACOUSTIC WAVE DEVICES 40
G. S. Kino and H. J. Shaw, Stanford University

LASERS 65
A. L. Schawlow, Stanford University

MATHEMATICAL AND INFORMATION SCIENCES

LINEAR PROGRAMING, PAST AND FUTURE 84
G. B. Dantzig, Stanford University

NEXT DECADE OF LOGISTICS RESEARCH 96
H. M. Wagner, University of North Carolina and McKinsey and Co.

AUTOMATION AND ARTIFICIAL INTELLIGENCE 110
M. Minsky, Massachusetts Institute of Technology

APPLIED STATISTICS 120
H. Solomon, Stanford University

PERSPECTIVES IN MODERN CONTROL THEORY 142
M. Athans, Massachusetts Institute of Technology

HYDROMECHANICS RESEARCH AND THE NAVY: A PROJECTION 155
W. E. Cummins, David W. Taylor Naval Ship Research and Development Center

BIOLOGICAL AND MEDICAL SCIENCES

CHEMICAL FACTORS IN THE BRAIN INVOLVED IN LIFE-SUSTAINING REGULATORY MECHANISMS 172
R. D. Myers, Purdue University

ELECTRICAL "WINDOWS" ON THE MIND: APPLICATIONS FOR NEUROPHYSIOLOGICALLY DEFINED INDIVIDUAL DIFFERENCES 188
E. Callaway, M.D., Langley Porter Neuropsychiatric Institute

CONTENTS

PSYCHOLOGICAL SCIENCES

HUMAN CONSIDERATIONS IN INTERACTIVE TELECOMMUNICATIONS 198
A. Chapanis, Johns Hopkins University, and E. Williams, University College London

TOWARDS UNDERSTANDING AND IMPROVING DECISIONS 220
P. Slovic, Oregon Research Institute

ORGANIZATIONAL CLIMATE AS A MEDIATOR OF ORGANIZATIONAL PERFORMANCE 241
S. B. Sells, Texas Christian University

EARTH SCIENCES

RADIO WAVE PROPAGATION IN THE SOLAR-TERRESTRIAL ENVIRONMENT: PERSPECTIVES FOR THE FUTURE 264
O. G. Villard, Stanford University

ARCTIC SCIENCE: CURRENT KNOWLEDGE AND FUTURE THRUSTS 277
N. Untersteiner, University of Washington; K. L. Hunkins, Columbia University; and B. M. Buck, Polar Research Laboratories

REMOTE SENSING OF ENVIRONMENT: ACHIEVEMENTS AND PROGNOSIS 298
O. K. Huh, Louisiana State University; and V. E. Noble, Naval Research Laboratory

SOLID EARTH PROPERTIES AND THEIR IMPORTANCE TO THE NAVY CURRENT KNOWLEDGE AND FUTURE PROSPECTS 324
J. G. Heacock, Office of Naval Research; J. E. Oliver, Cornell University; G. V. Keller, Colorado School of Mines; and G. Simmons, Massachusetts Institute of Technology

COASTAL SCIENCES: RECENT ADVANCES AND FUTURE OUTLOOK 346
J. M. Coleman and S. P. Murray, Louisiana State University

SUN-EARTH RELATIONSHIPS AND THE EXTENDED FORECAST PROBLEM 371
W. O. Roberts, Aspen Institute for Humanistic Studies, University of Colorado, and National Center for Atmospheric Research

MATERIALS SCIENCES

CERAMICS IN THE FUTURE 388
J. B. Wachtman, Jr., National Bureau of Standards, and J. R. Johnson, 3M Company

PROSPECTIVES FOR SURFACE CHEMISTRY 415
J. T. Yates, Jr., and T. E. Madey, National Bureau of Standards

FUTURE OF AIRBREATHING PROPULSION 433
S. N. B. Murthy, Purdue University

FUTURE DESIGN AND ANALYSIS OF NAVAL STRUCTURES: THE IMPACT OF COMPUTING TECHNOLOGY 463
J. L. Tocher, Boeing Computer Services, Inc.

OCEAN SCIENCE AND TECHNOLOGY

NUMERICAL MODELING AND GLOBAL OCEAN FORECASTING 480
A. R. Robinson, Harvard University

MONITORING THE OCEAN ACOUSTICALLY 497
H. W. Munk and P. Worcester, Scripps Institution of Oceanography, University of California, San Diego

IMPROVING THE CHEMICAL BEHAVIOR OF METALS IN THE OCEAN ENVIRONMENT 509
D. R. Kester, University of Rhode Island

MARINE BIODETERIORATION 517
J. D. Costlow, Jr., Duke University Marine Laboratory

CONTENTS

TECHNOLOGY

UNCONVENTIONAL VEHICLES FOR OCEAN RESEARCH 528
F. N. Spiess, Scripps Institution of Oceanography, University of California, San Diego

NONLINEAR ACOUSTICS: A NEW DIMENSION IN UNDERWATER SOUND 547
T. G. Muir, Applied Research Laboratories, University of Texas at Austin

FOREWORD

Rear Admiral R. K. Geiger, USN

Chief of Naval Research



Admiral Robert E. Geiger is Chief of Naval Research and also serves as Assistant Oceanographer of the Navy for Ocean Science and as adviser to the Secretary of the Navy and the Chief of Naval Operations on research and patent matters. Early in his career he served aboard the U.S.S. *Bairoko* (CVE-115), as project officer and project pilot for aircraft development work at Key West, Fla., and as ASW Officer based at Barbers Point, Hawaii. He served as aeronautical engineer on the REGULUS II Program and on ASW research and development at the Bureau of Aeronautics; as Project Manager of the A-NEW Project at the Bureau of Naval Weapons; as Deputy Director for Advanced Plans of the Air Force Directorate for Special Projects at El Segundo, Calif.; as Deputy Director for Programs, Office of Space Systems, in the Office of the Secretary of the Air Force; as Project Manager (PM-16) of the then newly chartered Navy Space Project of the Naval Material Command; as Project Manager (PME-106) of the Naval Electronic Systems Command; and as Director of the Space and Command Support Division (OP-986) of the Office of the Chief of Naval Operations. He has received numerous awards, medals, and citations for "exemplary managerial skills and technical competence in the development of programs vital to the Nation." Admiral Geiger was born in St. Joseph, Mo. He attended the Georgia Institute of Technology, graduated from the U.S. Naval Academy, graduated as a Naval Aviator at Pensacola, Fla., and received a B.S. in Ordnance Engineering at the Naval Postgraduate School, Monterey, Calif., and an M.S. in Aeronautical Engineering at the Massachusetts Institute of Technology.

Thirty years ago the Office of Naval Research was established by Congress with its charter to encourage scientific research and to disseminate its findings in the Naval interest. An essential part of the process of scientific research is to assess where we are and to seek directions for further penetration of the unknown. In pursuit of its charter, and on the occasion of its thirtieth anniversary, ONR is pleased to have participated in this process by having sponsored the efforts included in this volume to gain new prospectives in scientific fields of Naval interest. I believe the articles will prove to be landmarks in their respective areas. We appreciate the great difficulty of the kind of work represented here and its great value, for science and the Navy, in charting the way ahead. From this new vantage point, ONR looks forward to being involved with you in gaining better understanding of nature and of what can be done.

PREFACE



Edward I. Salkovitz has been Director of the Material Sciences Division of ONR since 1973. Dr. Salkovitz was Chief Scientist at ONR London (1970–72) and Head of the ONR Metallurgy Branch (1960–64). At the Naval Research Laboratory (1942–60), he organized the Metal Physics Branch and served as Acting Associate Superintendent of the Metallurgy Division. He served as Head of the Material Sciences Division of the Defense Advanced Research Projects Agency (1964–5). At the University of Pittsburgh, Dr. Salkovitz held joint professorial appointments in the Physics Department and the Metallurgical and Materials Engineering Department (1965–70) (1972–3) and served as Chairman of the latter department. Currently, he is Adjunct Professor in the School of Engineering, and has been a part-time lecturer at Howard University and the University of Maryland. In 1959, Dr. Salkovitz received the U.S. Navy Meritorious Civilian Service Award and in 1963 was Guest Fellow at Harvard. He has written 80 papers, primarily in metal physics, and was coeditor of the book *Dimensions of Biomedical Engineering*. He is on the editorial advisory board of the *Journal of Biomedical Materials Research* and *Treatise on Material Science and Technology*. He earned a B.S. degree and D.Sc. in physics at Carnegie Institute of Technology.

When it was established in 1946, the Office of Naval Research was the main channel for Federal support of science in the United States. With the creation of the National Science Foundation, which was founded on the ONR model, as well as the follow-on establishment of research contracting offices in other agencies, ONR restricted its primary mission to satisfying the needs of the Navy. Since there are few fields of science or technology that cannot be related directly or indirectly to Navy requirements, the real choice becomes one of emphasizing areas of particular interest where anticipated results may have a direct bearing on future naval activities.

Most research programs within ONR are organized along disciplinary lines, the main disciplines being the physical, mathematical, information, biological, medical, psychological, earth, material, and ocean sciences; but some programs center on such fields as aviation, vehicle, and sensor technologies.

PREFACE

The Physical Sciences Program pursues research on radiation, lasers, acoustics, optics, electronics, superconductivity, magnetism, and surfaces. Research in the Mathematical Sciences Program covers the mathematical and computer sciences, the design of techniques for logistics and systems analysis, and the mechanics of fluids. The objectives of Biomedical research are to understand principles essential to maintaining the health and work capacity of personnel, to prevent disease, and to reduce stress factors such as pressure in diving. The Psychological Research Program seeks a better basis for understanding, improving, and predicting human performance in military environments. Thus, the reduction of manpower costs and the betterment of personnel effectiveness are anticipated benefits from investments in man-job and man-machine designs. The Earth Sciences Program has the objective of providing comprehensive knowledge of physical environments in which the Navy and Marine Corps must operate. Approaches are devised to measure, predict, and modify such environments in order to facilitate naval communications and operations. The Material Sciences Program conducts research in metallurgy, ceramics, chemistry, structural mechanics, and power. Progress in these disciplines is crucial to Navy concerns with the design, construction, and operation of its vehicles and weapons. The Ocean Science and Technology Program seeks to provide an understanding of physical, chemical, biological, and geological phenomena in the oceans, primarily to understand their effects on underwater acoustics.

It seems appropriate, therefore, in observing ONR's 30th anniversary that we have assembled a group of papers that focus on some of the above pursuits. We have asked the distinguished authors not merely to review past accomplishments but to indicate where matters stand today and to assess the prospects for the Navy in their areas of expertise. Obviously, not all fields pertinent to the Navy could receive attention, nor was an attempt made to give equal space to all topics. The fact that the contributions come from a variety of disciplines and institutions reflects the need and desire of ONR to draw upon the expertise of scientists and engineers in government, industry, and the universities.

It would be unseemly not to thank the authors, many of whom sacrificed part of their summer vacation to meet our publication deadline. And many thanks go to members of the Editorial Committee: Dr. P. C. Badgley, M. Denicoff, Dr. G. Goldstein, H. Fitzpatrick, Dr. George Neece, Dr. J. J. O'Hare, Dr. D. W. Padgett, Dr. D. Paskausky, Dr. D. P. Woodward, and Mrs. Lois A. DeCatur.

E. I. SALKOVITZ
Chairman of Editorial Committee

INTRODUCTORY REMARKS

The Honorable Melvin Price

Chairman, Armed Services Committee House of Representatives

The Hon. Melvin Price, Representative in Congress of Illinois' 23rd Congressional District, is Chairman of the Armed Services Committee of the House of Representatives and of its Subcommittee on Research and Development. He is also a member of the Joint Committee on Atomic Energy and of the House Committee on Standards of Official Conduct. Mr. Price was a newspaper correspondent and later became secretary to former Congressman Edwin M. Schaefer (1933–1943). He was elected to the 79th Congress while serving in the Army as an enlisted man; he has been reelected to each succeeding Congress. In 1946, Congressman Price became a member of the present House Armed Services Committee. In the same year, he became a charter member of the Joint Committee on Atomic Energy and in past years has served as the Committee's Vice Chairman and Chairman. Since 1958, he has been Chairman of the Joint Committee's Subcommittee on Research, Development, and Radiation. He has also served as Congressional Advisor to several Disarmament Conferences and International Conferences on Peaceful Uses of Atomic Energy. He was an early advocate of nuclear-powered submarines, and he is recognized as one of the best informed members of Congress on matters relating to National Defense, International Security Affairs, and the Peaceful Uses of Atomic Energy.



During its first thirty years, the Office of Naval Research has served the Navy and the Nation well. ONR was established in 1946 by Public Law 588—an innovative and forward-looking action by the 79th Congress (the Congress to which I was first elected). By its action, Congress demonstrated its appreciation of the increasingly important role of science and technology in the future.

The principal responsibility assigned to ONR was to encourage, promote, plan, initiate, and coordinate Naval research to provide for the maintenance of Naval power and the preservation of national security.

The Act includes provisions that have stood the test of time: a measure of independence, essential for carrying out fundamental and innovative work; a due regard for the efforts by other groups; the need for outside review and advice; and, importantly, the special nature of contracting for research and

development. The success of these farseeing provisions and of the operation of the Office of Naval Research based on them have had a marked influence on similar activities of other government agencies now engaged in support of research.

The central thesis of the Act, the dependence of Naval power on scientific research, has been borne out in time, and ONR has played an important role in providing fundamental understanding leading to advances in many Naval capabilities: navigation, sensors, computers and communications, logistics, ocean measurement and prediction, and training . . . details are out of place here and no doubt will be found in the papers presented in this volume.

While in one respect circumstances have surely changed, with ONR moving from a central place in the national research picture to a lesser role, there is a parallel to the situation facing the Navy as a whole—while its resources are smaller, its responsibilities have never been greater. During the past decade the Soviets, if they have not surpassed us in terms of naval capability, have certainly closed the gap to where our ability to control the seas is questionable. We do as a Nation, however, possess the technology to reverse this trend, thanks to the efforts of research organizations such as the ONR.

The Navy must not, however, take on an air of complacency with regard to the ONR. It is imperative that the ONR and the Navy constantly assess and reassess the dynamics of the world situation from both an operational and technological viewpoint and insure that they maintain the vitality and capability needed to meet the challenge. In pursuit of our long-term research goals, the ONR has an excellent record and sound operating principles. On this basis I look forward to a continuation of ONR's remarkable record in meeting the future challenges in science for the Navy.

PHYSICAL SCIENCES



Herbert Friedman, Superintendent of the Space Science Division and Chief Scientist of the E. O. Hulburt Center for Space Research, has been associated with the Naval Research Laboratory throughout his professional career. He conducted his first rocket astronomy experiments in 1949. He has participated in numerous satellite programs and more than a hundred rocket experiments. These experiments traced the cyclic variations of solar X-rays and ultraviolet radiations, revealed the ultraviolet fluxes of early-type stars, and led to the discovery of X-ray stars, X-ray galaxies, and the X-ray pulsar in the Crab Nebula. Dr. Friedman has served on the President's Science Advisory Committee and as President of two international commissions—the Inter-Union Commission on Solar-Terrestrial Physics of the International Council on Scientific Unions and Commission 48 on High Energy Astrophysics of the International Astronomical Union. He has been granted some 50 patents and has published some 200 papers. He has received more than a dozen awards, among them the President's Award for Distinguished Federal Civilian Service, the Rockefeller Public Service Award, and the highest DOD and Navy awards. Dr. Friedman earned a B.S. from Brooklyn College and a Ph.D. in Physics from the Johns Hopkins University. He is a member of the National Academy of Sciences, the American Philosophical Society, the American Academy of Arts and Sciences, and the International Academy of Astronautics.

PHYSICAL SCIENCES

SOLAR-TERRESTRIAL PHYSICS

Herbert Friedman

E. O. Hulburt Center for Space Research
Naval Research Laboratory
Washington, D.C.

We have recently marked the 50th anniversary of the discovery of the ionosphere and the beginnings of the scientific discipline of solar-terrestrial physics. In 1924, radio waves were echoed from heights as great as 300 km by Edward Appleton and his colleagues in England. Within a few years, G. Breit and M. Tuve at the Carnegie Institution developed the pulse sounding technique, and E. O. Hulburt and A. H. Taylor at the Naval Research Laboratory (NRL) began to outline the features of diurnal control of the ionization by solar radiation. Subsequent theories attempted to relate hypothetical models of the structure of the upper atmosphere to an invisible spectrum of solar ionizing radiations. From those early years to the present time, the Sun and the upper atmosphere have been studied with sensors carried aloft with balloons and aircraft and, finally, with rockets and satellites.

EARLY HISTORY

Before the advent of modern rockets, only the lower 30 km could be directly sampled. In the 1920s, balloon instruments recorded atmospheric temperature and pressure into the stratosphere. Temperature decreased steadily up to about 12 km and then remained nearly independent of height. Pressure varied as expected in a fully mixed atmosphere of molecular oxygen and nitrogen, but diffusive equilibrium was believed to control the distribution of atmospheric gases at greater heights. Even though helium was only a trace constituent in ground-level air, it was expected to dominate the atmosphere about 100 km because its lower atomic mass would give it a scale height eight times as great as oxygen and nitrogen. This simplistic picture was soon challenged by studies of the luminous trails of meteoritic particles as they heated to incandescence in the Earth's atmosphere near 110 km and evaporated completely by 80 km. The meteor observations required that the air be denser at 100 km than expected if the temperature were the same as observed at 12 km. Accordingly, the temperature at an altitude of 100 km must have returned from the cold of the stratosphere to the warmth of ground level. At these higher temperatures helium would not dominate over oxygen or nitrogen until heights as great as 300 to 400 km.

Further evidence of the temperature structure of the atmosphere was obtained by observing the reflection of sound waves from explosions. From the arrival times of explosive sounds and their angles of incidence at distant points, it was deduced that temperature in the stratosphere was lower by 70°C than at ground but increased rapidly above 30 km until it greatly exceeded ground-level temperature.

While meteor and sound wave studies were giving new insights into the high-altitude temperature and pressure structure, theorists were also beginning to deal with the photochemistry of the upper atmosphere. Solar ultraviolet radiation is cut off at about 2900Å by a trace of atmospheric ozone. From studies of the change in the absorption limit wavelength near sunset, the center of the ozone layer was placed at about 25 km. Then, it was deduced that ozone was produced by the dissociation of molecular oxygen under the influence of solar ultraviolet in the Schumann-Runge bands, followed by recombination of oxygen atoms with O_2 to form O_3. Simple photoequilibrium theory implied that O_2 would be completely decomposed to atomic oxygen above 150 km. Thus, the high atmosphere would consist of molecular nitrogen, atomic oxygen, and helium. Hydrogen was not thought to be an important constituent.

The distribution of energy in sunlight from infrared to the ultraviolet ozone cutoff closely resembles a 6000°C black-body spectrum with a peak near 5000Å. At shorter wavelengths in the ultraviolet, the energy would be expected to decrease rapidly, and at X-rays it would be inconsequential. It was difficult to account for the ionization of the upper atmosphere with this input energy distribution. Soft X-rays would have the correct absorption profile to affect the E-region (90–150 km) and extreme ultraviolet would produce the F-region (> 150 km), but a 6000° Sun was not an adequate source of these energetic photons. Faced with this dilemma, ionospheric researchers grasped with enthusiasm the opportunity offered by the availability of German V-2 rockets after World War II for direct study of the Sun's short wavelengths.

The NRL Rocket Sonde Branch was established in January 1946, under the leadership of Ernst Krause, to begin preparations of scientific payloads for atmospheric, ionospheric, and cosmic ray research. E. O. Hulburt, who was then superintendent of the Optics Division, saw a great opportunity for studying directly the solar ionizing radiations that were absorbed in the ionosphere. The research program that was set in motion in 1946 has continued through the full three decades of the Office of Naval Research's history, marked by continuous support of the space science effort at NRL. The tradition of the study of solar-terrestrial physics initiated in Hulburt's era still runs strong in the laboratory that now bears his name, the E. O. Hulburt Center for Space Research at NRL.

THE ROCKET YEARS

Hulburt dusted off a small Hilger quartz spectrograph that had seen service in auroral research during the Second International Polar Year, 1932–1933, and offered to sacrifice it in a rocket flight to observe the solar ultraviolet below 3000Å. Richard Tousey and his colleagues quickly recognized that Hulburt's simple approach would not suffice. Within 3 months, an innovative design for a spectrograph was developed into a flight instrument and on October 10, 1946, it brought back the first solar ultraviolet spectrogram to a wavelength of 2200Å (Figure 1).

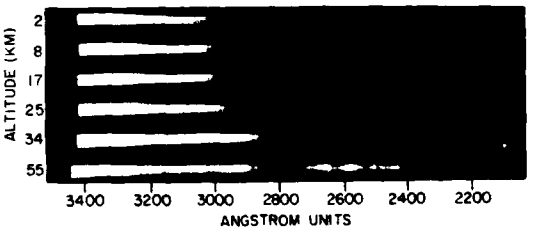


Figure 1—First solar spectrum obtained at high altitude, October 10, 1946 (NRL)

NRL was not alone in the early attempts to measure the solar ultraviolet spectrum. J. J. Hopfield and H. E. Clearman, at the Johns Hopkins University Applied Physics Laboratory, obtained excellent results only 6 months after NRL, but it was immediately apparent that extension of the spectrum to shorter wavelengths would require means of pointing the spectrograph at the sun from a stabilized platform. From one-axis stabilization (first used by NRL) to two-axis stabilization (developed by the University of Colorado for the Air Force) was a major technological development, but it took 6 years to complete. The biaxial pointing control was the most important single contribution to instrumentation for solar astronomy, until the Orbiting Solar Obser-

vatory series of NASA was developed a decade later.

Early models of the ionosphere predicted a simple succession of stratified layers, each controlled by an essentially monochromatic input. Rocket measurements quickly revealed a continuum of ionizing radiation marked only by slight inflections that gave the deceptive impression of layer structure in the reflection of radio pulses. Rocket-borne mass spectrometers showed that electron loss processes are controlled by complex ion chemistry, and trace constituents can dominate the reaction chains. Certainly the most surprising result of the early mass spectrometer observations of C. Y. Johnson and his colleagues at NRL was the discovery that nitric oxide ions dominated the E-region, even though the molecule is a minute trace constituent of the neutral atmosphere. Ionospheric weather is always disturbed by a variety of winds, waves, and drifts. "Very large traveling disturbances" follow magnetic storms; smaller scale disturbances are common on a day-to-day basis. Acoustic waves are generated by violent tropospheric storms and gravity waves propagate all the way from ground to well above 100 km.

The most elementary considerations of the solar corona require temperatures in the million degree range and an appropriate X-ray emission. From the outset of the NRL rocket astronomy program, detection of solar X-ray flux was given high priority. After some early attempts to detect X-ray blackening of film behind suitable filters and to excite thermoluminescence in a CaF_2:Mn phosphor, quantitative flux data versus altitude were obtained with photon counters carried aboard a V-2 rocket in 1949. The observed X-ray intensity (1–8Å) was sufficient to account for a major part of lower E-region ionization (Figure 2).

On the same flight, a hydrogen Lyman-α detector responded to a strong flux in D-region (75–90 km), which supported a hypothesis of M. Nicolet that ionization of a trace of nitric oxide by hydrogen Lyman-α was the effective electron production process. Radiation in the Schumann continuum (1450–1600Å) was detected above 90 km and increased steadily to the peak altitude of 151 km. Absorption by molecular oxygen was, therefore, not confined to a sharp

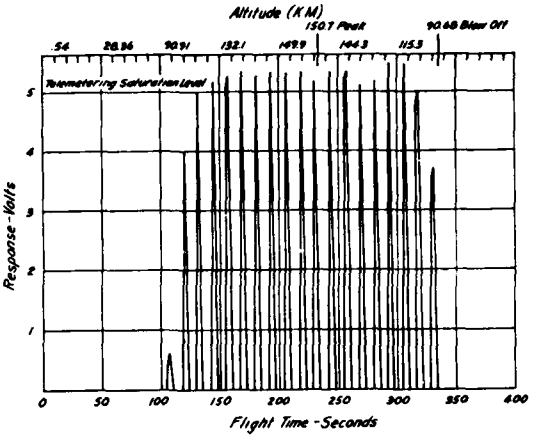


Figure 2—Telemetry record of solar X rays (1-8A) obtained from spinning V-2 rockets, September 20, 1949 (NRL)

equilibrium layer near 90 km as was predicted by photochemical equilibrium theory. Instead, it was clear that a significant concentration of O_2 persisted into the F-region. Following the 1949 measurements, broadband photometry of the X-ray spectrum was extended by the use of a variety of window materials and filters on the detectors. The spectrum was found to fit approximately with thermal radiation at a few million degrees. Over a period of years from minimum to maximum of the solar cycle, marked variability was observed—as much as a factor of 7 for X-rays (8–20Å). The flux variations were consistent with variations in ionospheric electron density and it appeared that X-rays were a controlling source of E-region behavior (Figure 3).

At the 1954 Cambridge Conference on the Ionosphere, Havens, Friedman, and Hulburt proposed a tentative model of the ionosphere based on early evidence of the full distribution of solar energy through the XUV and X-ray regions of the spectrum. Although no quantitative data existed on fluxes between the soft X-ray range and the hydrogen Lyman-α limit, it was proposed that most of the emission was attributable to the neutral and ionized helium resonance lines at 304Å and 584Å and the helium continuum. The E-region loss process was assumed to be dissociative recombination of molecular oxygen and an effective recombination coefficient at each altitude was computed on the basis of charge ex-

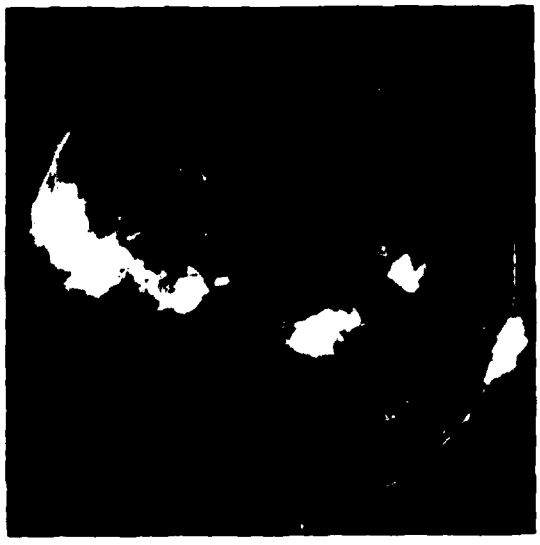


Figure 3—XUV image of the Sun in the wavelength band 150–600Å, obtained on January 15, 1974. Most of the emission originates in highly ionized atoms, Mg IX, Mg X, Si XII, Fe XIV, Fe XV, and Fe XVI, which are produced at plasma temperatures of 1–2.5 million K.

change between O^+ and O_2. At F-region altitudes the loss process reduced to simple photorecombination of atomic oxygen. Surprisingly, the equilibrium ionosphere derived from this elementary model was nearly correct. Subsequent observations filled in the solar spectrum in high-resolution detail from X-rays to near UV, but the essential model was not substantially altered. According to spectral flux measurements achieved during the decade of the 1950s (University of Colorado, Air Force Cambridge Research Laboratories (AFCRL), and NRL), F-region sources contained the Lyman continuum and the band 350–200Å, including He II (304Å), and adding up to about 1 erg cm^{-2} s^{-1}. Lines of He I (584Å), Mg X (625,610Å), and Si XII (520,500Å) added another 0.4 erg cm^{-2} s^{-1}. The short wavelength range, 170–211Å, is dense with lines of Fe VIII to Fe XIV that contribute another 1 erg cm^{-2} s^{-1}. In E-region (90–140 km), H-Ly β (1,025.7Å), C III (977Å), and part of the Lyman continuum (910–800Å) contributed as much energy as X-rays.

While the photoionization and loss processes in E- and F-regions could be well understood on the basis of early rocket studies, D-region processes still remain difficult to untangle. The lowest region of the ionosphere is the seat of all disruptions of HF communications that accompany the prompt radiation flash of a solar flare. Because of the greater density of D-region compared to E and F, the collision frequency is high and shortwave radio signals are strongly absorbed when the ionization is sharply increased. The various forms of sudden ionospheric disturbances are classified as

1. SWF—shortwave fadeout (5–20 MHz) as a result of absorption. It begins promptly (within a minute) of the rapid onset of the flare.
2. SCNA—sudden cosmic noise (background microwave emission of the galaxy) absorption detected at about 19 MHz on radiometers.
3. SPA—sudden phase anomaly. The sky wave changes phase with respect to the ground wave when ionization is produced at lower heights, sometimes as much as 16 km in a large flare.
4. SEA/sudden enhancement of atmospherics. Signal strength increases on very long waves (about 10 000 m) reflected from the bottom of D-region. The atmospherics are generated by tropical thunderstorms.
5. SFA—sudden field strength anomalies. Interference effects occur between sky wave and ground wave as the reflecting ceiling moves down or up.

The D-region is variable on a day-to-day basis as the result of atmospheric factors as well as the activity of the Sun. Perhaps the most striking variation is the winter anomaly at middle and high latitudes, where the electron concentration may increase as much as tenfold near 80 km. The enhancement appears to be statistically connected with increases in temperature of the stratosphere near 30 km. Contributing factors may include (1) the effect of large-scale mesospheric circulation on the distribution of ionizable minor constitutents; (2) a change in mesospheric temperature sufficient to change the rate coefficient for the formation of nitric oxide, the principal ionizable constituent. Altogether, the evidence is persuasive that meteorological factors have a strong influence on D-region variability.

What contributes to D-region complexity is the abundance of minor constituents, including H_2O, OH, H_2O_2, NO, NO_2, N_2O, CO, CO_2, and CH_4. An intense infrared airglow is produced by these

molecules. Only in recent years have the existence and importance of hydrated and conglomerate ions been recognized, largely through the work of R. Narcisi at AFCRL. Oxonium (H_3O^+) and hydronium ($H_3O^+H_2O$) play a role but also significant are O_4^+, $O_2^+(H_2O)$, $H_3O^+(OH)$, $H_3O^+(OH)(O_2)$, $H_3O^+(OH)(H_2O)$, and $H_3O^+(H_2O)_n$. Near the top of D-region, about 95 km, meteoritic debris collects in a layer of atomic ions. Whereas early theory was concerned only with the negative ion, O_2^-, present modeling includes O_3^-, NO_3^-, and NO_2^-, CO_3^-. Electron attachment to neutral particles forms negative ions at a rate comparable to ion-electron recombination, and at night the negative ion concentration has an especially important effect on the loss process. It is obvious why D-region has been called the "chemical kitchen" of the ionosphere.

There is special military interest in the disruption of D-region by the debris of a nuclear explosion. Energetic particles are trapped on magnetic field lines and oscillate back and forth between conjugate points. Some of the particles are dumped into D-region and induce intense HF radio wave absorption. As the electrons circulate about the field lines, they draft eastward. In about 1 hour they can blanket the earth with D-region absorption. Sometimes the phenomenon lasts for days as electrons slowly leak out of the geomagnetic trap. At the time of explosion, enormous amounts of nitric oxide are generated in the fireball. The ensuing decrease in stratospheric ozone can be very substantial over the entire globe.

X-RAY FLARES

In 1954 Friedman and Chubb proposed that sudden ionospheric disturbances (SID) were the result of enhanced solar flare X-ray emission and the attendant ionization of D-region down to 60 km. Most prior theories had assumed that the ionizing source was solar flare enhanced Lyman-α, the same radiation which produced normal D-region. but theoretical analysis showed that SID phenomena required that the Lyman-α intensity increase by a factor of 10^4, or a flux as high as 10^4 erg cm^{-2} s^{-1}. On the other hand, these effects could be produced by fluxes as low as 10^{-5} erg cm^{-2} s^{-1} of one to two angstrom X-rays. While the Lyman-α requirement is astrophysically impossible, the X-ray enhancement could readily occur if the solar flare heated a small volume of the corona to a few tens of millions of degrees.

To test the X-ray hypothesis, it was essential to achieve a rocket launching in coincidence with a solar flare. Since flares are relatively short lived and unpredictable, a quick-reaction rocket-launching capability was needed which could send a standby payload aloft at a moment's notice. V-2s, Vikings, and Aerobees, which comprised the stable of research rockets at the time, all used liquid propellants that could not be stored in the rocket in the launching tower for more than a few hours. A military rocket, the Deacon, 9 ft (2.7 m) long and 6 in. (0.15 m) in diameter, was the only solid-propellant vehicle available and it could reach a height of about 40 km when launched from the ground. If the Deacon were carried to 25 km on a ballon, however, it could be ignited at that altitude and would then climb to well above 100 km. J. Van Allen, who conceived of the combination of Deacon and Skyhook ballon, named the system the Rockoon.

In the pre-IGY year, 1956, an expedition called Operation San Diego-Hi was organized to study solar flare radiation. Rockoons were released from the deck of a Navy landing ship dock, the U.S.S. *Colonial*, about 400 mi (645 km) out to sea off the coast of southern California. In the early morning of each day, a 150,000 ft^3 (4245 m^3) polyethylene balloon carried a Deacon rocket aloft. As the rocket floated at an altitude of 25 km, it was followed by the ship, which could communicate via teletype to the High Altitude Observatory at Boulder, Colo., and the Sacramento Peak Observatory in New Mexico. When a message was received alerting the rocket experimenters of the start of a flare, the Rockoon could be fired by radio command. It would then climb above the absorbing atmosphere to measure the solar X-ray flash of the flare. Out of 10 tries, success was achieved on 1 day when a class 1 flare occurred. The enhancement of X-ray emission was very pronounced whereas Lyman-α was almost unchanged. The cause-effect relationship between flare X-rays and SID was thus established.

Beginning about 1957, two-stage combinations of solid-propellant rockets such as the Nike-Deacon replaced the Rockoon for flare studies.

The Nike served the booster function previously performed by the balloon. A sufficient number of solar flare X-ray measurements were made from 1957 to 1959 to show that large flares produced intense fluxes of very short wavelength X-rays. The highest energies, around 20 keV, penetrated as low as 43 km. For the most intense flares the emission spectra could be fitted approximately with thermal sources at temperatures as high as 10^8 K. In 1959, Peterson and Winckler observed with balloon-borne equipment a burst of X-rays which they estimated to have lasted about 18 s and whose energy reached 500 keV.

The decade of the 1960s brought in the NRL Solrad program and the NASA series of Orbiting Solar Observatories (OSO's). Solrad-1, 1960, immediately confirmed the solar flare X-ray control of SID behavior in the D-region. Threshold for SID was found to be 2×10^{-3} erg cm^{-2} s^{-1} (1–8Å). The OSO's were a sophisticated series of solar observatories that inaugurated a new era of solar physics, which reached its climax with the Apollo Telescope Mount on Skylab. The superb results of that mission will be analyzed for years and have already established a strong case for major future programs in solar physics.

THE LIGHT OF THE NIGHT SKY

The overhead sky is not totally black between the stars. On a dark moonless night, far from city lights, the eye can detect a faint glow. Much of this airglow is produced at heights from 60 to 300 km above the ground and can be identified with excited atoms and molecules of oxygen and nitrogen. Far more spectacular are the auroral lights, colored forms often seen in rapid motions across the arctic and antarctic skies in regions surrounding the magnetic poles of the Earth. Early triangulation measurements showed that the altitude was about a hundred kilometers. The auroral light thus provided a means of learning about the nature of the atmosphere at heights far above the reach of available experimental probes before rocketry was developed.

In the 1920s it was believed that the aurora was produced by electrons streaming from the Sun. As the electrons approached the Earth, the magnetic field would bend their paths so that they impacted the atmosphere uniformly on the day and night sides but concentrated in circular regions around each pole. The violent changes in auroral light implied corresponding variability in the flow of electrons from the Sun.

The spectrum of auroral light contained lines of molecular nitrogen and a green line, whose origin remained a mystery for many years. It did not appear in gaseous discharge tubes in the laboratory, but by 1924 it was finally identified with atomic oxygen. The reason for its absence in laboratory discharges is that its lifetime against emission is long, about 0.5 s, and it is deexcited by atomic collisions before it radiates. In the low pressure of the upper atmosphere, collisions are so infrequent that excited oxygen has time to radiate.

With the advent of rockets, it became possible to determine the altitudes of midlatitude airglows directly and it was immediately found that the early estimates were in error by factors as large as 2 or 3. An upward viewing photometer on a rocket sees the full airglow from below the emitting region, but the measured intensity decreases as the airglow layer is traversed. The differentiated curve of airglow intensity versus height typically shows a layer distribution. The green line of atomic oxygen (5577Å) is emitted in a sharp layer near 100 km and a weaker line (6300Å) in a broad range, maximizing at about 240 km.

Rocketry also made it possible to observe in the ultraviolet below 3000Å where the resonance transitions of most atmospheric gases occur. In 1955, an NRL rocket photometer measured hydrogen-Lyman-α (1216Å) above 85 km and found it to be more intense than all visible airglow. Atomic oxygen (1304Å and 1356Å) and molecular nitrogen Lyman-Birge-Hopfield bands (1300Å) to 1600Å) were observed in subsequent flights. The oxygen and nitrogen emissions are strong in the daylight hemisphere where they are excited to resonance by direct sunlight but are not detectable at night. Hydrogen Lymanα is intense at night because the hydrogen extends to very great altitudes in the form of an extended geocorona some 50,000 mi (80,450 km) radius. Sunlight scatters from far reaches of the geocorona back into the nightside shadow. Because hydrogen is light enough to escape gravity, the geocorona must be continuously replenished from below. The hy-

drogen comes from the photodissociation of water vapor and other hydrogen-bearing compounds by solar ultraviolet.

In recent years, airglow measurements have been extended to the extreme ultraviolet shortward of hydrogen Lyman-α, where neutral helium (584 Å) and ionized helium (304 Å) are detected. An excellent portrait of the airglow was obtained from an NRL far ultraviolet camera/spectrograph employed on the lunar surface in the Apollo 16 mission in April 1972. The electrographic camera photographed the earth in the ranges 1050–1600 Å to reveal the extended hydrogen Lyman-α corona, the auroral ovals, and equatorial airglow arcs of oxygen, 1304 Å and 1356 Å. Two bands of oxygen airglow stretch from opposite sides of the equator, converging toward the equator on the dark side. They are produced by combination of oxygen ions with electrons. It is believed that upper atmospheric winds drive the O^+ against the magnetic field so as to concentrate the ions into the belts (Figures 4, 5).

Downward-looking photometric observations in the far ultraviolet have revealed patchiness in the atmospheric airglow which is most likely caused by small-scale inhomogeneities in composition. Imaging devices could exploit the ultraviolet pattern as the basis for a "meteorology" of the high atmosphere, which may be important for understanding ionospheric irregularities and their impact on radio transmission.

Unlike the airglow, which arises from the flux of solar electromagnetic radiation on the atmosphere, the aurora is produced by the dumping of energetic charged particles (protons and electrons). Because charged particles can enter only along magnetic field lines, auroral phenomena usually are confined to well-defined zonal rings, or ovals, surrounding the magentic poles. Observing the aurora from space opens up the entire electromagnetic spectrum and provides total geographic perspectives very difficult to achieve from the ground. An entire auroral oval can be photographed, and from sufficiently high altitudes both the northern and southern auroral ovals are observed simultaneously. In the ultraviolet, the aurora can be detected on the sunlit side of the Earth because day airglow is relatively very weak. Also, no ultraviolet emerges from levels below 90 km and the Earth looks nearly black

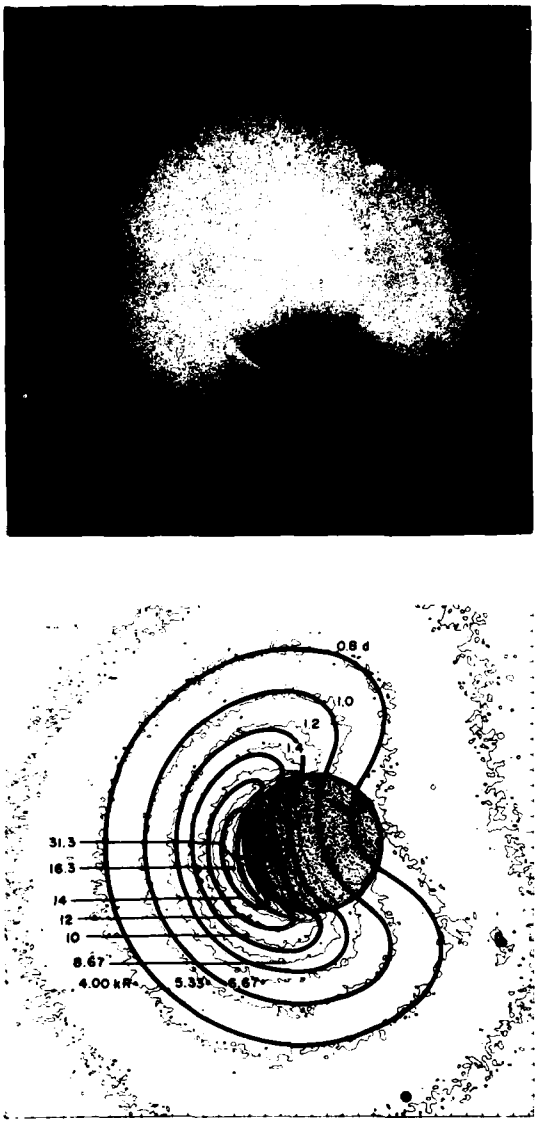


Figure 4—(A) Photograph of Earth in hydrogen Lyman-α, taken from Moon on Apollo 16 mission, with Carruthers electrographic camera. (B) Fit of radiation intensity contours to theoretical model of resonant scattering. (NRL)

underneath the high-altitude aurora. Most of the incoming particle energy is transformed to soft X-rays, which emerge freely from the atmosphere. Hence, a total X-ray albedo measurement in space can be an accurate gage of input energy. Such measurements are being attempted from the Solrad-Hi satellites now in orbit.



Figure 5—Photograph of Earth in oxygen resonance line (1304Å) shows magnetically controlled equatorial emission bands attributed to oxygen recombination radiation, on both sides of the dip equator. Aurora and dayglow are heavily overexposed. (NRL)

Satellite far UV observations have distinguished clearly between electron-induced auroras and proton auroras. Lyman-α is emitted strongly in the latter and negligibly in the former. NRL experiments on the Orbiting Geophysical Observatory, OGO-4, observed that hydrogen Lyman-α intensity is depressed over the polar caps. It appears that hydrogen ions are escaping along the "open" magnetic field lines in the polar regions. This polar wind escape route may greatly increase the rate of loss of hydrogen from the terrestrial atmosphere and, perhaps, be even more important for helium. With high-resolution imaging in various UV colors, using a Carruthers electrographic camera, we would have a powerful means of studying auroral phenomena.

IONOSPHERIC IRREGULARITIES

Until comparatively recent years, aeronomy was content to fit grossly averaged data on solar radiation and atmospheric composition to a standard model of solar-terrestrial relationships. That picture was as incomplete as any model of the lower atmosphere would be without dynamic weather changes. With more global data and greater temporal detail, we are coming to recognize the great influence of weather in every level of the upper atmosphere. Winds, waves, and drifts distort the largest scale features of static models and produce the fine-scale irregularities that are of such importance to modern communications. Ground observers have known for many years of the movement of ionospheric disturbances from high latitudes toward the equator. It is now clear that auroral heating generates huge high-altitude waves that produce these traveling ionospheric disturbances.

Local irregularities in plasma density spread the F-region return of radio signals into a multiplicity of echoes, a phenomenon known as "spread-F." An early indication of these irregularities came from radio astronomical observations of scintillating signals from pointlike sources such as quasars. The phenomenon is analogous to the optical twinkling of stars that results from refractive index irregularities due to turbulence in the lower atmosphere. The F-region irregularities produce strong scintillation and fading on higher radio frequencies, even the GHz frequencies associated with communication satellites, which were once thought to be the answer to trouble-free communications. The practical implications of F region scintillation are manyfold. Starfish, a high-altitude nuclear burst, produced worldwide spread-F. Video pictures from meteorological satellites are often blurred by scintillation. At times of scintillation, navigational satellites have had difficulty inserting ephemeris data. Ionospheric tilt in the polar cap scintillation region leads to deterioration in communications over the pole from geostationary satellites.

F-region irregularities are a constant phenomenon over the polar regions but also frequently affect the equatorial ionosphere at night, especially near the equinoxes. Satellite observations show typical variations of three orders of magnitude in the amplitude of plasma inequalities over a single polar orbit. In some regions the plasma is almost perfectly smooth; in others, incredibly rough. At high latitudes we are, undoubtedly, seeing the effects of particle precipitation and the small-scale electric fields associated with auroras. The equatorial behavior, as yet, has no satisfactory explanation. Sophisticated modeling pro-

grams at NRL may be expected eventually to reveal the appropriate interactions between winds and plasmas that lead to formation of irregular distributions of blobby plasma.

The current series of NASA Atomspheric Explorers deliver a startling panorama of irregularities in ionospheric structure. Gross variations, as much as two orders of magnitude in plasma density, appear over horizontal distance of only a few kilometers. On a microscale, 50% changes appear over just a few tens of meters. Although many hypotheses are offered to explain F region irregularities, none is clearly correct. It is interesting that large irregularities are most frequently noted when meteoritic debris (Mg^+, Fe^+, Si^+, Na^+) is abundantly present. Theorists have shown that a few metal ions per cubic centimeter at 150 km can have a greatly amplified effect on the movement of an entire plasma tube at higher altitudes compared to the much higher concentration of atomic oxygen ions within the tube at 300 or 400 km.

At lower altitudes, sporadic-E is a frequent irregularity. It can reflect waves that would normally be transmitted on high frequency and cause the signals to be received as far away as 2000 km from the source on a single hop. In summer, sporadic-E is the cause of severe interference on TV broadcasts. With rocket probes, the form of sporadic-E has been defined as a sharp stratum of ionization near 100 km. It usually extends over a radius of 100 to 200 km, but its thickness is only 2 or 3 km. In midlatitudes, sporadic-E is common near midday in the summer. The layer is populated by meteoritic ions, but the detailed mechanism of how they concentrate in such sharp layers is not well understood (Figure 6).

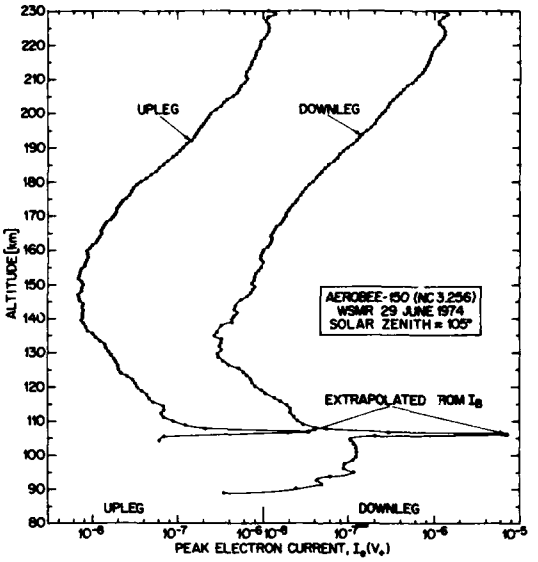


Figure 6—Sporadic-E structure detected with pulsed plasma probe carried aboard Aerobee rocket launched at White Sands, N. Mex. (NRL)

THE SOLAR WIND

Prior to 1957 it was believed that the Sun's influence on the Earth's atmosphere was primarily via photoionization which created the ionosphere and, sporadically, by streams of charged particles which produced ionospheric and magnetic storms and auroras. Solar magnetic fields appeared to confine the solar corona primarily to the near vicinity of the Sun and the Earth's magnetic field served to bind ionized gas to the earth. Interplanetary space was assumed to be highly empty, although a very tenuous extension of an essentially static corona could reach past the Earth's orbit.

We now know that a solar wind streams steadily from Sun to Earth and at times gusts strongly. The outward expansion of the Sun's atmosphere creates the wind, which is supersonic throughout the interplanetary medium. Beyond a few solar radii, the rarified atmosphere is nearly collision free and electron currents may flow with almost negligible resistance. In this manner, magnetic field is "frozen" into the solar wind and carried into the interplanetary medium. At the same time that solar magnetic field is drawn outward by the wind, solar rotation twists the stream into an Archimedes spiral as seen from above the ecliptic plane.

In Parker's development of the hydrodynamic theory of the solar wind, the flow begins in the lower corona. The velocity increases steadily up to about 400 km/s at about 20 solar radii, and the particle concentration reaches about 8 cm^{-3}, primarily hydrogen, with 2 to 4% He and a trace of heavy elements. These parameters fluctuate in time and space. Because the theory assumes a

spherically symmetric expansion, it offers no detailed model of the long-lived "plasma streams," which have very different velocities and densities.

Measurements with magnetometers aboard space probes show that the Earth's field resembles the dipole field of a simple bar magnet, decreasing inversely proportional to the cube of the radius out to about 13 Earth radii. At that distance the field becomes turbulent and drops to a much smaller value. At 20 Earth radii it decreases markedly once again but then becomes smooth and essentially constant with distance. The innermost region is the magnetosphere; its turbulent boundary is the magnetosheath. Where the supersonic wind first encounters the Earth's field a bow shock is observed.

The magnetized plasma carried by the wind is characterized by large-scale discontinuities, such as shock waves and neutral sheets, but its gross structure is dominated by a sector pattern. Within each sector, the polarity of magnetic field is predominantly toward or away from the Sun. Four or five sectors may fill a circumference of the interplanetary medium at the Earth's orbit and field reversal at a sector boundary is rather sharp. Sector patterns are characteristically stable on a time scale of a year or two, but change in polarity can come abruptly in the time of a single solar rotation. Some scientists believe that the existence of the sweeping pattern may have implications for geomagnetic perturbations that are somehow coupled to lower atmosphere pressure patterns.

It is difficult to trace phenomena within the sector pattern back to large-scale photospheric fields. There may possibly exist an overall solar magnetic pattern fundamentally separate from the mechanism responsible for sunspots and small-scale field structures. The latter are believed to be related to a basic poloidal field, which is stretched into a toroidal field by differential rotation of the solar atmosphere. When kinks in tubes of magnetic force emerge through the photosphere, they produce magnetic loop structures rooted in sunspots. One model of the solar magnetic field suggests that much of the area of the photosphere and inner corona is magnetically closed, as evidenced by the tightly knit structure of small loops seen in the Skylab X-ray and XUV photographs. The remaining large areas are open "holes" in the corona from which magnetic lines reach into the interplanetary medium and allow the solar wind to escape. Perhaps these holes are the hypothetical M-regions that Julius Bartels named many years ago as the features responsible for 27-day recurring geomagnetic activity.

THE MAGNETOSPHERE

The magnetosphere trap for charged particles over a wide range of energies from thermal to hundreds of MeV. Under varying pressure of the solar wind, the huge volume of plasma can balloon outward or contract. On the sunward side the solar wind pushes the magnetospheric boundary toward the Earth and combs the lines of force around the earth downwind into a stretched-out tail. The entire bag of plasma quivers in a quasi-periodic mode with characteristic time constants as though it fills and empties like a relaxation oscillator. When hit by the blast wave of a large solar flare, the sudden compression leads to a violent shakeup of the particle population accompanied by dumping of energetic particles into the auroral zones and the ionosphere.

Compression propagates an increase in geomagnetic field strength all the way to ground, producing the phenomenon of a "sudden-commencement geomagnetic storm." Energetic particles somehow manage to enter the trapped radiation belts where they oscillate back and forth in latitude and at the same time drift in longitude—electrons to the east and protons to the west. An equatorial ring current is thus generated at a distance of 3 or 4 Earth radii. Its accompanying magnetic field represents the "main phase" of a magnetic storm.

As particles leave the magnetosphere and find their way into the ionosphere on the nightside of the auroral oval, the auroral lights come on. A polar electrojet current develops which produces magnetic substorms at ground level. The energetic particles that shake out of the magnetosphere and enter the ionosphere far exceed the energy content of the solar wind. They are believed to have been stored in the magnetosphere and accelerated to high energy upon being triggered to release by the outburst of particles arriving from the Sun. Acceleration may occur in the magnetotail, which stretches for nearly a hundred Earth radii in

the antisolar direction. Its lines of force return to the polar regions of the Earth.

Lightning flashes generate radio noise, which propagates in the form of whistlers along geomagnetic field lines back and forth between the northern and southern hemispheres. The name "whistler" describes the audiotone descending rapidly in frequency which results from dispersion along the ducting path in which the wave is trapped. Whistler studies first identified a sharp decrease in electron density at about 4 Earth radii. This boundary was named the "plasmapause." It encloses the toroidal-shaped plasmasphere, a volume of relatively dense cool hydrogen extending upward from the top of F-region. Beyond the plasmapause the plasma undergoes a transition sharply to very low density and much higher temperature.

Magnetic micropulsations have been known from the time of early studies of the Earth's field with delicately suspended compass needles. With modern fast magnetometers, pulsations can be observed as fast as 0.1 s. Longer periods range up to 100 s. The pulsations are natural oscillations of the magnetosphere. They may arise from instabilities created on the surface of the magnetosphere as the solar wind sweeps over it. For a period of 1 s, the wavelength is about 1000 km in the magnetosphere and longer periods mean still longer wavelengths. Micropulsations provide a variety of information about magnetospheric structure and how the plasmapause moves during a substorm.

The equatorial radius of the plasmapause varies with local time and solar activity. Although only a small component of helium exists relative to hydrogen in the plasmasphere, the resonant glow of He^+ 304Å in sunlight provides direct evidence of movement in the plasmapause. The STP-72-1 satellite carried photometers which measured He^+ 304Å radiation not only in the ionosphere but also in the plasmasphere and detected oscillations of the plasmapause that accompanied magnetospheric substorms. Image converters operating in the XUV offer promising means of observing the "breathing" of the plasmasphere.

The International Magnetospheric Study (IMS), which will span 1976–1978, was inspired by the need to unravel the time sequence of magnetospheric events from spatial changes. Pairs of satellites are required to make simultaneous measurements across magnetospheric boundaries. NASA and the European Space Agency (ESA) are collaborating in a mother-daughter satellite mission called ISEE, for International Sun-Earth Explorer. During the lifetime of the IMS, many other coordinated observations will take advantage of various spacecraft in orbit. The two NRL Solrad-Hi satellites in their 65,000-n.mi. orbits will normally be spaced with one outside the magnetosphere and one inside. Following the IMS the approach to magnetospheric studies will shift from passive observations, such as the above, to active experiments carried by the shuttle or released and controlled from the shuttle. By deliberately perturbing various instabilities in a precisely controlled manner, it should be possible to interpret the resultant responses according to detailed models.

SOLAR PHYSICS

The study of the Sun itself is central to all aspects of solar-terrestrial physics. Flares, differential rotation, 11- and 22-year sunspot cycles, the hot corona, the flow of solar wind, the ejection of relativistic particles, and the missing neutrinos illustrate the diversity of baffling phenomena that have challenged solar physicists from past to present. The gross features of solar activity take large forms, easily visible from the ground, but their detailed mechanisms are driven by small-scale phenomena that can best be observed with spaceborne instruments, which achieve the highest spectral and spatial resolution. Before we can hope to predict solar variability and its ionospheric and tropospheric consequences, we must have a better understanding of it.

Sunspots reveal the complex hydrodynamics of the solar atmosphere. Historical records show a puzzling absence of spots from 1645 to 1710 and no evidence of a corona. From the relative drift of large spots at different solar latitudes we infer the differential rotation of the photosphere. At the same time, weak field regions seem to exhibit rigid rotation.

Recognition of the solar wind came only two decades ago and its association with coronal holes much more recently. Coronal holes also seem to

exhibit rigid rotation. The prediction and confirmation of the existence of the solar wind have led to a dynamical picture of a far-reaching solar corona that stretches throughout the solar system. From Apollo Telescope Mount (ATM) photographs, it appears that field lines are closed over young active regions. The plasma density is higher inside these regions and the corona is largely bound to the sun by these magnetic fields. But the corona is perforated with holes at all latitudes, especially near the poles, where the field lines are carried outward into interplanetary space by the expanding solar wind. The area of Sun covered by coronal holes is directly proportional to geomagnetic activity at earth.

Although the basic source of solar wind must be a fluidlike expansion at the base of the corona, there has been little theoretical effort to model the large-scale forms of the wind deep in the solar system. Exploration of the wind in the interplanetary space has been confined almost entirely to the neighborhood of the ecliptic. It is essential to study the wind at midsolar latitudes where solar activity is strongest but also is important to observe the wind directly over the poles where it emerges in a relatively undisturbed way. An out-of-the-ecliptic mission should have high priority for solar physics.

The flare mechanism may involve a variety of plasma instabilities and requires detailed study of all the available data from the ATM Skylab mission. Our present understanding of flare phenomena can be summarized briefly. The energy before flare release may be stored in unstable, current-carrying magnetic fields. The larger the flare, the longer the lapse time before the energy reservoir is refilled to permit another flare in the same region. Although the rapid build-up of flare radiation implies impulsive particle acceleration, there is often evidence of a preceding gradual heating phase which can be detected in radio, visible, XUV, and soft X-ray activity. High-resolution magnetic field observations reveal early changes in this preflare period.

The impulsive phase is generally characterized by hard X-ray and microwave bursts generated by the passage of highly accelerated particles through the corona. A major part of the energy of a flare must be carried by energetic electrons. In the main phase, the radio and X-ray emission is accompanied by evidence of mass motions—surges, eruptive prominences, and expanding clouds of nonthermal particles.

Mass ejection is vividly shown in ATM coronagraph pictures and in the telemetered images from the coronagraph aboard OSO-7. At the start of the main phase, a shock wave is sometimes observed which precedes the cloud of very energetic plasma. Radioheliograph observations show that accelerated particles pass through the corona but become trapped in very large magnetic loops. Further studies are needed of the propagation, trapping, and escape into the interplanetary medium.

Flare X-rays and radio bursts provide evidence of the acceleration of particles to high energies in solar flares, but the complexity of solar cosmic ray phenomena may yield newer insights into the flare mechanism and its attendant acceleration processes, as well as evidence of particle propagation within the solar atmosphere and nuclear reactions near the surface. The first observation of gamma ray emission lines, obtained from the OSO-7 satellite in 1972, suggests that much can be learned about surface nuclear reactions with more sophisticated gamma ray spectrometers. The deviation in composition of solar cosmic rays below 10 MeV per nucleon from normal cosmic ray abundance is a particularly intriguing puzzle. Exceptionally high deuterium and tritium accompanied by He^3 abundances that exceed He^4 are sometimes observed. Recurrent streams of meV protons are an almost constant phenomenon and have been found to persist over several solar rotations. Their lifetimes go well beyond the typical life of a flare-active region.

SOLAR MONITORING

Among its ultimate objectives, the study of solar-terrestrial physics seeks to relate observable phenomenology of the sun to prediction of its effects on communications. At the present time sudden ionospheric disturbances cannot be predicted with much certainty more than a matter of minutes to an hour before occurrence, but the duration of radio blackout can be estimated to within 5% from observation of the initial few minutes of rapid rise to maximum X-ray brightness at the outbreak of a flare.

Much progress can be expected in the capability of predicting the ionospheric and magnetic storminess that normally follows the electromagnetic flare outburst for several days. The Navy Solrad program, initiated in 1960, has produced a series of solar-monitoring satellites with progressively more sophisticated instrumentation. Solrad-Hi is a pair of satellites now in circular orbit at 65,000 n.mi. and spaced 180° apart that offers very nearly full-time observation of the sun over the ultraviolet and x-ray spectrum and in a broad range of particle energies carried by the solar wind. These satellites provide an operational system directly coupled to the fleet communications community. At the same time, Solrad is a research satellite which may be expected to reveal new, useful indices for prediction of the impact of solar activity on HF communications. A complementary program, Solwind, is being designed for STP-78-1. It will exploit the capability demonstrated by the OSO-7 coronagraph and the Skylab ATM XUV spectroheliograph to monitor solar plasma flow as it leaves the Sun. Solrad and Solwind combined represent a very promising approach to operational solar monitoring.

THE SHUTTLE ERA

With the advent of the shuttle, a variety of passive and active experiments of great diagnostic power can be carried out. The shuttle will permit the performance of mother-daughter experiments involving a "captive probe" or subsatellite released from the shuttle and reporting back to the shuttle and a comparable probe aboard the shuttle. Multiple probes may be released which will extend the simultaneous spatial coverage of time- and space-dependent phenomena. Active experiments will involve modification of ionospheric parameters by excitation of artificial airglow and aurora. It will become possible to perturb magnetospheric particle distributions through wave particle interaction processes.

The Earth's outer atmosphere is an excellent natural plasma laboratory free of the wall effects and attendant sheaths that complicate fundamental plasma studies in the ground-based laboratory. Its large dimensions and proportionate time scales for phenomena to develop make it possible to simulate laboratory plasma problems in ways that simplify study. In deeper space, collision-free plasma conditions are unique for studies of collision-free shock waves. In situ observations of wave-particle and wave-wave interactions, as well as wave-guide properties, can be conducted without perturbing the phenomena being investigated. Specific experiments designed for fundamental plasma physics studies may be performed with the capabilities offered by the shuttle.

Among the injection devices that are suitable for use aboard or release from the shuttle are the arc jet plasma gun, high-energy electron and ion accelerators, and low-energy beam injection devices. Plasmas of energies 10 eV to 1 keV may be injected in microsecond to millisecond pulses with output energies of more than 10 kilojoules (kJ) per pulse at repetition rates of several per minute. Existing designs for plasma propulsion, such as the magnetic-plasma dynamic arc, are suitable. A typical beam from this arc could carry 200 eV argon ions at 10,000 A. Electron or ion accelerators may operate up to 50 keV and deliver about 1kJ per pulse. In the lower energy range, hundred milliampere currents of electrons less than 10 eV can be provided easily.

The injection of dense plasma streams along geomagnetic field lines could heat the ionospheric plasma to thousand-degree temperatures and generate shock waves. Interaction of the plasma beam and shock wave with the neutral gas would then produce artificial airglow, for example, O I 6300Å. The airglow could be used to trace the dynamics of the atmospheric wind system in the F-region, where typical winds of 100 m/s are found. By firing the plasma gun repetitively, a train of glowing tracer clouds could be generated all along the orbit of the shuttle. Observers at ground level would have a means of observing the behavior of the ionospheric wind system on a global scale.

Barium oxide has been released from rockets to provide visible Ba ion tracers at 4554Å, the resonance line made visible by scattering sunlight. These tracer experiments reveal the orientation of magnetic and electric fields and the drift of plasma under their control. The luminosity is sufficient for good TV imaging. Artificial auroras can be induced with controlled energy ranges of electrons and ions. The spatial patterns of luminous

trails that they induce could reveal much information about waves and current sheets as well as plasma instabilities. In addition to the excitation lines detectable from the ground, such as O I (6300Å, 5577Å, and 8446Å), N II (3914Å), and Ba II (4554Å), the UV resonance line of O I at 1304Å can be observed from the shuttle.

It is hoped that the shuttle will carry a diversified traffic of free-flyer payloads. Already in the preliminary design stage are a series of Electrodynamic Explorers for the 1980s. These satellites will be paired—one in a circular orbit near 500 km, the other coplanar and in an eccentric orbit with adjustable apogee from 3 to 6 Earth radii. Coordinated measurements should provide a great deal of information about couplings between the magnetosphere and the ionosphere. At the present time we have only the sketchiest ideas of how the solar wind interacts with the magnetosphere and indirectly perturbs the ionosphere.

In recent years, horizon scanning from satellites has been an effective means of measuring concentrations of various atmospheric constituents. The simplest versions of instrumentation are narrow band photometers which observe the extinction of sunlight through the atmosphere. With the size and weight of equipment that can be carried on the shuttle, a vertical resolution of about 1 km should be attainable for NO, OH, O, O_2, and O_3. With Fabry-Perot interferometers for the infrared, similar accuracy should be possible for CH_4 and H_2O.

Between 60 and 140 km, the upper atmosphere is cooled by infrared radiation. Downward-looking infrared observations from spacecraft can determine the CO_2 and O_3 composition and the atmospheric temperature profile versus altitude. For the shuttle, infrared interferometers are being planned to cover the 1 to 5 μm and 5 to 150 μm ranges with cooled optics and detectors.

Lidar is a very promising technique for probing the atmosphere from the shuttle and, eventually, from shuttle-launched spacecraft. Operating as an optical analog of a pulsed radar in the middle ultraviolet (tunable 2200–3000Å), it can observe the time-delayed returns by Rayleigh scattering at different altitudes. At selected wavelengths in the absorption bands of specific constituents, their abundances versus altitude will become apparent. Among the candidate molecules for Lidar detection are O_2, O_3, NO, NO_2, N_2O, H_2O, OH, H_2, CO_2, CO, and CH_4.

European scientists have been considering an arrangement of one or two lasers with average power about 2.5 kW and a 1 m telescope to receive the backscattered radiation. The system would operate in the 0.2 to 10.6 μm range. For a first try on Spacelab, the telescope would be rigidly mounted so that scanning would require movement of the shuttle. Later flights could provide a rocking motion normal to the shuttle's longitudinal axis.

CONCLUSION

The many aspects of variability in solar radiation and solar wind combine with the dynamics of the magnetosphere and ionosphere to produce doubly complex patterns of solar-terrestrial relationships. To improve our understanding of the interactions, the Sun itself must remain a prime object of study. In the next generation of solar observatories, order-of-magnitude improvements need to be sought in spatial resolution at all wavelengths. Out-of-the-ecliptic missions and simultaneous measurements in all regions of the Sun-Earth system will be necessary to unravel the chain of interactions that accompany the propagation of radiation and plasma from Sun to Earth. The interpretation of such observations and the development of predictive capabilities will be greatly aided by modeling with advanced computers.

BIBLIOGRAPHY

S. Akasofu and S. Chapman, editors, *Solar-Terrestrial Physics*, Oxford Press, 1972.

C. DeWitt, J. Hieblot, and A. LeBeau, editors, *Geophysics, the Earth's Environment*, Gordon and Breach, New York, 1963.

John M. Goodman, editor, *Symposium on the Effect of the Ionosphere on Space Systems and Communications*, Jan. 22, 1975 NRL, 1975.

H. Odishaw, editor, *Research in Geophysics*, M.I.T. Press, Cambridge, Mass., 1964

"Physics of the Earth in Space," Space Science Board, NAS-NRC, 1968.

J. A. Ratcliffe, editor, "Fifty Years of the Ionosphere," *J. Atm. & Terrest. Phys.* 1975 Symposium on the Effect of the Ionosphere in Space, Jan. 22, 1975. **36** (1974).



Norman F. Ramsey is Higgins Professor of Physics at Harvard University. After temporary periods at the Carnegie Institution of Washington, the University of Illinois, the MIT Radiation Laboratory, and Los Alamos, he became an associate professor at Columbia University and Head of the Physics Department at Brookhaven National Laboratory before joining the faculty at Harvard in 1947. He has been on part-time leave from Harvard since 1966, when he became President of the Universities Research Association, which operates the Fermi National Accelerator Laboratory. In 1973, he was the Eastman Professor at Oxford University. Dr. Ramsey's work has ranged from molecular beams to particle physics, and he has concentrated on precision measurements of electric and magnetic properties of nucleons, nuclei, atoms, and molecules. He and his associates discovered the deuteron electric quadrupole moment, proposed the first successful theory of the chemical shift for the magnetic shielding of nuclei in nuclear magnetic resonance, and developed high-precision methods of molecular beam spectroscopy, including the atomic hydrogen maser. He received the Lawrence Award, the Davisson-Germer Prize, and the Presidential Certificate of Merit. His books include *Experimental Nuclear Physics* (1953), *Nuclear Moments* (1953), *Molecular Beams* (1956), and *Quick Calculus* (1965). Dr. Ramsey was born in Washington, D.C. He received the A.B. and Ph.D. degrees from Columbia University; the B.A., M.A., and Sc.D. degrees from Cambridge University; and the M.A. and D.Sc. degrees from Oxford. He also received an M.A. (hon.) from Harvard University; a D.Sc. (hon.) from Case-Western Reserve University; and a D.Sc. (hon.) from Middlebury College. He is Vice President-Elect of the American Physical Society.

ATOMIC AND MOLECULAR STANDARDS OF TIME AND FREQUENCY

Norman F. Ramsey

Harvard University
Cambridge, Mass.

In discussing the history and the prospectives of atomic and molecular standards of time and frequency, two alternative approaches are available. One is to treat all devices in parallel on a year-by-year basis. The other is to discuss each alternative device in succession. It is clear that the latter approach is the most suitable and will be followed here, but frequent cross-references will be given to other devices. In following this procedure, it is clear that the first technique discussed should be the molecular and atomic beam magnetic resonance method; historically it was the first, it stimulated the invention of the other methods, and it still remains one of the most effective time standards.

EARLY HISTORY OF THE MOLECULAR BEAM RESONANCE METHOD

The molecular beam magnetic resonance method arose from a succession of ideas, the earliest of which can be traced back to 1927, although it was rather remote from the idea of resonance. In 1927 the physicist Sir Charles Darwin [1]—the grandson of the great evolutionist—discussed theoretically the nonadiabatic transitions that make it possible for an atom's angular momentum components along the direction of a magnetic field to be integral multiples of h both before and after the direction of the field is changed an arbitrary amount. Inspired by Darwin's theoretical discussion, Phipps and Stern [2] in 1931 performed the first experiments on paramagnetic atoms passing through weak magnetic fields whose directions varied rapidly in space. Güttinger [3] and Majorana [4] developed further the theory of such experiments. Frisch and Segre [5] continued atomic beam experiments with adiabatic and nonadiabatic transitions of paramagnetic atoms and found, in agreement with Güttinger's and Majorana's theories, that transitions took place when the rate of change of the direction of the field was larger than or comparable to the Larmor frequency,

$$\omega_o = \gamma_I H_o \,, \tag{1}$$

which is the classical frequency of precession of a classical magnetized top with the same ratio γ_I of magnetic moment to angular momentum. Transitions did not take place when the rate of change to the direction of H was small compared to the Larmor frequency. However, some of the results of Frisch and Segre were not consistent with theoretical expectations. Rabi [6] pointed out that these discrepancies arose from the effects of the nuclear magnetic moments since some of the transitions were performed in such weak fields that strong or intermediate coupling between the

nuclei and the electrons prevailed. The transitions in such circumstances were quite different from those for which the effects of the nuclear spins could be neglected. Rabi showed that the results of Frisch and Segre were consistent with expectations if the effects of the nuclei were included. Rabi also pointed out that such nonadiabatic transitions could be used to identify the states and hence to determine the signs of the nuclear magnetic moments. Motz and Rose [7], Rabi [8], and Schwinger [9] in 1937 calculated the transition probability for molecules that passed through a region in which the direction of the field varied rapidly.

In all of the above experiments, however, the direction of the field varied in space and the only time variation arose as the atoms in the atomic beam passed through the region. Since the atoms possessed a Maxwellian velocity distribution, the atomic velocities varied and the apparent frequencies of the changing field were different for different velocities. Furthermore, the change in field direction ordinarily went through only a portion of a full cycle. For both of these reasons no sharp resonance effects could be expected. No suggestion was made initially to use an oscillatory magnetic field, i.e., a field that varied in time instead of space so that the apparent frequency would be the same to all the atoms, independent of their velocities. It is surprising that this possibility was not immediately recognized after Rabi's brilliant theoretical paper [8] in 1937. To simplify the theoretical analysis, Rabi assumed in 1937 that the field was actually oscillatory in time. As a consequence the results are all applicable to the resonance case with oscillatory magnetic fields even though the possibility of actually using fields oscillatory in time was not then recognized. Consequently this paper, without alteration, still provides the fundamental theory for molecular beam magnetic resonance experiments with oscillatory fields, even though the oscillatory field method was only invented by Rabi [10, 11] a year or so after the fundamental theoretical paper was written.

Gorter [12] in 1936 had suggested that nuclear transitions in solids could be induced by an oscillatory field from a radio-frequency oscillator. He proposed to detect the transitions by the absorption of the radio-frequency radiation and by the rise in temperature of solids subject to such oscillatory fields. Although Purcell et al. [13] and Bloch et al. [14] in 1946 successfully detected the absorption of such transitions by the reaction of the radiation on the radio-frequency circuits, Gorter's experiments [12] were unsuccessful in 1936.

Following a visit by Gorter to Columbia University in September 1937 in which he described his unsuccessful experiments, Rabi [10, 11] proposed the use of an oscillator-driven magnetic field as the transition-inducing field in a molecular beam resonance experiment. Two successful molecular beam devices using this method were soon constructed by Rabi [10, 11], Zacharias [10, 11], Kusch [10], Kellogg [11], and Ramsey [11]. A schematic view of these [10] is shown in Figure 1. In these experiments the atoms and

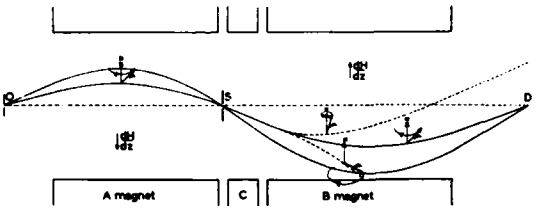


Figure 1—Schematic diagram [10] showing the principle of the first molecular beam resonance apparatus. The two solid curves indicate two paths of molecules having different orientations that are not changed during passage through the apparatus. The two dashed curves in the region of the B magnet indicate two paths of molecules whose orientation has been changed in the C region so the refocusing lost due to the change in the component of the magnetic moment along the direction of the magnetic field.

molecules were deflected by a first inhomogeneous magnetic field and refocused by a second one. When a resonance transition was induced in the region between the two inhomogeneous fields, the occurrence of the transition could easily be recognized by the reduction of intensity associated with the accompanying failure of refocusing. For transitions induced by the radio-frequency field, the apparent frequency was almost the same for all molecules independent of molecular velocity. As a result, sharp resonances were obtained whenever Eq. (1) was satisfied.

Rabi et al. [11] soon extended the method to the molecule H_2, for which the resonance frequencies depended not only on Eq. (1) but also on internal

interactions within the molecule. The transitions in this case occurred whenever the oscillatory field was at a Bohr frequency for an allowed transition

$$h\nu = E_1 - E_2 . \qquad (2)$$

For the first time these authors began speaking of their results as "radio-frequency spectroscopy."

MOLECULAR BEAM MAGNETIC RESONANCE EXPERIMENTS

By 1939 the new molecular beam magnetic resonance method had demonstrated its usefulness sufficiently well that it appeared to Rabi, Kellogg, Ramsey, and Zacharias to be of possible value for the definition of standard magnetic fields and for use as a time and frequency standard. In 1939 they discussed these possibilities with some scientists at the Bureau of Standards—whose names are fortunately no longer remembered—and found little interest there in the use of subtle molecular beam technique for such practicaal purposes as standards of magnetic field, time, or frequency.

In most respects the molecular beam technique in 1939 was more suitable as a standard of magnetic field than of frequency or time since the observed resonances at that time were largely dependent on the externally applied magnetic field. From the point of view of frequency control, it was consequently a great step forward when in 1940 Kusch, Millman, and Rabi [15, 16] first extended the method to paramagnetic atoms and in particular to $\Delta F = \pm 1$ transitions of atoms where the relative orientation of the nuclear and electronic magnetic moments were changed, in which case the resonance frequencies were determined dominantly by fixed internal properties of the atom rather than by interactions with an externally applied magnetic field. The first resonance measurements of the Cs hyperfine separation, which has been so extensively used in frequency control, were reported [16] in 1940.

In 1941 the research with the atomic beam magnetic resonance method was mostly interrupted by World War II and did not resume until 1946. In 1949 Kusch and Taub [17], in research supported in part by the Office of Naval Research (ONR), pointed out the possibility of observing the hyperfine resonances at magnetic fields such that the resonance frequency was an extremum, in which case the frequency to first order was independent of the strength of the magnetic field.

In 1949 Ramsey [18, 19] invented the separated oscillatory field method for a molecular beam resonance experiment on molecular hydrogen, which was supported by the Office of Naval Research. In this new method the oscillatory field, instead of being distributed uniformly throughout the transition region, was concentrated in two coherently driven oscillatory fields in short regions at the beginning and end of the transition region. The theoretical shape of a resonance curve with this apparatus is shown in Figure 2.

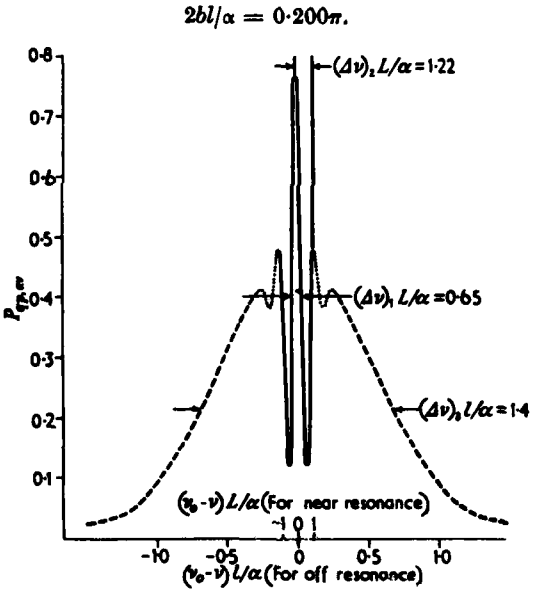


Figure 2—Theoretical shape for separated oscillatory field resonance pattern [18]

Ramsey pointed out that this method has the following advantages: (1) the resonances are 40% narrower than even the most favorable Rabi resonances with the same length of apparatus; (2) the resonances are not broadened by field inhomogeneities; (3) the length of the transition region can be much longer than the wavelength of the radiation, provided that the two oscillatory re-

gions are short, whereas there are difficulties with the Rabi method due to phase shifts when the length of the oscillatory region is comparable to the wavelength; (4) the first-order doppler shift can mostly be eliminated when sufficiently short oscillatory field regions are used; (5) the sensitivity of resonance measurements can be increased by the deliberate use of appropriate relative phase shifts between the two oscillatory fields. All of these characteristics are of great value for atomic beam resonance devices used as precision frequency and time standards. An early molecular beam apparatus [20] using this method is shown in Figure 3.

ATOMIC BEAM FREQUENCY STANDARDS

With the above developments, it was apparent to most molecular beam researchers by 1949 that atomic beam methods could be highly effective for precision frequency control. However, this was less clear to others who believed that crystal frequency control techniques had advanced so far that atomic devices could not be enough better to justify the extra cost and effort. However, in 1952, Sherwood, Lyons, McCracken, and Kusch [21,22] reported briefly on atomic beam resonance research supported by the National Bureau

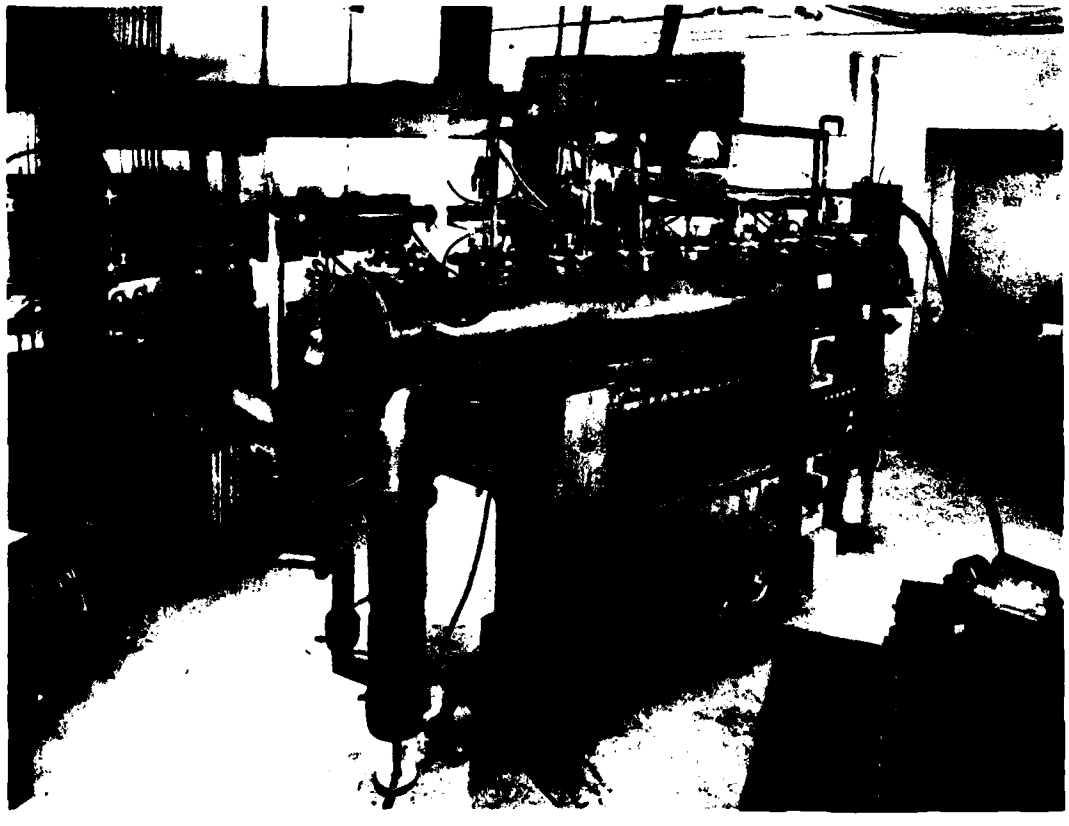


Figure 3—Apparatus for which separated oscillatory field was first proposed

of Standards and directed primarily toward the development of an atomic beam clock. A schematic diagram of a proposed atomic beam clock at that time is given in Figure 4. The financial support for such work soon dwindled due to advances in the then new field of microwave spectroscopy and to the view then held at the National Bureau of Standards that a molecular clock based on the microwave absorption by ammonia at its inversion frequency would be simpler and more promising.

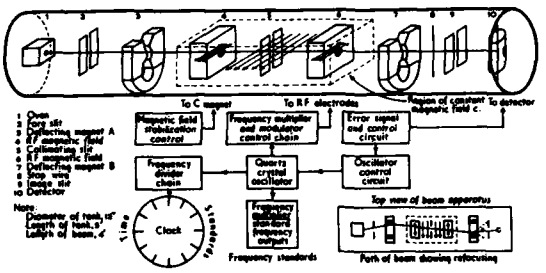


Figure 4—Schematic diagram of a proposed atomic beam clock [22]

A few years later, in work supported in part by the ONR, Zacharias [23,24] stimulated renewed interest in an atomic beam cesium clock. His initial concern was for an entirely new type of cesium beam in which ultrahigh precision would be obtained by the use of extremely slow molecules moving upwards in a vertical apparatus at such low velocities that they would fall back down by the action of gravity. Although this fountain experiment eventually failed due to the unexpected deficiency of the required ultraslow molecules emerging from the source, it stimulated Zacharias to develop and to urge others to develop well-engineered atomic beam frequency standards using normal atomic velocities. The unsuccessful fountain experiment of Zacharias illustrates the value to science even of some unsuccessful experiments; the existence of this unsuccessful effort directly and indirectly stimulated three quite different but important developments: (1) the use of conventional but well-engineered atomic beams for frequency control; (2) the development by Kleppner, Ramsey, and others [27–27] of the stored-atom technique, which eventually led to the hydrogen maser; and (3) high-precision resonance experiments with ultraslow neutrons [26]. The first report on an atomic beam frequency standard was that of Zacharias at the 1955 ninth Frequency Control Symposium. Zacharias claimed a short-time stability of 1 part in 10^9 for his atomic cesium frequency standard.

In 1955 Essen and Parry [28] of the British National Physical Laboratory successfully operated the first practical laboratory atomic cesium beam apparatus that was extensively used as an actual frequency standard. Their construction and effective use of this device provided a major impetus to the subsequent development of atomic beam cesium frequency standards.

In 1956 the first commercial model of an atomic beam frequency standard appeared on the market. This was National's Atomichron developed [29] by Holloway and Orenberg in collaboration with Zacharias and further improved by McCoubrey and Daley. This device used Ramsey's separated oscillatory field method for increased precision, a special design of cesium oven that could be operated several years without exhaustion, titanium pumping to permit permanent sealing off of the evacuated beam tube, and many other features generally necessary for an effective commercial device. The first commercial Atomichron is shown in Figure 5. The development of the Atomichron was supported financially largely by the U.S. Signal Corps at Ft. Monmouth, N.J., and the Office of Naval Research, although some support came from the Air Force. A purchase order by the Signal Corps for a relatively large number of Atomichrons made possible the development of mass-production techniques and improved engineering to permit sufficient reliability and reductions in price to assure commercial success.

The early atomic beam frequency standards were subject to various frequency shifts dependent on the amplitude of the radio-frequency power used and on other variables. To account for these results, Ramsey, with the aid of computational analysis supported by the Office of Naval Research and by the National Company, investigated the various possible distortions that would occur in an atomic beam resonance [30]. The elimination of radio-frequency phase shifts and other sources of distortion made possible the

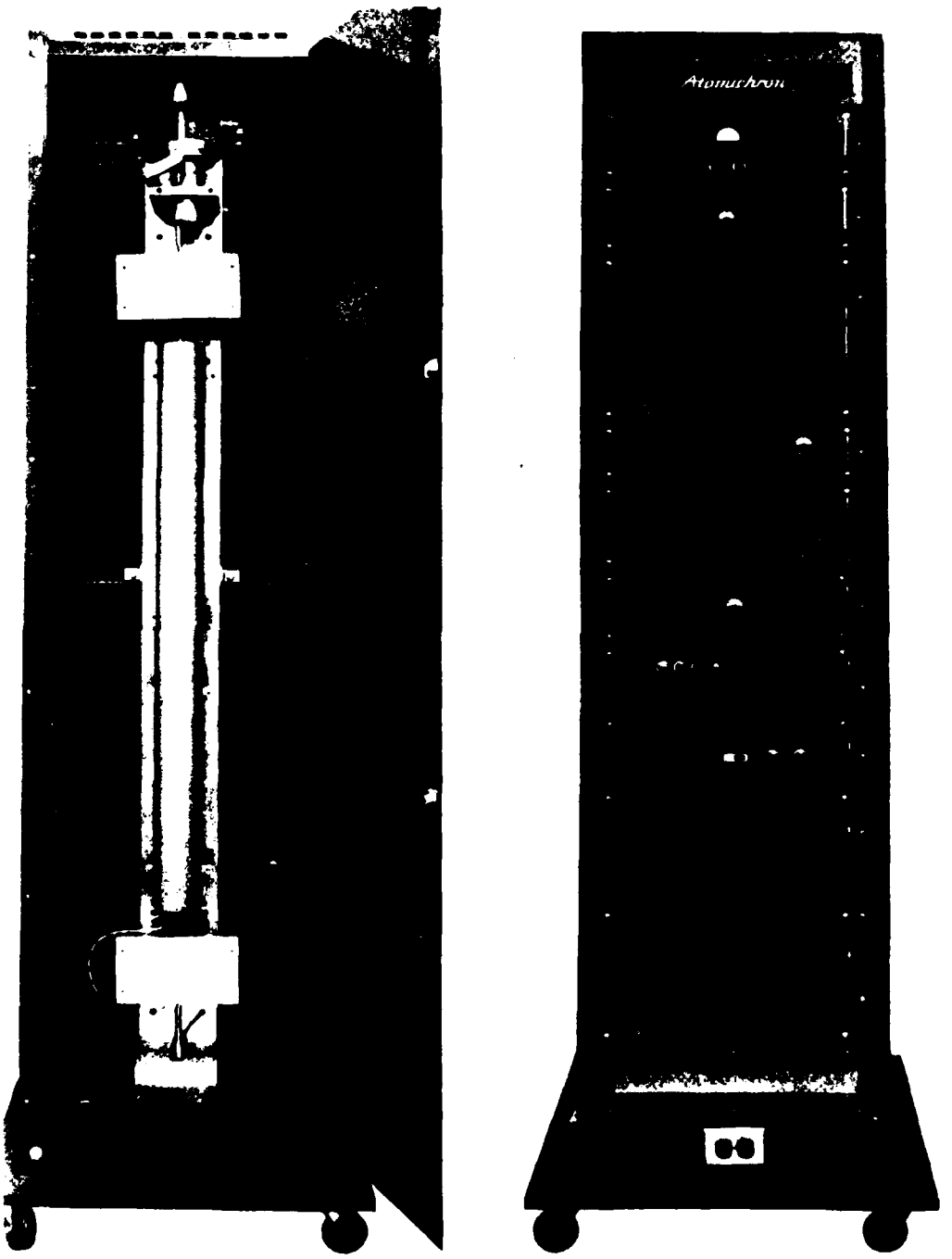


Figure 5—First commercial atomic beam frequency standard. National's Atomichron [29]

marked increases in accuracy that have been obtained with the atomic beam frequency standards.

From 1956 on, the atomic beam frequency standards developed rapidly. Mockler, Beehler, and Barnes [29,31] developed an atomic cesium frequency standard at the National Bureau of Standards in Boulder, Colo. Other commercial organizations such as TRG, Bomac, Varian, and Hewlett-Packard became involved. Many laboratories outside both England and the United States either constructed or purchased atomic beam frequency standards including those in Canada, France, and Germany and the laboratories of Kartaschoff [31] and Bonanomi [31] in Switzerland, Reder, Winkler, and others [31] at Ft. Monmouth and Markowitz at the Naval Observatory sponsored various worldwide studies of the comparison of atomic clock frequencies and the synchronization of clocks. Extensive studies were made of other atoms such as thallium for use in the atomic beam tubes, and various molecular resonances were studied for possible use in a molecular beam electric resonance apparatus for frequency control purposes. A Tl^{205} frequency measurement accurate to 2 parts in 10^{11} was reported by Bonanomi [32]. However, atomic cesium remains the most widely used substance in molecular or atomic beam frequency control devices. Particularly effective atomic beam cesium clocks were developed and sold by Hewlett-Packard, which also developed a "flying clock" particularly suitable for the intercomparison of atomic clocks in different laboratories. A typical beam tube for an atomic cesium frequency standard is shown in Figure 6. Accuracies as high as 1 part in 10^{13} have been claimed for some laboratory cesium standards [31].

In 1967, the 13th General Conference of Weights and Measures resolved that the unit of time in the International System of Units should be the second defined as follows: "The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium atom 133," a definition that is still retained.

MICROWAVE ABSORPTION SPECTROSCOPY

Microwave absorption spectroscopy had an early start in the experiments of Cleeton and Williams [33,34] in 1934. They observed the absorp-

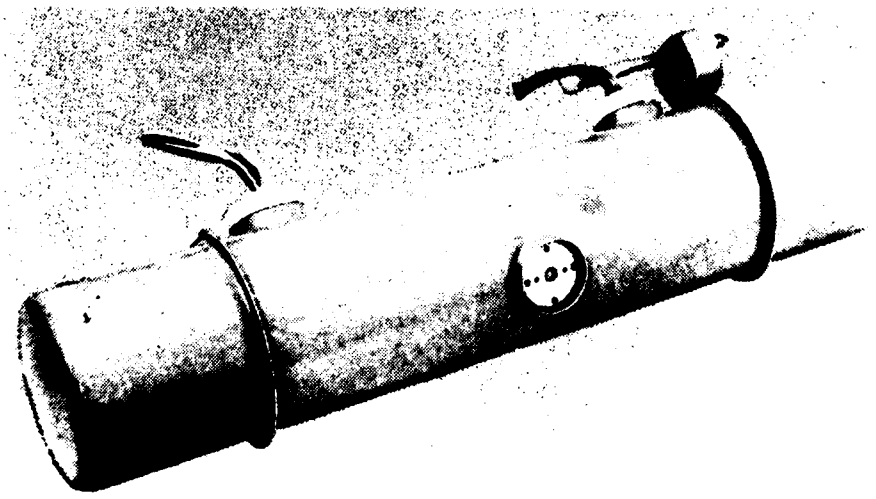


Figure 6—Beam tube for atomic cesium standard manufactured by Varian Associates for Hewlett-Packard [29]

tion of microwave radiation at the NH_3 inversion frequency. However, research on microwave absorption was inhibited at that time by the lack of suitable microwave oscillators and circuits so there was no further development of microwave absorption spectroscopy until after the development of microwave oscillators and waveguides for radar components in World War II. Immediately following World War II there was a great burst of activity in microwave absorption spectroscopy. Although there were no publications on experimental microwave spectroscopy in 1945, in the single year of 1946 there were a number of important publications from many different laboratories including reports by the following authors [35]: Bleaney, Penrose, Beringer, Townes, Dicke, Strandberg, Dailey, Kyhl, Van Vleck, Wilson, Dakin, Good, Coles, Hershberger, Lamont, Watson, Roberts, Beers, Hill, Merritt, and Walter, and in 1947 there were more than 60 published papers on this subject including a number of publications by Gordy and Jen, those with reports the previous year, and others. A typical microwave absorption experiment at this time is shown schematically in Figure 7.

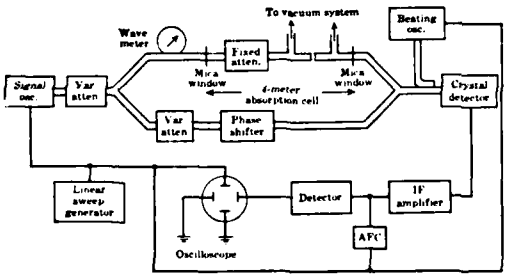


Figure 7—A typical microwave absorption experiment using a radio-frequency bridge and heterodyne detection

Microwave absorption techniques were quickly recognized to be of potential value for frequency standards. In 1948 a group of workers [22] at the National Bureau of Standards built an ammonia clock that was completed in 1949 and is shown in Figure 8, and it eventually achieved an accuracy of 1 part in 10^8. Rossell [22] in Switzerland and Shimoda in Japan devised an improved ammonia absorption clock good to a few parts in 10^9.

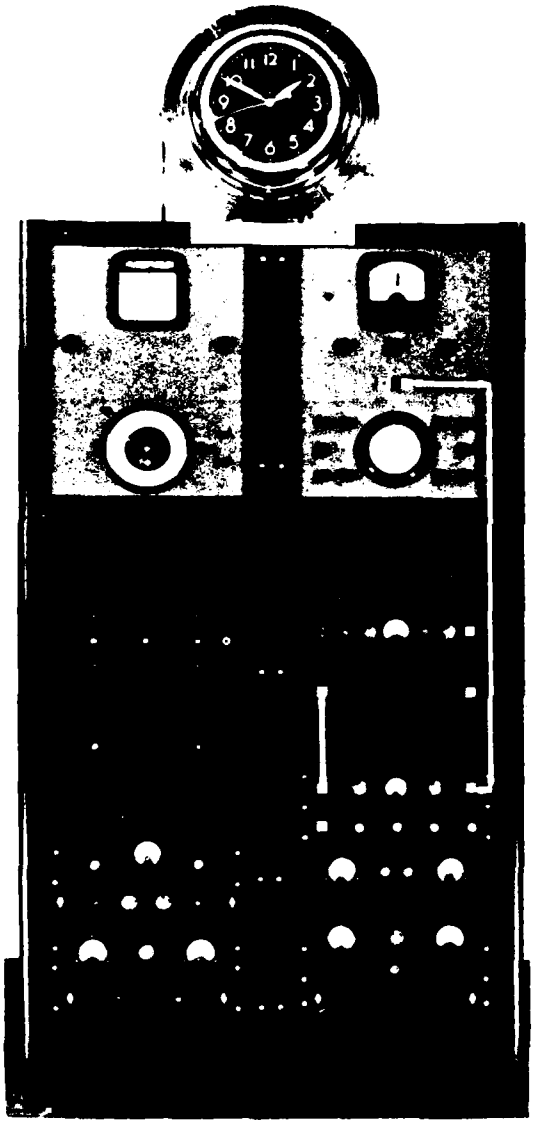


Figure 8—National Bureau of Standards ammonia clock [22]

The first report pertaining to microwave atomic and molecular frequency standards was that of Dicke [31,36] at the 1951 fifth Frequency Control Symposium. In the seventh, eighth, and ninth symposia he, Carver, Arditi, and others described the continuation of this work at both Princeton and the Radio Corporation of America (RCA) with the financial support of the Signal Corps and the Office of Naval Research [31,36].

The microwave absorption studies soon merged with the optical pumping techniques described in the next section, since the intensities of the resonances were greatly enhanced by the use of optical pumping.

OPTICAL PUMPING

The starting point of all research on optical pumping was a paper by Bitter [37] in 1949, which showed the possibility of studying nuclear properties in optically excited states. Kastler [38,39] showed the following year that this technique could be effectively combined with the double resonance method he and Brossel [38] had developed. Both optical pumping and optical detection techniques served the purpose of increasing the signal-to-noise ratio of the resonator output signal: the optical pumping greatly enhances the population of certain states so the signal is not weakened by stimulated emission nearly canceling absorption, and the optical detection increases the signal-to-noise ratio because of the lower noise level of optical detectors over microwave detectors.

The combination of optical pumping techniques with the buffer gas method for reducing doppler shift developed by Dicke [29,36] provided gas cells of real value as frequency-control devices. Although many different atoms have been used in such gas cells, Rb^{87} soon became the favorite in most such devices. Extensive work in optically pumped gas cells for frequency control has been done at Princeton, RCA, International Telephone & Telegraph (ITT), Space Technology Laboratory, the National Bureau of Standards, Clauser Technology Corporation, Varian Associates, and many other commercial, university, and government organizations in the United States and abroad. Figure 9 shows a typical optically pumped rubidium frequency standard.

The optically pumped gas cells have the advantages of simplicity, relatively low cost, large

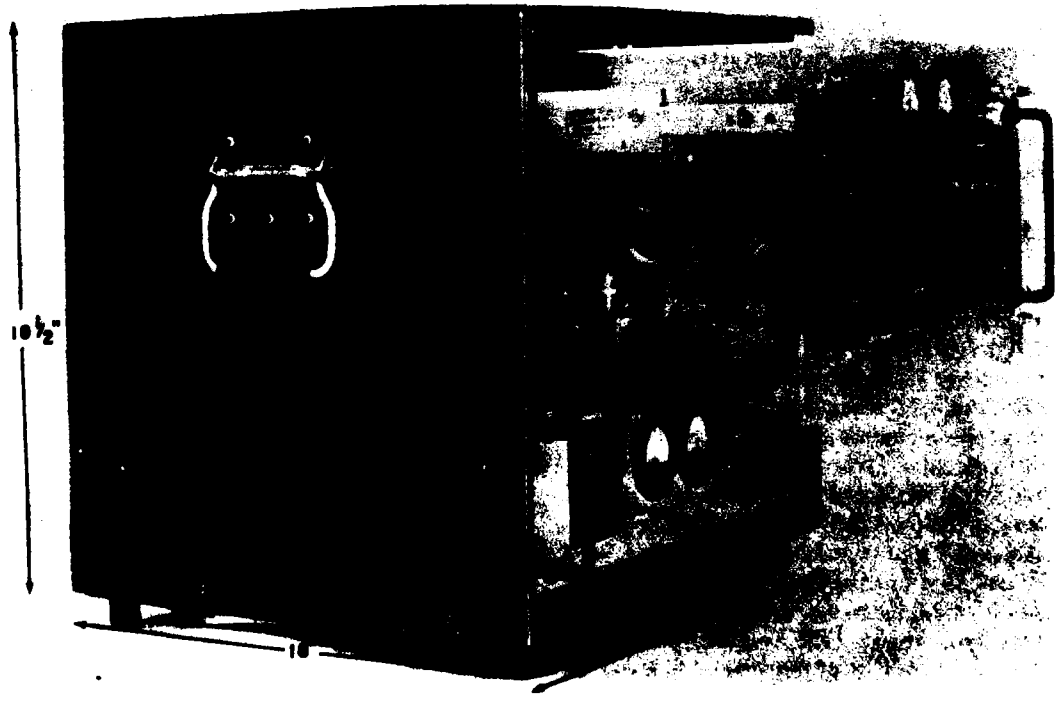


Figure 9—Rubidium frequency standard

signal-to-noise ratio, and good spectral purity. Unfortunately the relatively large shift in frequency due to numerous buffer gas collisions is dependent on purity, pressure, and temperature. Changes in the light intensity shift due to variations in the pumping lamp intensity or spectrum may also be a problem. As a result, the stability of rubidium gas cells over a period of several months is ordinarily no better than a few parts in 10^{10}. These pressure shifts prevent the optically pumped gas cells from being primary frequency standards, but the gas cells are used as frequency control devices when too much accuracy is not required. Research is currently in progress in a number of laboratories to improve the stability of optically pumped gas cells; Bouchiat, Brossel [40], and others, for example, have eliminated the buffer gases and, as in the hydrogen maser, have used collisions with suitable coated walls to retain the atoms and reduce the effect of the first-order doppler shift.

MOLECULAR MASERS

In 1951 Pound et al. [41], in experiments supported by the Office of Naval Research, studied nuclear spin systems with inverted populations and noted that such systems in principle were intrinsic amplifiers rather than absorbers. The first suggestions actually to use systems with inverted populations as practical amplifiers and oscillators were made at closely the same time in 1953–1955 and indepe ıdently by Townes [42], Weber [43], and Basov and Prokhorov [44]. The first such amplifier was successfully constructed in 1955 by Gordon, Zeiger, and Townes [42] and called a maser (Microwave Amplifier by Stimulated Emission of Radiation). The device used inhomogeneous electric fields to focus the higher energy molecular inversion states of ammonia molecules in a molecular beam. These molecules then emitted coherent stimulated radiation in passing through a cavity tuned to the 24-GHz ammonia inversion transition. A schematic diagram of the first ammonia maser is shown in Figure 10. A report by Gordon on the new ammonia maser was a major attraction at the special meeting on atomic and molecular resonances sponsored by the Signal Corps Engineering Laboratory in 1956.

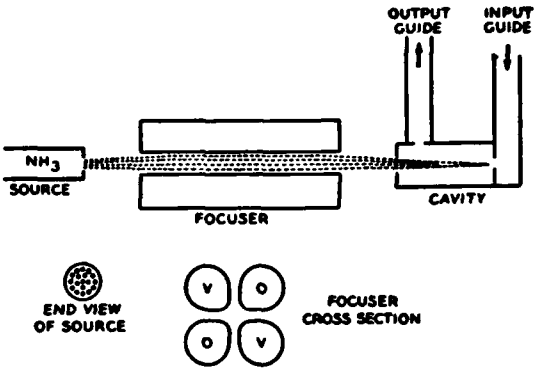


Figure 10—Schematic diagram of original ammonia maser [42]

In that year Bloembergen [45] proposed the three-level solid-state maser and in 1958 Townes and Schawlow [46] pointed out the possibility of masers at the infrared and optical frequencies.

Since the announcement of the first successful ammonia maser in 1955 there has been tremendous research and development activity by scientists and engineers in many countries. Masers at infrared or optical frequencies (lasers) have great potential for frequency control. Further discussion of lasers will be deferred to a later section of this report. Molecular maser developments for the purposes of frequency control soon became intense and went in many directions including the search for more suitable molecules than ammonia, the development of two cavity masers analogous to the separated oscillatory field method [18] for molecular beams, use of ammonia of different isotopic composition, and so forth. A value of the $N^{15}H_3$ frequency accurate to 5 parts in 10^{11} has been obtained by de Prins and confirmed by Barnes [32]. However, after a few years of intense molecular maser activity, the interest in such masers for frequency control waned since the molecular masers on the one hand lacked the simplicity and low cost of optically pumped rubidium gas cells and on the other hand lacked the high precision of either atomic cesium beams or atomic hydrogen masers.

ATOMIC MASERS

In 1957 Ramsey [31] proposed to increase the accuracy of the atomic beam magnetic resonance

method by retaining the atoms for a much longer time between the two separated oscillatory fields, thereby obtaining much narrower resonances. His first thought was to confine the atoms with inhomogeneous magnetic fields in a large ring. However, it soon became apparent that the inhomogeneous confining magnetic fields, which acted on the atoms for long periods of time, would hopelessly broaden the resonances. In fact, it became clear that the frequencies would be much less perturbed by a confinement force that was present for only a short fraction of the time even though the force might be stronger when it was applied. The obvious limit of such a device was confinement of atoms in a box with suitably coated walls. Although many wall bounces would be required to achieve marked narrowing of the resonance by long storage time, the first experiments involved only a few wall collisions, since most scientists at that time believed that even atoms in an S state would undergo hyperfine transitions at even a single wall collision. The first experiments of Kleppner et al. [25] involved only a few wall collisions and the experiment was appropriately called a "broken atomic beam resonance experiment." Cesium atoms and a Teflon-coated wall were used in these first experiments.

Goldenberg, Kleppner, and Ramsey [27] then made an atomic beam resonance apparatus that stored atoms of cesium for a longer time and they investigated alternate wall-coating material in experiments supported by the Office of Naval Research and the National Science Foundation. They found that when the storage bulb was coated with a paraffin-like substance called Paraflint [27], resonances could be observed after as many as 200 wall collisions. It was recognized that atomic hydrogen would probably be a more suitable atom than cesium because of the low electric polarizability and the low mass of hydrogen, but cesium could be much more efficiently detected than hydrogen.

Kleppner and Ramsey [25, 27, 47] therefore proposed detection of the emitted radiation rather than of the atom. In particular, they noted that atoms of hydrogen in the higher energy hyperfine state could be focused into a suitably coated storage bulb by a six-pole magnet while atoms in the lower state would be defocused. They showed that if such a storage bulb were surrounded by a microwave cavity tuned to the 1420 MHz hyperfine transition frequency, then maser oscillation should occur. In 1960, Goldenberg, Kleppner, and Ramsey [47] constructed and operated the first atomic hydrogen maser. This apparatus is shown in Figure 11. Although the total microwave power was small—approximately 10^{-12} W—the stability was so high that the output was concentrated into an extremely narrow band with a consequently favorable signal-to-noise ratio.

Although the first hydrogen masers used wall coatings of Paraflint or of Dri-Film (dimethyldichlorosilane [27]), it was soon found that with atomic hydrogen, in contrast to cesium, Teflon-coated walls gave longer storage times and smaller frequency shifts from wall collisions [48]. Bender [49] soon pointed out that spin exchange collisions of hydrogen atoms could not be neglected and might produce a significant frequency shift, but Crampton [50] noted that the normal tuning technique would cancel out such an effect. Later Crampton [51] pointed out the existence of a smaller additional spin exchange effect that would not be canceled by the normal tuning method. This effect was omitted in earlier theories due to their neglect of the contribution of the hyperfine interaction during the time of the short duration of the collision. Crampton [51] developed a technique for measuring the spin exchange effect. Crampton also pointed out the existence of a small frequency shift [51] due to magnetic field inhomogeneities; this small shift is often called the Crampton effect. Both of these effects are so small they did not affect past measurements and they can be further reduced by suitable apparatus design.

A commercial hydrogen maser [52] was developed by Vessot, Peters, Vanier, McCoubrey, Levine, and Cutler. The work was started at Bomac and successively transferred to Varian Associates and Hewlett-Packard. It has also been carried on at the Smithsonian Astrophysical Observatory by Vessot and his associates; at the Goddard Space Flight Center by Peters, Reinhardt, and others; and at the Jet Propulsion Laboratory. The H-10 maser developed by Vessot and his associates is shown in Figures 12 and 13. The masers are being built chiefly for long baseline interferometry in radio astronomy, which benefits greatly from the high stability of

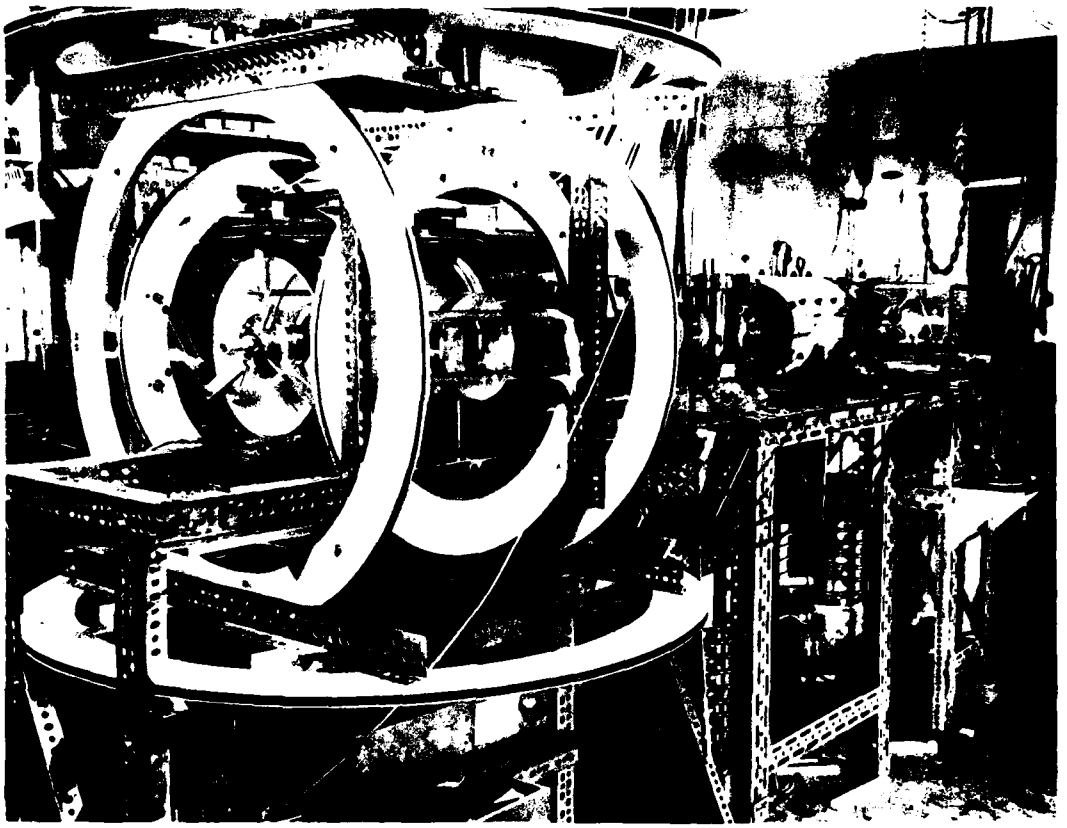


Figure 11—Original hydrogen maser [47]. The large coils are to cancel external magnetic fields. In later hydrogen masers, these were replaced by two or three concentric cylinders of high-permeability magnetic shielding.

the hydrogen maser. Vessot [53] has recently flown a hydrogen maser in a high-altitude rocket to test the gravitational red shift. Research and development on hydrogen masers has also been carried out in the laboratories of Vanier in Canada; Kartaschoff [31] in Switzerland; Audoin [54] and Grivet [31] in France; Crampton [51], Kleppner [52], Wang [51], Hellwig [55], Ramsey [56], and others [57] in the United States; and in a number of other countries. Audoin and his associates [54] introduced a useful double-focusing technique that eliminates undesired atoms from the focused beam.

The hydrogen maser eliminates first-order doppler shifts and photon recoil effects by virtue of the confinement of the atoms in a box where the *average* velocity is essentially zero and by absorption of recoil momentum by the confining box. The hydrogen maser also benefits from the relatively long storage time with the resulting narrow beam and from the low noise characteristic of maser amplification. It shares with most other atomic or molecular frequency standards the need for correcting for the small second-order doppler shift.

The chief disadvantage of the hydrogen maser for time and frequency control has been the existence of a small frequency shift due to collisions of the atoms with the Teflon-coated walls of the storage bulb. With a 16-cm diameter bulb this wall shift is about 2 parts in 10^{11} and can be measured by using bulbs of two different diameters. How-

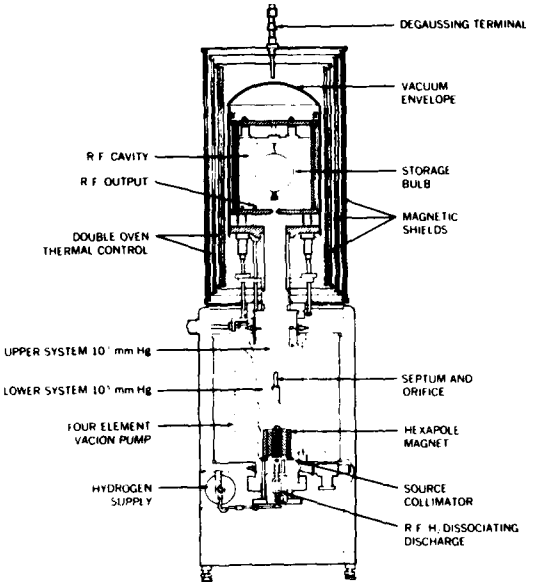


Figure 12—Schematic diagram of a commercial hydrogen master developed by Vessot and his associates [52]

ever, until recently the measurements of the wall shifts have been limited to accuracies of a few percent by variations in different wall coatings. However, Uzgiris and Ramsey [56] at Harvard have reduced the wall shift by a factor of 10 by the use of an atom storage vessel 10 times larger in diameter (1.5 m). In the same laboratory, Brenner [58] and Debely [59] have developed a technique to measure the wall shift in a single storage bulb: they were able to change the bulb's volume by deforming its shape. Since a single bulb is used in this method, it is free from the uncertainties in the nonreproducibility of the wall coatings of different bulbs. Although this method was first used on hydrogen masers with normal-size storage bulbs, Reinhardt [60] has applied it to the large storage bulbs as well. Zitzewitz [61] has shown that a temperature of about 80°C the wall shift passes through zero; it is thus possible to operate the hydrogen maser at a temperature such that the wall shift vanishes and to select this temperature by the deformable bulb technique. With these new methods, absolute accuracy better than 1 part in 10^{13} should be attained.

Although the hydrogen maser is the most stable atomic maser over long periods of time, Novick,

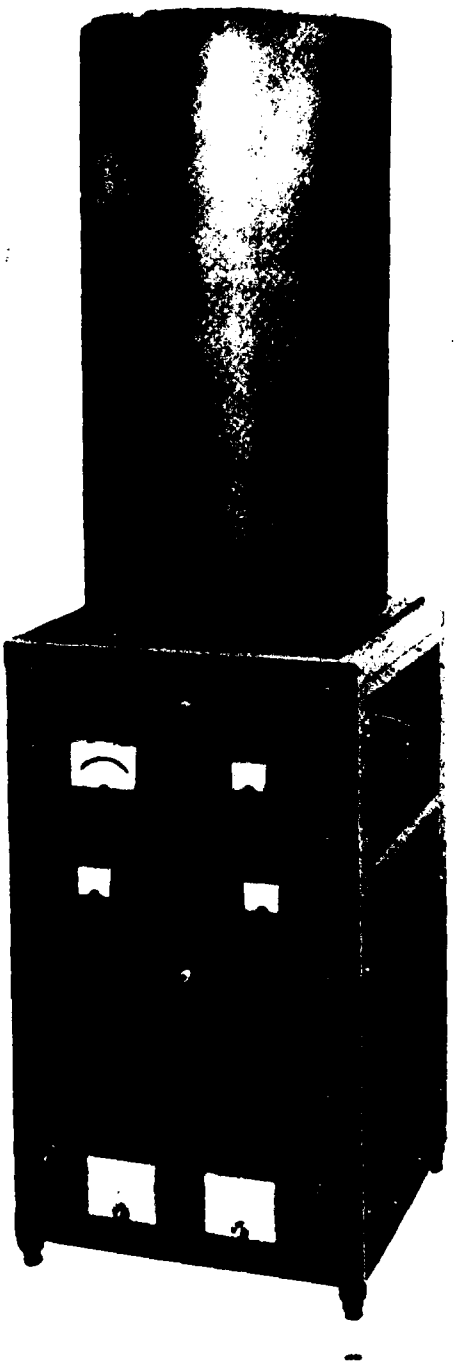


Figure 13—Commercial hydrogen maser [52]

Vanier, and others [31] have developed a high-power optically pumped atomic Rb^{85} maser whose relatively high output power is useful for short-term stability.

LASERS

Townes and Schawlow [46] pointed out that masers could be produced at infrared and optical frequencies. The first optical maser or laser was successfully made from ruby by Maiman [62]. Subsequently there was a great burst of activity in this field and lasers were made of a wide variety of materials and at high pulsed power. From the point of view of frequency control the laser using a helium-neon gas mixture developed by Javan [63] and his associates was the first one of interest as a time standard because of its potential stability.

As absolute time standards most lasers suffer from the fact that the output frequency is primarily determined by the distance between two mirrors since the first-order doppler broadening of the atomic or molecular resonance exceeds the resonance width of the interferometer. This characteristic contrasts with a microwave maser where the frequency is determined primarily by the atomic transition with only a relatively small amount of pulling from mistuning of the microwave cavity. However, various methods from diminishing the first-order doppler spread and thereby for determining the laser frequency more by atomic or molecular properties have been developed. These methods usually depend upon nonlinear effects.

One method that has been particularly effective is laser-saturated absorption spectroscopy developed by J. L. Hall [64, 65], Schawlow [66], Hansch [66], and others [67, 68]. In such a device, laser light is passed in opposite directions through, say, a CH_4 or I_2 absorption cell. There is a minimum of absorption at a frequency corresponding to no first-order doppler shift, since stationary molecules absorbing at that frequency absorb the light from both directions equally well and hence are more readily saturated than are the moving molecules which respond at most to light from a single direction. A schematic view of a laser-saturated absoprtion device is shown in Figure 14. An absorption cell containing methane

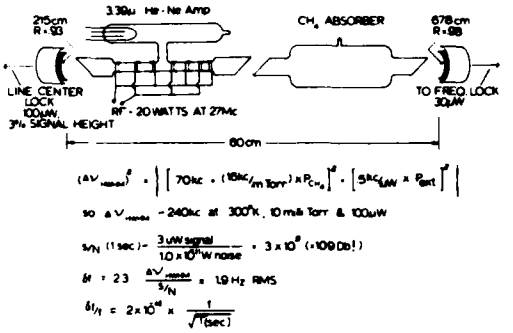


Figure 14—Laser-saturated methane frequency reference [64]

is included in the optical path between the two mirrors of a helium-neon laser. Since approximately equal amounts of laser light are going in each of two directions between the laser mirrors, the methane molecules are subjected to the two opposite beams of light. For a moving methane molecule, the frequency of the two beams will appear to be slightly different due to first-order doppler shifts. However, for those few methane molecules that are not moving significantly along the direction of light propagation, there will be no doppler shift so the two beams will appear to be at the same frequency. Such molecules are consequently subjected to double the intensity of resonant radiation and their ability to absorb radiation is more quickly saturated with a corresponding loss in absorptive power. The laser will oscillate at the frequency of least absorption, namely that of the molecules with no component of velocity along the direction of the laser beams so the first-order doppler broadening is eliminated. Stabilities of a few parts in 10^{14} have been achieved by Hall and others [64] with 3.39 μm He-Ne laser-saturated absorption in CH_4 but the reproducibility is only about 1 in 10^{11}. Although this technique markedly reduces first-order doppler broadening, it does not automatically remove all shifts associated with molecular recoil. Also power shifts and second-order doppler effects remain. A combination of saturated absorption with an atomic beam of calcium has recently given encouraging results [69].

Double-resonance [68, 70, 71] and two-photon doppler-free absorption spectroscopy [72] eliminate first-order doppler broadening by requiring

the absorption of two photons. If these photons come from opposite directions and are at different frequencies appropriate to an intermediate real energy level, the different first-order doppler shifts would prevent simultaneous absorption of both photons except for the absorption by molecules moving with approximately zero velocity along the direction of the laser beam, since for these molecules the first-order doppler shifts are approximately zero. Two-photon doppler-free absorption spectroscopy is particularly effective when the two photons moving in opposite directions are at the same frequency, even though in this case the intermediate transition is to a virtual level since it is unlikely that a real level will fall exactly halfway between the initial and final states. Since the doppler shift in one direction is equal and opposite to that in the opposite direction, the sum of the two frequencies is independent of the molecular velocity, so molecules at all velocities can contribute to the two-photon doppler-free spectrum. Since the two photons move in opposite directions with equal momentum, there is no recoil of the molecule and hence no doppler or recoil broadening. High laser power levels, however, may be required so power shifts may be a problem, but they can be reduced with a suitable experimental arrangement. In common with most other methods, the second-order doppler shift is not eliminated in two-photon spectroscopy.

A major advance in recent years has been the development of frequency multiplying techniques to the optical region by Javan [63] and others [73]. With these techniques it is possible to compare laser frequency standards with the cesium beam standards used in the definition of the second. With such devices, both the frequency and the wavelength of CH_4 (and CO_2) stability lasers have been measured and thereby a precision value for the velocity of light of $299\,792\,458.3 \times 1.2$ m/s has been obtained [64, 73, 74].

TRAPPED IONS

Dehmelt [75] in 1959 first used electromagnetic ion traps in radio-frequency resonance studies. The intrinsic width of the resonances as determined by the uncertainty principle can be very narrow since the ions are retained in the apparatus for very long periods of time. Dehmelt and his associates [76] have constructed a successful trapped-ion experiment for measuring g-2 of the electron with a single electron in the trap to avoid space charge; they have called this device a mono-electron oscillator. They have also proposed a barium or thallium mono-ion oscillator as a possible oscillator of high stability [76]. However, until recently these devices have suffered from the relatively high velocity of the ions in the trap (approximately of 1 e V of kinetic energy) with the correspondingly larger broadening due to the second-order doppler shift. Initial efforts by Dehmelt et al [77] to diminish this were only partially successful and trapped-ion devices have not as yet provided frequencies as stable as those of the best alternative frequency standards. However, as discussed below in the section on doppler broadening, Wineland and Dehmelt [76] have recently proposed an ingenious technique for resonant radiation cooling of trapped ions. If this technique is fully successful, trapped-ion resonance devices should become highly promising frequency standards, although their stability is degraded by the low signal-to-noise ratio which results from the space charge limitation on the number of ions that can be studied simultaneously. Ion recoil ordinarily causes no difficulty since the recoil momentum is absorbed by the trapping field.

SUPERCONDUCTING CAVITIES

High-stability superconducting cavity oscillators have recently been made by Stein and others [78] at Stanford University with the support of the Office of Naval Research. Although such oscillators do not strictly come within the scope of this report, their stability, especially for short times, is so great that they should be discussed here at least briefly even though they are not suitable as absolute standards since the frequency depends on cavity dimensions instead of a characteristic atomic or molecular transition frequency. A schematic view of such a superconducting cavity is shown in Figure 15. Stabilities of the order of 10^{-15} have been attained with such oscillators as discussed in the next to the last sec-

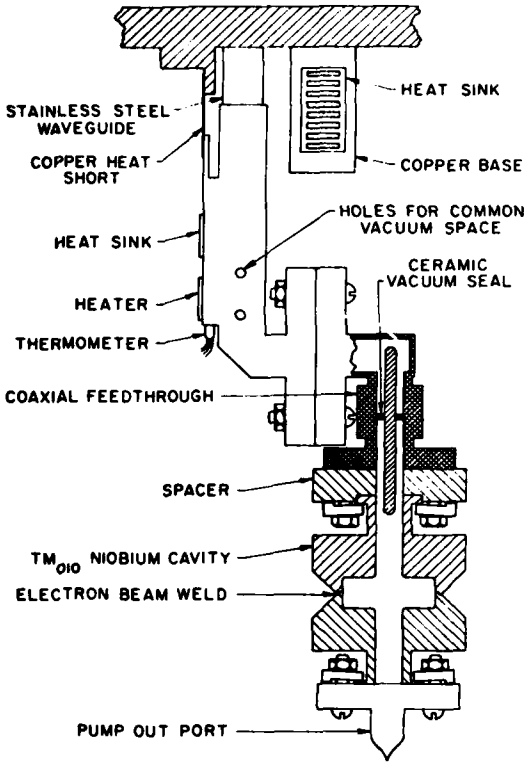


Figure 15—View of superconducting cavity showing the mounting [78]

tion. Since short-term stability increases with oscillator power and since superconducting cavities can be operated at a relatively high power level, they have particularly favorable short-term stability which makes them particularly useful in providing a fundamental frequency that is highly multiplied to reach the laser range.

DOPPLER BROADENING

The atoms or molecules in atomic and molecular frequency standards are in thermal motion and hence subject to both first- and second-order doppler shifts or broadening. The first-order doppler shift—the familiar increase in the frequency received from an approaching radiation source—is proportional to $v/c \approx 3 \times 10^{-7}$ so any competitive frequency standard must provide a means for eliminating the first-order doppler shift. Consequently, this essential feature of the different frequency standards can most simply be given by describing the way that each one eliminates both frequency shifts and resonance broadening from first-order doppler shifts.

In the cesium and molecular beam devices the first-order doppler shift is eliminated by the use of two separated oscillatory fields of coherent radiation of the same phase. In the hydrogen maser, the first-order doppler shift is eliminated by confining the hydrogen atoms to a small volume which is traversed many times during the radiation process of each atom so the velocity averages to zero. In trapped-ion spectroscopy, the first-order doppler shift is eliminated for the same reason. With laser-saturated molecular absorption devices, double-resonance spectroscopy, and two-photon spectroscopy the first-order doppler shift is removed by the requirement of the absorption of two or more photons moving in opposite directions, as discussed in the sections on these devices.

However, even after the first-order doppler shift is eliminated, there remains in atomic and molecular oscillators a second-order doppler shift whose magnitude is of the order $(v/c)^2 = 10^{-13}$. If much progress is to occur beyond the present accuracy of a few parts in 10^{-13}, means must be found to reduce the magnitude of the second-order doppler shifts, i.e. to reduce the velocities. New possible techniques for reducing the magnitudes of the velocities have been proposed by Hansch and Schawlow [79] and by Wineland and Dehmelt [76], but the proposals are so far mostly untested. The basic idea is to cool, say, trapped ions by shining on them intense laser light at a frequency slightly below the resonance frequency. This light can be absorbed by an ion whose motion provides the appropriate first-order doppler shift. The subsequent emission, however, is in all directions and hence on the average at the normal resonance frequency. By conservation of energy the ion must therefore lose kinetic energy. In this fashion the trapped ions can be cooled by many successive absorptions and reemissions. It will be of great interest during the coming years to see if these techniques for reducing the second-order doppler shift are successful and to see if they lead to marked increases in the accuracy of clocks and frequency standards.

ACCURACY, REPRODUCIBILITY, AND STABILITY

In discussions of time and frequency standards it is necessary to distinguish between three different but related properties of the standards: accuracy, reproducibility, and stability. Accuracy measures the degree to which a standard independently agrees with the value specified in the definition of the unit of time. Reproducibility is a measure of the extent to which properly adjusted independent devices of the same design agree. Stability is a measure of the degree to which the same device gives the same result in successive intervals of time. The stability is conventionally measured by the parameter $\sigma_y(\tau)$ which is the square root of the two-sample Allan variance [80] for adjacent samples which in turn is one-half of the mean square of the fractional differences of the frequencies measured in adjacent intervals of time duration τ.

For different applications, different characteristics are the most relevant. Thus, for absolute standards of frequency, the accuracy is the most important property. On the other hand, for many measurements, such as long baseline interferometry in radio astronomy, stability is of primary concern.

The stability $\sigma_y(\tau)$ is plotted as a function of the time interval τ for a number of different oscillators in Figure 16.

The need for accurate timing has been recognized for many centuries and the development of better clocks has been vigorously pursued throughout that time. However, the truly spectacular advances in that field have occurred only

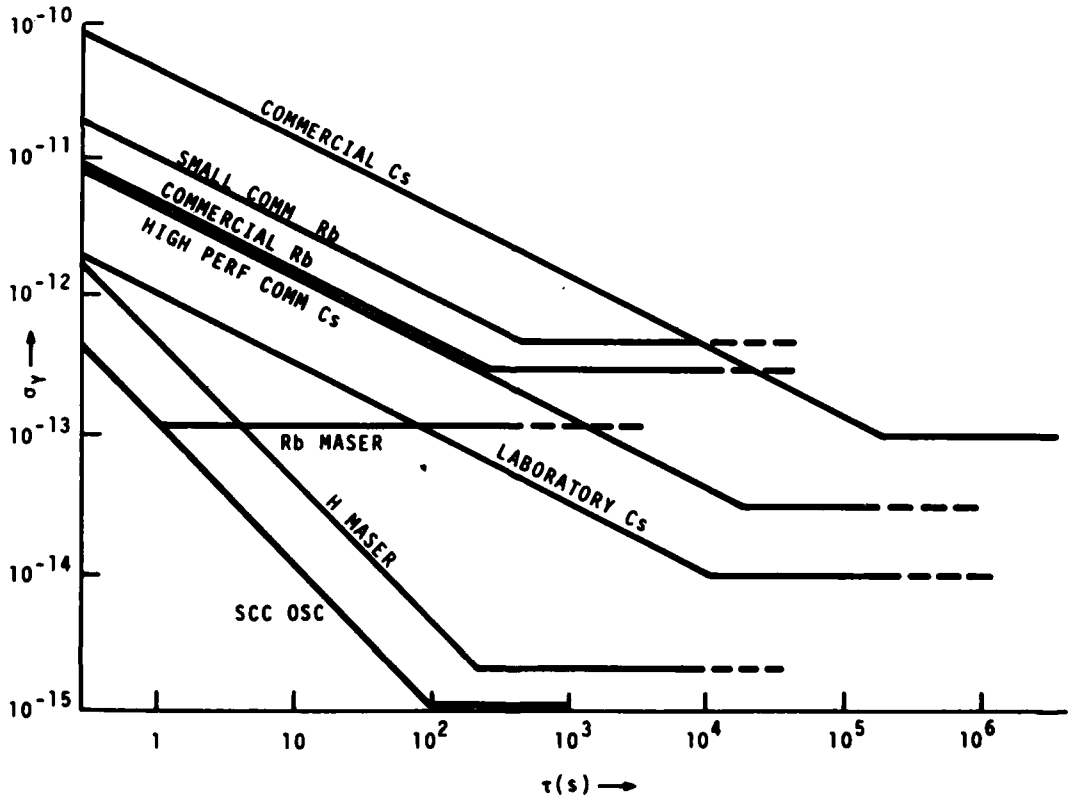


Figure 16—Stabilities of various frequency standards [55, 78]

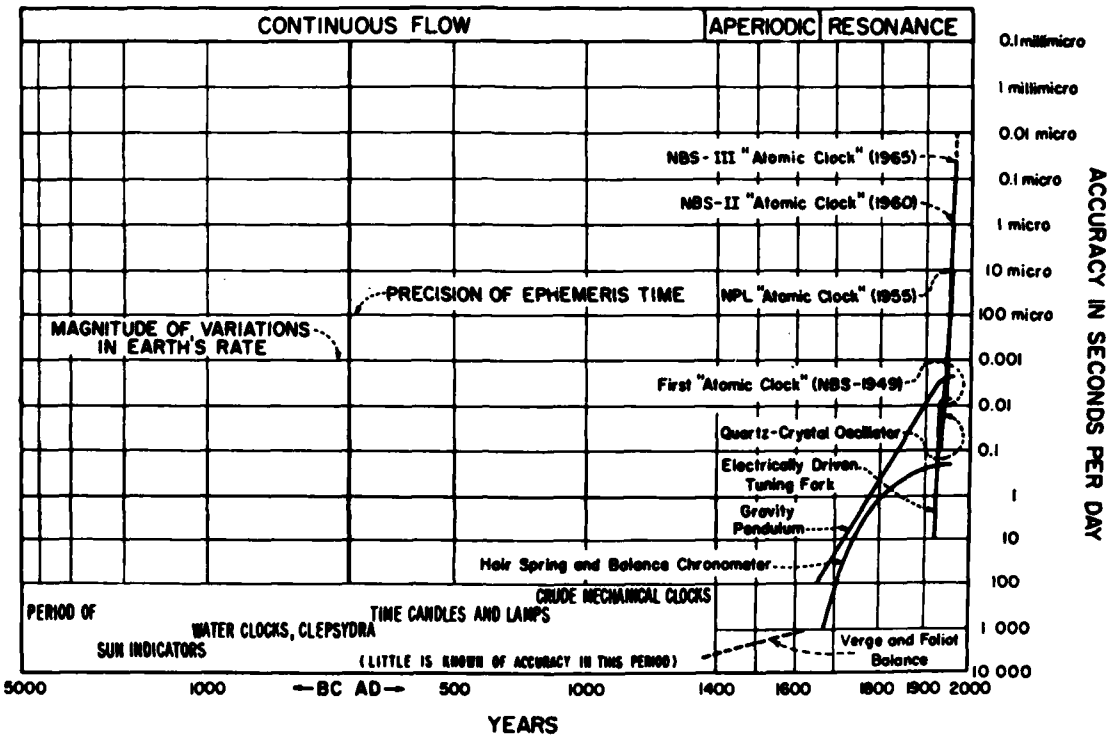


Figure 17—The accuracy of timing through history

in the past few decades, as is illustrated in Figure 17, which shows the development of the accuracy of timing through history.

FUTURE PROSPECTS

Although atomic and molecular frequency and time standards have been a reality for a number of years, new developments are occurring at a relatively rapid rate. As a result it is impossible to forecast reliably the future developments that will lead to the most major subsequent advances. However, a number of prospective developments for the different devices are included in the above discussions of these devices. For highest stability and reproducibility the most promising of these prospects may be summarized as the following: (1) Further improvements on the existing atomic and molecular beam methods such as the widely used cesium frequency standard, including better velocity definition to reduce uncertainties due to second-order doppler shifts and including possible new molecules with higher frequency resonances. (2) Hydrogen maser improvements including combinations of the deformable bulb technique with either the large box maser or operation at a temperature where the wall shift vanishes. The use of electronic cavity tuning should provide increased stability, and interesting studies have been undertaken at the National Bureau of Standards of the use of atomic hydrogen as a passively operating frequency standard [55]. (3) Improved stored-ion devices, especially if the newly proposed techniques [76, 79] for cooling the trapped ions work well and thereby markedly diminish the second-order doppler broadening. Development of mono-ion oscillators [69]. (4) The use of lasers especially when combined with nonlinear spectroscopy techniques which eliminate first-order doppler broadening, such as saturated molecular absorption and especially two-photon doppler-free spectroscopy. (5) The

reduction of the second-order doppler broadening for any of the methods by the resonance cooling techniques discussed earlier [79,79]. (6) Improvements in frequency multiplying techniques to connect the microwave and optical regions. (7) Improvements in superconducting cavity oscillators. (8) Combinations of various techniques such as saturated absorption laser spectroscopy with atomic beams or use of a superconducting cavity as a slave oscillator for an atomic resonance device.

However, if past precedents are followed, there will in addition be many unexpected new ideas and developments that drastically improve existing techniques or lead to totally new methods of atomic or molecular frequency control.

REFERENCES

1. C. Darwin, *Proc. Roy. Soc.* **117**, 258 (1927).
2. T. E. Phipps and O. Stern, *Z. Phys.* **73**, 185 (1931).
3. P. Güttinger, *Z. Phys.* **73**, 169 (1931).
4. E. Majorana, *Nuovo Cimento* **9**, 43 (1932).
5. R. O. Frisch and E. Segre, *Z. Phys.* **80**, 610 (1933).
6. I. I. Rabi, *Phys. Rev.* **49**, 324 (1936).
7. L. Motz and M. Rose, *Phys. Rve.* **50**, 348 (1936).
8. I. I. Rabi, *Phys. Rev.* **51**, 652 (1937).
9. J. Schwinger, *Phys, Rev.* **51**, 645 (1937).
10. I. I. Rabi, J. R. Zacharias, S. Millman, and P. Kusch, *Phys. Rev.* **53**, 318 (1938) and **55**, 526 (1939).
11. J. M. B. Kellog, I. I. Rabi, N. F. Ramsey, and J. R. Zacharias, *Phys. Rev.* **55**, 729 (1939); **56**, 728 (1939); and **57**, 677 (1940).
12. C. J. Gorter, *Physica* **3**, 503 and 995 (1936).
13. E. M. Purcell, H. G. Torrey, and R. V. Pound, *Phys. Rev.* **69**, 37 (1946).
14. F. Bloch, W. Hansen, and M. E. Packard, *Phys. Rev.* **69**, 127 (1946) and **70**, 474 (1946).
15. P. Kusch, S. Millman, and I. I. Rabi, *Phys. Rev.* **57**, 765 (1940).
16. S. Millman and P. Kusch, *Phys. Rev.* **57**, 438 (1940).
17. P. Kusch and H. Taub, *Phys. Rev.* **75**, 1477 (1949).
18. N. F. Ramsey, *Phys. Rev.* **76**, 966 (1949); *Molecular Beams*, New York; Oxford, 1956 (1969); and *IEEE Trans. on Instr. and Meas.* **IM-21**, 90 (1972).
19. N. F. Ramsey and H. B. Silsbee, *Phys. Rev.* **84**, 506 (1951).
20. H. G. Kolsky, T. E. Phipps, N. F. Ramsey, and H. B. Silsbee, *Phys. Rev.* **80**, 483 (1950).
21. J. E. Sherwood, H. Lyons, R. H. McCracken, and P. Kusch, *Bull. Amer. Phys. Soc.* **27**, no. 1, 43 (1952).
22. H. Lyons, *Ann. N.Y. Acad. Sci.* **55**, 831 (1952) and *Sci. Amer.* **196**, 71 (Feb. 1957).
23. J. R. Zacharias, private communication and *Phys. Rev.* **94**, 751 (1954). (R. Weiss and R. Vessot were associated with Zacharias in the experimental work on the "fountain" experiment.)
24. J. R. Zacharias, J. G. Yates, and R. D. Haun, M.I.T., Res. Lab. Electron., Cambridge, Mass., Quart. Prog. Rep. 30, Jan. 1955, and "An Atomic Frequency Standard," *Proc. IRE* (Abstract) **43**, 364 (Mar. 1955).
25. D. Kleppner, N. F. Ramsey, and P. Fjelstadt, *Phys. Rev. Lett.* **1**, 232 (1958).
26. J. K. Baird, P. D. Miller, W. Dress, and N. F. Ramsey, *Phys. Rev.* **179**, 1285 (1969).
27. H. M. Goldenberg, D. Kleppner, and N. F. Ramsey, *Phys. Rev.* **123**, 530 (1961).
28. L. Essen and V. L. Parry, *Nature* **176**, 280, 284 (1955).
29. F. H. Reder, "Atomic Clocks and Their Applications," USASRDL Tech. Rep. 2230 (AD 265452), 1961.
30. N. F. Ramsey, *Phys. Rev.* **100**, 1191 (1964); 109, 822 (1958); *J. Phys.* (Paris) **19**, 809 (1958); and I. Esterman, editor, *Recent Research in Molecular Beans*, 107, Academic Press, New York, 1959.
31. Proc. Frequency Control Symposia 1964–1976; *IEEE Trans. Instrum. Meas.* **IM-13** (1964); *IEEE Trans. Instrum. Meas.* **IM-15** (1966); *IEEE Trans Instrum. Meas.* **IM-19** (1970); also *IEEE Trans. Quantum Electron.* **QE-5** (1969). R. E. Beehler, *Ann. Freq. Control Symp.* **25**, x (1971).
32. J. Bonanomi, *Quantum Electronics III*, Columbia Univ. Press, New York, 1964.
33. C. E. Cleeton and N. H. Williams, *Phys. Rev.* **45**, 234 (1934).
34. E. V. Condon and H. Odishaw, *Handbook of Physics*, McGraw-Hill, New York, 1967.
35. C. H. Townes and A. L. Schawlow, *Microwave Spectroscopy*, McGraw-Hill, New York, 1955.
36. R. H. Dicke, *Phys. Rev.* **89**, 472 (1953).
37. F. Bitter, *Phys. Rev.* **76**, 833 (1949), and M. H. T. Pryee, *Phys. Rev.* **77**, 136 (1950).
38. J. Borssel and A. Kastler, *C. R. Acad. Sci.* (Paris) **229**, 1213 (1949).
39. A. Kastler, *J. Phys.* (Paris) **11**, 225 (1950), and *J. Opt. Soc. Amer.* **47**, 460 (1957).

40. M. A. Bouchiat and J. Brossel, *Phys. Rev.* **147**, 41 (1966).
41. R. V. Pound, E. M. Purcell, and N. F. Ramsey, *Phys. Rev.* **81**, 156, 278, 279 (1951) and **103**, 20 (1956).
42. J. P. Gordon, H. Z. Zeiger, and C. H. Townes, Columbia Rad. Lab. Prog. Rep., Dec. 1951; *J. Commun. Eng. Japan* **36**, 650 (1953); and *Phys. Rev.* **95**, 282 (1954); also *Phys. Rev.* **99**, 1264 (1955).
43. J. Weber, "Amplification of Microwave Radiation by Substances Not in Thermal Equilibrium," *Trans. IRE Electron Devices* **ED-3**, 1–4 (June 1953).
44. N. G. Basov and A. M. Prokhorov, *Zh. Eksp. Teor. Fiz.* **27**, 431 (1954) and **28**, 249 (1955); or *JETP Lett.* **1**, 184 (1955).
45. N. Bloembergen, *Phys. Rev.* **104**, 324 (1956).
46. A. L. Schawlow and C. H. Townes, *Phys. Rev.* **112**, 1940 (1958).
47. H. M. Goldenberg, D. Kleppner, and N. F. Ramsey, *Phys. Rev. Lett.* **8**, 361 (1960).
48. D. Kleppner, H. M. Goldenberg, and N. F. Ramsey, *Phys. Rev.* **126**, 603 (1962), and H. C. Berg and D. Kleppner, *Rev. Sci. Instr.* **33**, 238 (1962).
49. P. L. Bender, *Phys. Rev.* **132**, 2154 (1963).
50. S. B. Crampton, *Phys. Rev.* **158**, 57 (1967).
51. S. B. Crampton, J. A. Duvivier, G. S. Read, and E. R. Williams, *Phys. Rev.* **A5**, 1752 (1972), and S. B. Crampton and H. T. M. Wong, *Phys. Rev.* **A12**, 1305 (1975); *Bull. Am. Phys.* **18**, 709 (1973) and **19**, 83 (1974).
52. D. Kleppner, H. C. Berg, S. B. Crampton, N. F. Ramsey, R. F. C. Vessot, H. E. Peters, and J. Vanier, *Phys. Rev.* **138**, A972 (1965).
53. R. F. C. Vessot, NASA reports and private communications, 1976.
54. C. Audoin, *Rev. Phys. Appl.* **1**, 2 (1966) and **2**, 309 (1967); *Phys. Lett.* **28A**, 373 (1968); C. Audoin, M. Desaintfuscien, P. Petit, and J. P. Schermann, *Nucl. Instrum. Methods* **69**, 1 (1969); "Design of a Double Focalization in a Hydrogen Maser," *IEEE Trans. Instrum. Meas.* **IM-17**, 351–358 (Dec. 1968) (this work utilizes a useful double-focusing method to eliminate the undesired $F = 1 mF = 1$ state from the focused beam); *Electron. Lett.* **5**, no. 13 (1969); *C. R. Acad. Sci.* (Paris) **264**, 698 (1967) and **270**, 906 (1970); "Double-Resonance Method for Determination of Level Populations," *IEEE J. Quantum Electron.* **QE-5**, 431–434 (Sep. 1969); and S. Haroche, C. Cohen-Tannoudji, C. Audoin, and J. P. Schermann, *Phys. Rev. Lett.* **24**, 861 (1970).
55. H. W. Hellwig, *Proc. IEEE* **63**, 212 (1975); *Metrologia* **6**, 56 (1970); and *NBS Technical Note* **616**, 1 (1972) and **662**, 1 (1975).
56. E. Uzgiris and N. F. Ramsey, *Phys. Rev.* **A1**, 429 (1970).
57. Laboratories that have engaged in hydrogen maser studies include Harvard University, Massachusetts Institute of Technology, Bomac Laboratories, Varian Associates, Hewlett-Packard, the National Bureau of Standards, Goddard Space Flight Center, the Jet Propulsion Laboratory, U.S. Electronics Command, Hughes Research Laboratory, Laboratoire de l'Havloge Atomique, Orsay (France), PTB (Braunschavey, Germany), the National Research Council and Laval University (Canada), R.R.L. (Tokyo, Japan), LSRH (Neuchatel, Switzerland), and the Lebedev Institute (Moscow, U.S.S.R.).
58. D. Brenner, *J. Appl. Phys.* **41**, 2942 (1970).
59. P. E. Debely, *Rev. Sci. Instr.* **41**, 1290 (1970).
60. V. Reinhardt and J. Lavanceau, *Proc. Annu. Symp. on Freq. Control* **28**, 379 (1974).
61. P. W. Zitzewitz and N. F. Ramsey, *Phys. Rev.* **A3**, 51 (1971).
62. T. H. Mainman, *Nature*, **187**, 493 (1960).
63. A. Javan, W. Bennett, and D. R. Herriott, *Phys. Rev. Lett.* **6**, 106 (1961); L. O. Hocker, J. G. Small, and A. Javan, *Phys. Rev. Lett.* **29A**, 321 (1969).
64. R. L. Barger and J. L. Hall, *Phys. Rev. Lett.* **22**, 4 (1969); *Appl. Phys. Lett.* **22**, 196 (1973); *Atomic Masers and Fundamental Constants* **5**, 322 (1976) (Plenum Press).
65. J. L. Hall and C. Bordé, *Phys. Rev. Lett.* **30**, 1101 (1973).
66. T. W. Hänsch, M. D. Levenson, and A. L. Schawlow, *Phys. Rev. Lett.* **26**, 946 (1971).
67. K. M. Evenson, et al., *Phys. Rev. Lett.* **29**, 1346 (1972).
68. R. G. Brewer, *Science* **178**, 247 (1972).
69. R. Z. Barger, T. C. English, and J. B. West, *Annu. Symp. on Freq. Control* **29**, 316 (1975) (U.S. Army Signal Corps. Ft. Monmouth, N.J.).
70. H. R. Schlossberg and A. Javan, *Phys. Rev. Lett.* **17**, 1242 (1966).
71. T. W. Hänsch, I. S. Shahin, and A. L. Schawlow, *Phys. Rev. Lett.* **27**, 707 (1971).
72. L. S. Vasikenko, V. P. Chebotaev, and A. V. Shishaev *JETP Lett.* **12**, 113 (1970); D. Pritchard, J. Apt, and T. W. Ducas, *Phys. Rev.* **32**, 641 (1974); M. D. Levenson and N. Bloembergen, *Phys. Rev. Lett.* **32**, 645 (1974); F. Birahan, B. Cagnac, and G. Grynberg, *Phys. Rev. Lett.* **32**, 643 (1974); T. W. Hänsch, et al., *Opt. Comm.* **11**, 50 (1974).
73. K. M. Evenson, J. S. Wells, F. R. Petersen, B. L. Danielson, and G. W. Day, *Appl. Phys. Lett.* **22**, 192 (1973) and **20**, 296 (1972); *Phys. Rev. Lett.* **31**, 573 (1973).

74. B. W. Jolliffe, W. R. C. Rowley, K. C. Shotton, A. J. Wallard, and P. Z. Woods, *Nature* **251**, 46 (1974).
75. H. G. Dehmelt, *Phys. Rev.* **109**, 381 (1959); *Advances in Atomic Molecular Physics* **3**, 53 (1967) and **5**, 109 (1959).
76. D. Wineland and H. Dehmelt, *Bull. Am. Phys. Soc.* **18**, 1521 (1973) and **20**, 60, 61, 637 (1975).
77. H. G. Dehmelt, F. Major, E. N. Fortson, and H. A. Schuessler, *Phys. Rev. Lett.* **8**, 213 (1967); *Phys. Rev.* **170**, 91 (1968) and **187**, 5 (1969).
78. S. R. Stein and J. P. Turneauve, *Proc. Annu. Symp. Freq. Control* **27**, 414 (1973), and HPL 741, Stanford High Energy Physics Laboratory, Stanford, Calif.
79. T. W. Hänsch and A. L. Schawlow, *Opt. Commun. Netherlands* **13**, 68 (1975).
80. D. W. Allan, *Proc. IEEE* **54**, 221 (1966).



Gordon S. Kino is a Professor of Electrical Engineering at Stanford University. Dr. Kino has carried out experimental and theoretical work and has published more than a hundred papers in such fields as microwave triodes, traveling wave tubes, klystrons, microwave tubes, magnetrons, electron guns, wave propagation in plasmas, solid-state oscillators and amplifiers, microwave acoustics, and acoustic imaging devices for medical instrumentation and nondestructive testing. He has given invited talks on acoustic waves at conferences in the United States, Australia, Japan, and England. Dr. Kino was born in Melbourne, Australia; earned B.Sc. and M.Sc. degrees in mathematics at London University in England and a Ph.D. at Stanford University; and received a Guggenheim Fellowship in 1967. He is a Fellow of IEEE and of the American Physical Society and is a member of the National Academy of Engineering.



H. J. Shaw is a Adjunct Professor at Stanford University, has been a consultant to a large number of electronic firms, and in 1968–1969 was liaison scientist for the Office of Naval Research in London, England. Dr. Shaw's present work involves research on real-time acoustic imaging systems, acoustic nondestructive testing, and microwave and optical devices for inertial rotation sensing. Earlier, he was engaged in research on microwave antennas and high-power microwave tubes; microwave ferrite devices involving resonance and spin waves; microwave acoustic devices including thin-film transducers, bulk wave delay lines, and acousto-optic signal processors; and surface acoustic wave devices including transducers, delay lines, amplifiers, convolvers, matched filters, and optical scanners. Dr. Shaw was born in Seattle, Wash., and earned a B.A. at the University of Washington and a Ph.D. at Stanford University. He is a Fellow of IEEE, past chairman of the Professional Group on Electron Devices and of the San Francisco Section IRE, and past member of the Administrative Committee of the IEEE Group on Sonics and Ultrasonics.

DEVELOPMENT OF SURFACE ACOUSTIC WAVE DEVICES

Gordon S. Kino and H. J. Shaw

Edward L. Ginzton Laboratory
W. W. Hansen Laboratories of Physics
Stanford University
Stanford, Calif.

PREFACE

This article describes a new technology and group of devices which offer new dimensions in data processing, data storage, and delay. This new technology using surface acoustic waves provides the system designer with components in very compact form which are capable of performing certain system functions at data rates probably not achievable in any other way.

Although surface acoustic waves (SAW) have been known to science for a long time, investigation and interest in them has been largely confined to the seismologist (!), who was interested in wavelengths of kilometers whereas for electrical applications we are interested in wavelengths of micrometers. It is only in recent years that the electrical engineer has discovered surface waves and has opened up a whole new range of applications.

Much of the pioneering work in this field was done at Stanford in the late 1960s under Joint Services Electronics Program (JSEP) auspices (as well as support from independent Department of Defense agencies, particularly RADC, ECOM, and ONR). The work at Stanford was associated particularly with the names of Gordon Kino, Calvin Quate, and John Shaw and their students. Key elements for this research were drawn from some earlier work done at several other laboratories, particularly at Bell Telephone Laboratories and the University of California, Berkeley. Subsequent to this early work, a whole new range of experimentation has developed, with a large variety of devices, applications, and behavior being studied in many industrial and government laboratories.

Details of what can be done with such devices are given in the body of this paper. Here we would just like to list the advantages and possibilities of this new technology in summary form to furnish some perspective of what might be possible. Basically, surface acoustic waves provide a unique means for storage, delay, and complex parallel processing of long-duration, wideband signals. Quantitatively, one can operate at frequencies up to 500 MHz quite easily and with some care up to 1000 MHz or so, with bandwidths of the order of 100 MHz or better. Therefore, correspondingly, one can talk of data rates of 100 megabits per second. This is far higher than any competitive devices intended for similar purposes. One can store signals of a millisecond duration for a millisecond and, with some additional refinements, one can store such signals for several milliseconds. Further, one can have access to all or any portion of such an extended signal and can calculate correlation between one set of signals and another,

one of which may have been stored for several milliseconds before the comparison, and do other similar kinds of processing. One can, for example, think of performing Fourier transforms on such groups of data. The basic property which makes all this possible is that any such long-duration signal is spread out over a relatively short path length and completely exposed on an open surface so that one can have access to all or any portion of this for processing both linearly and nonlinearly. Of course, this is all a consequence of the high-frequency acoustic characteristics of special materials and the short wavelengths of the acoustic waves which make it possible to have such long wave trains in a small region of the surface. The fact that it is on the surface makes it accessible through suitable transducers.

The most important thing about all of these possible uses is that the devices involved are all planar, they are generally miniaturized so as to be compatible with microcircuits and semiconductor devices which can be used in association with them. The accuracy required for surface wave devices, transducers, filters, couplers, and so forth can be achieved by standard photolithography, and so we have a natural phenomenon which gives us these desirable characteristics of bandwidth, storage, and so forth but compatible with a technology which has been perfected for other purposes and can be applied very nicely to surface waves.

Aside from these characteristics of high-density storage, accessibility, precision, and natural miniaturization, one might mention at least one other application and that is that one can design transducers used for launching surface acoustic waves as filter elements. In a sense, a transducer of this kind acts like an end fire array and just as in an end fire array, one can feed all the elements of the array in parallel but selecting the dimensions and efficiency of individual elements and their spacing in such a way that the radiated signal demonstrates the required filter characteristics. This results in being able to produce quite complex characteristics in a much simpler form than with normal circuit elements. This property is also something which has now led to widespread application.

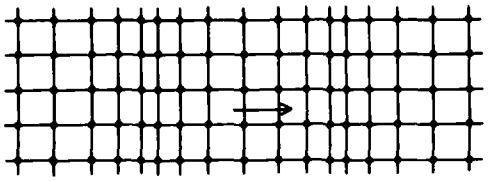
The main body of the text discusses these various applications, including linear and nonlinear characteristics, correlation, convolution, integrated amplifiers, and combinations of surface wave media with semiconductors to provide improved devices of a wide variety. It is the opinion of the authors that we are still far from having reached all the possible applications of this very significant new technology.

M. Chodorow

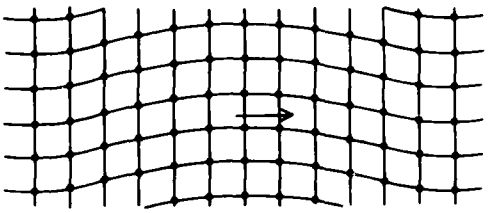
INTRODUCTION

The existence of surface acoustic waves was predicted by Lord Rayleigh about a century ago. In the intervening years their principal importance was in seismology. The uses we want to consider here are concerned mainly with signal processing, involving very much higher frequencies, typically in the UHF range. In this range, surface acoustic waves in crystals have unique properties which have resulted in a substantial amount of research and development on devices for signal processing, communication, and instrumentation using these waves. Their small wavelength is compatible with microcircuit dimensions, their slow propagation velocity allows a very small wave train to store a very large amount of information on a small crystal, and their relatively low attenuation in certain crystals at very high frequencies allows them to handle large amounts of analog or digital information at very high data rates.

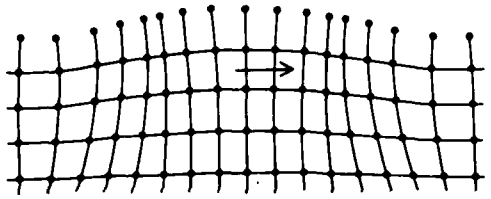
The most common type of acoustic wave, which can exist in gases, liquids, or solids, is the longitudinal wave, in which the medium is alternately compressed and expanded along the direction of propagation, as indicated schematically in Figure 1(a). A second type of wave, which generally exists only in solids, is the transverse or shear wave, in which the material particles move transversely to the propagation direction (Figure 1(b)). This type of wave possesses the characteristic of polarization in the transverse plane, analogous to the polarization of an electromagnetic wave. The Rayleigh wave or surface acoustic wave exists only near the free surface of a solid (Figure 1(c)).



(a) LONGITUDINAL WAVE



(b) TRANSVERSE WAVE OR SHEAR WAVE



(c) RAYLEIGH WAVE OR ACOUSTIC SURFACE WAVE

Figure 1—Schematic representation of basic acoustic wave types

Its particle motion is more complex, in having both longitudinal and shear components, both of which are required to satisfy the boundary conditions at the surface.

The study of high-frequency acoustic waves in crystals was motivated by the interesting propagation characteristics for bulk waves displayed by some crystals. It was demonstrated some 15 years ago that bulk acoustic waves with frequencies in the GHz range could propagate for distances of several centimeters in a quartz crystal [1]. This result led to a substantial amount of research on the room-temperature propagation characteristics of both longitudinal and shear waves in a number of crystals and the development of techniques for evaporating thin piezoelectric films of cadmium sulfide, zinc oxide, and other materials that could be used as transducers for the excitation and detection of bulk waves. These efforts were very successful and led to a series of delay lines which were able to operate efficiently at frequencies in the GHz microwave range, up to X band, and also delay lines capable of bandwidths of the order of 1000 MHz. Although various investigators went on to demonstrate more sophisticated signal-processing functions which can be performed within bulk wave delay lines, in practice their use has been limited largely to applications in which they perform as simple two-port delay lines in radar and computer systems.

Not long after this growth of activity in microwave bulk acoustic wave devices there were demonstrations of surface acoustic wave propagation on piezoelectric crystal surfaces, and the interdigital transducer for the excitation of such surface acoustic waves was introduced. For some period of time, there was apathy concerning surface waves because it appeared that interdigital transducers would necessarily have either very high insertion loss or severely limited bandwidth and also that surface waves would be excessively sensitive to surface scratches and imperfections, contamination, atmospheric loading, and the like. However, in the late 1960s it was shown [2], using new crystals then becoming available, principally lithium niobate [3], that these apprehensions were unfounded. By applying electrical and acoustic circuit engineering techniques, it was possible to achieve both low insertion loss and large bandwidth in practical interdigital transducers and to demonstrate reliable, low-loss propagation in delay lines constructed of these crystals, using standard optical polishing techniques in processing the delay line surface.

There was a major advantage in such surface acoustic wave delay lines. The propagation velocity of acoustic waves in solids is ordinarily some five orders of magnitude lower than that of electromagnetic waves. This means that small acoustic devices, having dimensions of the order of centimeters, can have propagation delay of the order of tens of microseconds and more. As a result we can have, on the face of a small crystal, a wave train containing an enormous amount of information, which would occupy a distance in space of a fraction of a mile if it were carried by an electrical cable. In a bulk wave delay line, where the wave is buried within the crystal, there are no practical means for gaining access to this information, except at the single output port. However, in a surface wave delay line most of the wave energy

is contained within a distance of a few micrometers below the surface. Thus, we have the information in a compact format where we can "read" it all at the same time by means of transducer "taps" located on the surface and apply to it one or more of the important operations which come under the general heading of signal processing. This is not meant to minimize the potential of bulk wave systems, with their unique capability for high-speed in-band operation in the GHz frequency range; it is to say, however, that the surface acoustic wave delay line opened a whole new range of devices not accessible to bulk wave systems.

The surface acoustic wave art is still relatively new. By comparison with the funds spent on the development of silicon technology, the amount of support which has gone into surface acoustic wave devices is very small. As has been the case with other materials-dependent fields, a key item is the price of materials involved in device construction. This is, of course, tied to the device volume and the usual iterations of device development and materials development, each spurred by the other, are necessary before practical markets are achieved. In the surface acoustic wave case, the materials development has been in progress at a modest rate for a number of years now and appears to be accelerating in response to potential growths arising from promising developments in commercial filters for radio and TV, retrofitting of surface acoustic wave components into radar and communication systems, and the development of completely new surface acoustic wave elements for future systems. For example, round, thin, polished wafers of oriented single crystal lithium niobate are available from suppliers which allow the construction of frequency filters with a materials cost in the range of tens of cents per filter, which is two or three orders of magnitude cheaper than the cost of materials for the same device when surface acoustic wave device development was in its infancy. Production techniques for surface acoustic wave frequency filters and similar components can typically begin with a circular wafer designed to fit into standard production silicon wafer holders, followed by photolithographic development of a large 2 in. by 2 in. (50.8 by 50.8 mm) matrix of separate filters and by dicing of the wafer into small complete filters on rectangular chips with face dimensions of the order of 1/4 in. (6.35 mm) which can easily fit inside a standard integrated circuit flat pack. Time-delay filters often use crystal plates which are larger but still small as compared to alternate approaches for accomplishing the same function.

The next section is devoted to devices whose operation depends on linear, passive interactions between electrical RF signals and the surface acoustic waves, including a variety of types of filters. The final section deals with amplifiers and convolvers involving active interactions and nonlinear interactions between surface acoustic waves and semiconductors.

LINEAR PASSIVE DEVICES

Surface acoustic waves are well suited to a variety of types of filters, and it is convenient to divide these into two categories, which will be referred to as frequency filters and time-delay filters. In the former, emphasis is on synthesizing detailed variations of insertion loss as a function of frequency for applications where, usually, it is desired to minimize the time delay. In the latter, substantial time delay is an essential characteristic. In the following paragraphs we will discuss filters of these two types, as well as other passive devices which depend on filter characteristics. We begin with a brief description of the basic interdigital transducer which, even in its simplest form, has characteristics of a bandpass filter.

Interdigital Transducers and Basic Delay Lines

The technology of surface acoustic waves began expanding rapidly with the development of the interdigital transducer, an efficient type of transducer for converting an electrical signal into an acoustic wave or vice versa. An interdigital transducer is normally placed on the surface of a piezoelectric material. When an RF electric field is applied to a piezoelectric material, the material will vibrate in unison with the field and an acoustic wave will be generated. The required electric field can be produced at the surface of a piezoelectric crystal by applying an electric potential between a

pair of parallel metal electrodes deposited on the surface of the crystal, as in Figure 2(a). This results in excitation of a surface acoustic wave that can be reconverted to an electrical signal at a second pair of similar electrodes. A single pair of electrodes is inefficient, and it is customary to use an array of electrodes in an interdigital pattern, as in Figure 2(b). Each pair of electrodes excites a surface acoustic wave and the transducer array is designed so that the waves reinforce one another, to provide a lower insertion loss. This is accomplished by choosing the spacing between adjacent pairs of "fingers" to be one wavelength so that a surface acoustic wave will travel that distance in just the time required for the excitation to be reinforced at the next finger pair.

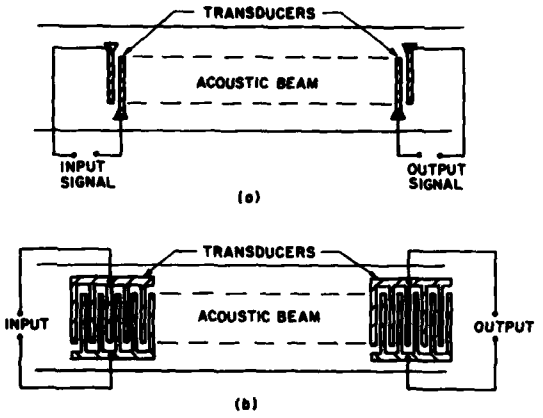


Figure 2—Schematic of surface acoustic wave delay line
(a) Simple single finger-pair transducers
(b) Multiple finger-pair interdigital transducer arrays

Early demonstrations of the interdigital transducer concept were made at Bell Telephone Laboratories and the University of California at Berkeley. The first example of surface acoustic wave delay lines having large bandwidth and low insertion loss was at Stanford under U.S. Air Force support, and this was followed shortly thereafter by the first demonstration of an amplifier for surface acoustic waves having large gain, under the JSEP program. The latter development was an outgrowth of a research program on acoustic amplifiers using bulk waves, which had succeeded in demonstrating the first stable acoustic bulk wave amplifiers under the same Stanford JSEP program.

Figure 3 shows the performance of one of the original two-port surface acoustic wave delay lines of the type of Figure 2(b), fabricated on the surface of a lithium niobate crystal. Each interdigital transducer consists of five identical finger pairs with widths and spacings of 8μm, which gives a center frequency of operation slightly above 100 MHz. For other frequencies, these dimensions are scaled inversely with frequency.

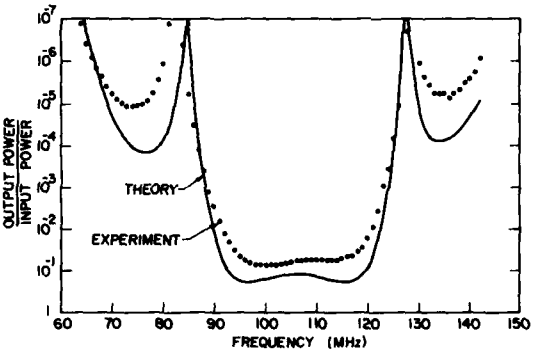


Figure 3—Bandpass frequency response of an early two-port surface acoustic wave delay line

The interdigital transducer, which has been the keystone in the development of surface acoustic wave devices, is typically formed by vacuum depositing aluminum or gold to fractional micrometer thicknesses. Lapping techniques borrowed from the optical polishing art are used to prepare the substrate surfaces, and photolithographic procedures borrowed from the semiconductor integrated electronics industry are used to define the electrode geometries. In recent years there have been substantial improvements in these procedures, which were required because surface wave patterns often involve larger surface areas than in microelectronics. The various piezoelectric crystals available, such as quartz, bismuth germanium oxide, lithium niobate, and lithium tantalate, offer various possibilities with regard to propagation loss, diffraction effects, time delay per unit length, upper frequency, bandwidth, temperature dependence, and so forth. Higher piezoelectric constant usually affords a larger

product of efficiency and bandwidth. Low values of insertion loss are achievable by proper design. The minimum insertion loss achievable with a simple uniform interdigital transducer is 3 dB, which is a basic limitation associated with the fact that the transducer is acoustically symmetrical and radiates equally in both directions along the surface of the delay line. Although transducers can be designed which are unidirectional, they have some fabrication and performance limitations and have not yet had wide acceptance. The total insertion loss also involves dissipative losses in both the electrical and acoustic circuits associated with the transducer, but by good design the theoretical minimum insertion loss can be approached to within less than 1 dB over a wide range of frequencies. Transducer designs have become increasingly more sophisticated, and engineers now have a large number of techniques for designing arrays, including complicated profiles of electrode widths, lengths, and spacings; choice of material combinations for the electrodes and substrate; use of intervening dielectric films; serrated electrodes; and so forth.

Frequency Filters

One of the most important areas of research and development in surface wave devices and the first area to have a civilian commercial application is that of frequency filters. As seen in Figure 3, the simplest form of interdigital transducer has bandpass filter characteristics. This property can be extended with considerable generality to synthesize filters having desired passband and stopband characteristics. This can be done by tailoring the lengths of individual fingers (measured perpendicular to the acoustic propagation axis) and adjusting the locations of individual fingers along the propagation axis. This profiled array then presents a geometrical pattern when viewed by eye, and we can loosely say that this pattern is the Fourier transform of the frequency response of the device. In this way and in others, it is possible to relate the detailed frequency characteristics of the device to the geometry of the array. This is a very important situation, because it transfers the problem of filter construction to the fabrication of geometrical electrode profiles. Once designed and tested, the transducer can then be replicated endlessly using photolithography, with very high precision and low cost. The resulting filters emerge from the assembly line pretuned, with no alignment procedures required. This is an example of a basic philosophy of surface acoustic wave devices. One incorporates the complicated aspects of a device design into the geometrical design of an electrode array.

Surface acoustic wave filters are of direct interest at this time to the radio and TV industry, as well as for a variety of more sophisticated applications. They are applicable in channel selection and filtering in both RF and IF channels and as frequency discriminators. They are applicable in various spread spectrum and frequency agile systems for radar and communication. Banks of surface acoustic wave filters can be used for frequency multiplexing, frequency sorting, frequency synthesis, and so forth. In all of these applications, surface acoustic wave filters are capable of better characteristics and smaller size than conventional devices.

Figure 4 illustrates the principles discussed previously. An input signal can be fed into either of the transducers and the filtered output taken from the other. The coupling strengths of the various electrode pairs are determined by the locations of cuts in the individual fingers, and the geometrical pattern of finger lengths illustrated in the right-hand array is tailored to produce a rectangular passband which can have an accurately flat and nondispersive response within the passband and high rejection outside this band.

Many novel special techniques have been developed to improve and simplify filter design, including means for changing electrode coupling strengths without varying their lengths, procedures for designing electrode shapes and positioning to minimize spurious acoustic reflections, and so forth, leading to patterns which, while sometimes complex, can be readily handled by photolithography.

Time-Delay Filters

The simplest form of time-delay filter using surface acoustic waves is the uniform tapped delay line, which contains an array of interdigital trans-

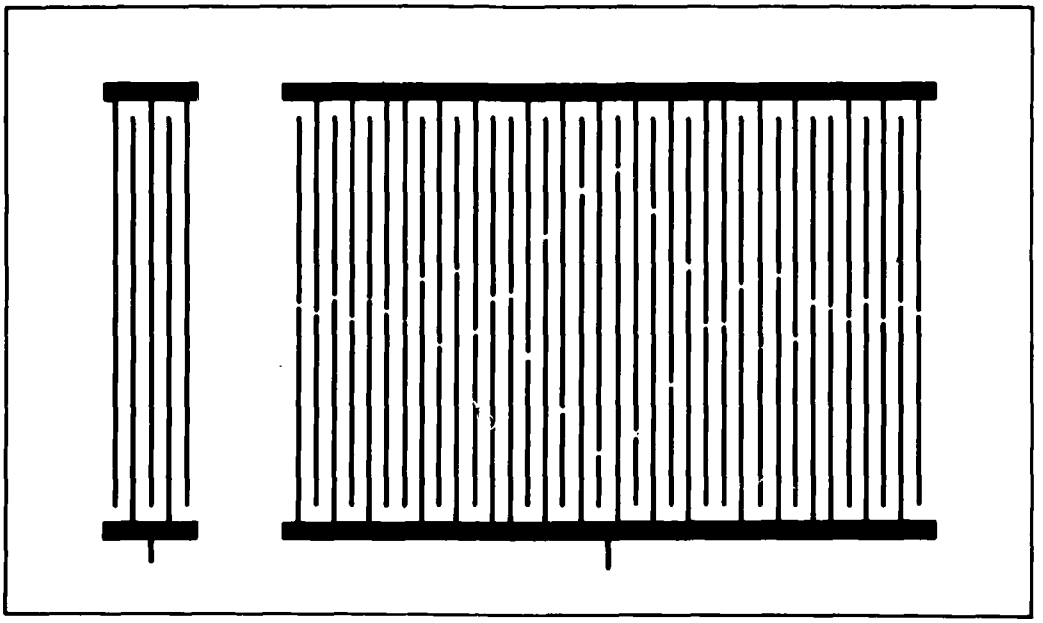


Figure 4—Schematic of surface acoustic wave bandpass filter

ducers equally spaced along the path of the signal, as in Figure 5. This is an example of a transversal filter, in which one can sample the signal at intermediate points along its path and combine these samples such as to achieve a desired transformation or processing of the signal. A signal in the form of a short RF pulse, introduced into an input transducer at one end of the delay line of Figure 5, produces electrical output pulses at all of the tapping transducers as it travels past them one at a time. If the electrical terminals of all of these transducers are connected together, the output will be a train of successive pulses, in which arbitrary pulses can have their polarities reversed

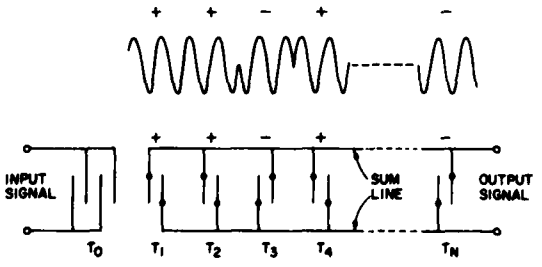


Figure 5—Schematic representation of tapped delay line

with respect to the others by reversing the connections to the corresponding transducers. For example, in the case shown, the output pulse from the third tapping transducer will have a polarity opposite to that from the first two transducers. Thus, the transducer array can be designed to give a coded electrical output signal. In this particular case, for purposes of illustration, we have chosen a so-called biphase coded digital signal, in which the code information is impressed upon the signal by varying the polarity of the RF waveform from pulse to pulse in an arbitrary way. If this same coded digital signal is introduced into the delay line as the input signal, then as the train of pulses passes under the output transducers the polarities of the individual pulses will correspond exactly with the polarities of the transducers at one particular time, and there will be a large output pulse at that instant. This pulse is decreased in length, with respect to the length of the input signal, by a factor equal to the product of the bandwidth and the time delay of the filter, which is the so-called pulse compression ratio. The magnitude of the output pulse is increased over that of the input pulse by the same ratio, and this is referred to as processing gain. This represents a gain in signal

strength with respect to the strength of background noise, i.e., a gain in signal-to-noise ratio. If a different signal, having a different sequence of positive and negative RF pulses, is introduced into the delay line, there will be no instant at which it will match the polarities of all of the tapping transducers, and the peak output signal will be reduced. This is an example of the ability of transversal filters to perform pattern recognition, by selecting a signal of given code from all other signals. Experimental input and output signals from a tapped delay line of this type are shown in Figure 6. This delay line [4] has 127 taps, spanning a time delay of 25.4 μs, and the pulse compression ratio is approximately 100.



Figure 6—Experimental tapped delay line waveforms

Surface acoustic wave tapped delay lines have much in common with analog shift registers. The clock rate is fixed in the case of delay lines, since data stored on the delay line propagates automatically along the line at the constant, frequency-independent propagation velocity of surface acoustic waves on the material in question. In general, as compared to the principal electronic circuit devices which can operate as analog shift registers, namely charge transfer devices such as bucket brigade devices (BBD) and charge coupled devices (CCD), surface acoustic wave devices operate at higher frequencies and higher data rates. Charge transfer devices are generally concerned with frequencies below 10 MHz, and surface wave devices are generally concerned with frequencies above 10 MHz. Charge transfer devices generally operate on baseband signals, while surface wave devices operate in a bandpass mode. Both devices can be tapped to form transversal filters. CCDs tend to be more defect sensitive than surface acoustic wave devices, in that a single bad cell can render the entire register inoperative, while surface acoustic wave devices are less prone to such effects, although a surface scratch can partially scatter the surface wave column and degrade the performance of the device.

A second important example of signal processing within a surface acoustic wave delay line is one which uses an analog signal of the type illustrated in Figure 7. The signal shown is a so-called

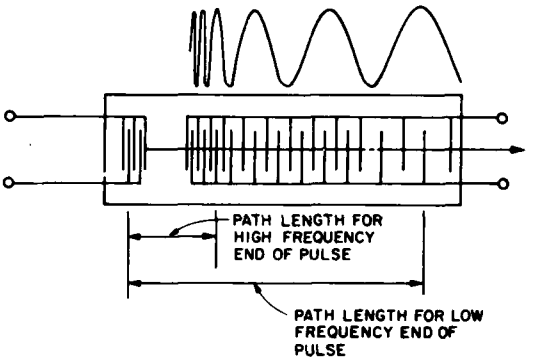


Figure 7—Schematic representation of chirp pulse compression filter

chirp pulse, whose amplitude is constant but whose instantaneous frequency varies linearly with time. The finger spacing of the transducer array is varied along its length to match the frequency variation across the chirp. The left-hand end of the array responds to the highest frequencies and the right-hand end to the lowest frequencies. At the instant shown, the chirp signal, which is traveling to the right, registers exactly with the array, much as in the case of Figure 5, and an intense output burst results at the right-hand terminals. This is a dispersive filter, in which the low-frequency end of the signal is delayed more than the high-frequency end, allowing the trailing edge of the long input pulse to catch up with the leading edge, thus collapsing the pulse. Pulse compression techniques of this type are of great importance in a variety of systems, perhaps the best known being radar systems using pulse compression, in which the transmitted signal from the radar is chirped and, after returning from a target, is passed through a pulse compression filter which

compresses it into a short pulse. In this way it is possible to use a long pulse, containing large energy for long-distance ranging, and to compress it in the receiver into a short intense pulse for accurate timing and range resolution. Similarly, if an ultrasonic ranging system is used to probe objects in living tissues, it is possible by the same techniques to limit the peak power to nondestructive levels and still obtain accurate distance resolution and discrimination against interfering signals. Also, in secure radar and communications systems, chirping represents one form of coding of a signal, in which a listener needs to know the code, in this case the chirp rate, and needs to have a chirp compression filter which operates at this rate, in order to receive the signal. Indeed, the signal can be below the thermal or background noise level and, when received by a compressive receiver containing a filter matched to its chirp rate, the signal can be extracted from the background noise level with some desired signal-to-noise ratio. At the same time, an ordinary receiver would not be aware of the presence of the signal. Surface acoustic waves fit naturally in this picture in operations requiring high chirp rates, that is, a high rate of change of frequency versus time across the chirp, together with a large total frequency excursion or bandwidth. Compression ratios in the range of 1000 to 10000 can be reached with surface acoustic wave systems. Surface acoustic wave pulse compression filters have been built with bandwidth exceeding 500 MHz with time delays (chirp length) of the order of a microsecond. The range resolution of a radar system is the inverse of the bandwidth, and this bandwidth corresponds to a target range resolution capability of 1 ft (0.31 m). At smaller bandwidths larger time delays have been achieved.

Chirp arrays of the above type are also applicable to nonscanning spectrum analyzers. In fact, they are applicable to calculating the complete complex Fourier transform of an arbitrary incoming analog signal. In this case, use can be made of an algorithm known in signal processing for some years, to the effect that one can calculate the Fourier transform of a signal by modulating that signal onto a chirp carrier, followed by convolution of the resulting modulated signal with a chirp, followed finally by multiplication by a chirp. The modulation and multiplication operations can be carried out using ordinary electronic mixers. For the convolution operation, chirp filters of the above type are very attractive, in cases where one wants to calculate Fourier transforms in real time involving substantial bandwidths and high analog or digital data rates. In this connection we should point out the general property that, when an arbitrary signal is incident upon an interdigital surface acoustic wave transducer from either the electrical or the acoustic side, the output of the transducer is the cross-convolution between the input signal and the geometrical pattern of the transducer.

A major addition to the surface acoustic wave art was recently made [5] with a device termed the reflective array compressor (RAC). This device is an alternative form of transversal filter in which arrays of parallel grooves on the delay line surface perform the function of tapping normally performed by interdigital electrode arrays. The key idea is such that a groove acts as a tap for the surface acoustic wave, because a portion of the reflected surface acoustic wave arising at the groove can be collected elsewhere by an interdigital surface, having width and depth in the micrometer range. We can get an idea of the operation of such devices by considering two extreme surface acoustic wave paths. A signal starting at interdigital transducer A and having a frequency mutually parallel and arranged in a herringbone pattern, with uniformally increasing groove-to-groove spacing proceeding from left to right. A surface acoustic wave encountering a groove in the surface will scatter a portion of the surface wave into other surface waves and into bulk acoustic waves. As applied to high-frequency surface acoustic wave devices, the groove is no more than an accurately fabricated scratch on the crystal surface, having width and depth in the micrometer range. We can get an idea of the operation of such devices by considering two extreme surface acoustic wave paths. A signal starting at interdigital transducer A and having a frequency such that the average spacing between grooves in the vicinity of B is one wavelength, will be partially reflected into a surface acoustic wave traveling from B to C, where it will again be partially reflected and travel to interdigital transducer D. The total time delay for this signal will be propor-

tional to the path length ABCD. Similarly, a signal of lower frequency will follow the path AEFD and experience a longer time delay. Thus this device accomplishes the same type of dispersion characteristic, wherein time delay is a function of frequency, as for the chirped interdigital structure of Figure 7, and can be used to perform the same functions. The folded paths in Figure 8 give twice the time delay for a given substrate length, and chirp pulse compression filters with time delay exceeding 100 μs are achievable by this means.

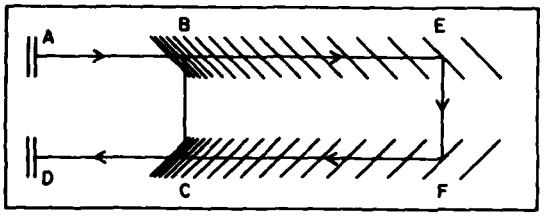


Figure 8—Schematic representation of reflective array compressor

Reflective arrays are less defect sensitive than interdigital arrays, and there is also an advantage in terms of higher frequency operation in that the grooves in reflective arrays are generally spaced the order of a wavelength as compared to the quarter-wavelength spacing which is more characteristic of interdigital arrays so that the dimensional requirements are less stringent. The grooves are conveniently fabricated by ion etching.

Long Delay Lines

We have seen that delay lines can be designed to have large bandwidth. Progress has also been made in increasing the time delay to further increase the time-bandwidth products available. The problem is to obtain a long propagation path on a crystal of manageable overall size. Several approaches are indicated schematically in Figure 9. At (a) is a so-called wraparound delay line plate in which a surface acoustic wave beam makes multiple helical transits around the periphery of a flat crystal plate, being carried from the top surface to the bottom surface by means of carefully rounded and polished end faces on the plate. At (b) is a delay line in which a surface acoustic wave beam travels in a folded path to build up long delays. The path folding is achieved by means of so-called surface acoustic wave track changers, which can transfer an acoustic surface wave beam, traveling along one path, over to an adjacent parallel path. The usual track changer is a form of so-called multistrip coupler, which is another surface acoustic wave component based on deposited electrode technology of the same type used in construction of ordinary interdigital transducers. It provides a directional coupler for surface acoustic waves and can be configured to perform various useful functions via either parallel or antiparallel track changing, including compensation for effects of beam spreading of surface acoustic waves due to diffraction in surface acoustic wave delay line devices. At (c) is a so-called disk delay line which might, for purposes of visualization, be regarded as a wraparound delay line as in (a) rotated about the central axis normal to its flat faces. A surface acoustic wave from a transducer on one of the flat faces travels around in a crisscrossing path, again being carried between the top and bottom surfaces of the disk by traveling around the rounded and polished edges. The curved edge of the disk acts as a converging lens which can be designed to counteract the effects of diffraction spreading of the wave, thereby decreasing the insertion loss over large total pathlengths. The forms in (a) and (b) are potentially applicable to general forms of signal proces-

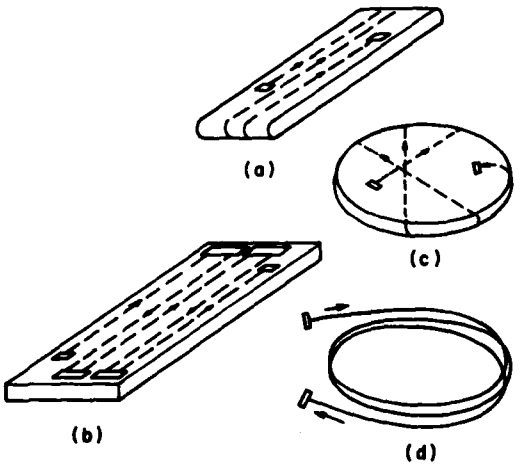


Figure 9—Schematic representation of long delay lines

sing if suitable techniques for fabricating long arrays of phase-coherent taps can be devised, which could be used to form transversal filters as discussed previously in connection with shorter, single-path delay lines. Also long delay lines without intermediate taps have potential use as volatile memory stores in computers and for real-time storage of video information in TV systems. Time delays in the range of 100 μs to 1 ms and time-bandwidth products extending to as high as 6×10^4 with good dynamic range have been demonstrated [6]. By including a surface acoustic wave amplifier directly on the delay line surface, it has been possible to reach time delays up to 20 ms [7].

Surface Acoustic Waveguides

Several approaches to low-loss waveguiding structures have been domonstrated which can contain a surface wave column in a path of constant width [8]. Waveguide types generally break down into two classes, one consisting of topological waveguides in which grooves, ridges, slots, or the like, running parallel to the desired surface acoustic wave beam edges, are fabricated on the delay line surface. The second type consists of thin film waveguides which operate by perturbing the propagation velocity of the surface acoustic waves, such that the velocity is slightly less in the region occupied by the waves than it is in the surrounding areas of the crystal surface. They thus operate on the same basic wave-slowing principle used in dielectric waveguides for RF and optical systems, based on the well-known principle that a wave traveling at a velocity slower than that of the surrounding medium does not radiate into that medium. With waveguides, it is possible to contain the surface wave column into a ribbon or strip whose width is just a few acoustic wavelengths, typically in the range of tens to hundreds of micrometers. This allows a higher density of beams to be placed in a given crystal area on delay lines such as in Figure 9 (a) and (b) before encountering excessive cross-talk levels between adjacent beams [6]. Design tradeoffs which must be considered include a tendency for increased propagation loss over that for unguided waves and dispersion, which is always associated with waveguides of finite width for any type of wave propagation, as opposed to the completely nondispersive propagation which is characteristic of unguided surface acoustic waves. In Figure 9(d) is shown an alternate type of long delay line which employs waveguiding. This consists of a length of fused silica fiber which is not completely unlike fibers used for optical waveguiding and is fabricated using similar fiber drawing techniques, except that its transverse dimensions are somewhat larger than in the optical waveguide case. Such fibers can support various bulk acoustic wave modes when fabricated in the form of cladded solid fibers, but, when fabricated in the form of hollow capillaries, they can support surface acoustic wave modes on their inner surface. Time delays up to the order of a half millisecond have been observed in capillaries of this type [6].

Resonators

The conducting electrodes used to form interdigital arrays cause small reflections of surface waves which can lead to low-level spurious signals, and, in the design of filters, steps are taken to limit these to specified levels. On the other hand, it is possible to optimize these reflections to make constructive use of them to form surface acoustic wave resonators which have considerable promise for use in electronic circuits where it is advantageous to have very small, inexpensive, accurate, pretuned resonators which can be mass produced by photolithographic methods. Arrays of isolated parallel conducting strips deposited on the surface as in Figure 10 can behave as efficient reflectors for surface waves if a large number of strips (typically hundreds) are used and if they are spaced such that the reflected waves from individual strips reinforce each other. The surface waves are then trapped between the two reflectors of Figure 10, making multiple transits between them and creating a standing wave, like electromagnetic waves in a cavity resonator or optical waves in a Fabry-Perot interferometer. Interdigital transducers are used to couple to this resonant standing wave from an external electrical circuit. Etched grooves, as described in connection with reflective array compressors, can also be effectively used to form the reflective gratings.

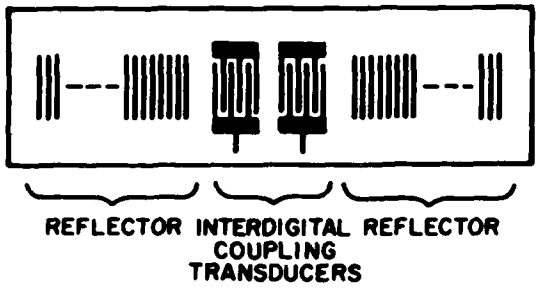


Figure 10—Schematic of surface acoustic wave resonator

Values of resonant Q up to the order of 20000 have been achieved with such resonators at UHF and VHF frequencies, limited by the finite reflectivity of the arrays, as the propagation loss on lithium niobate substrates is low enough to allow approximately another order-of-magnitude increase. Research is also underway to use these resonators as circuit elements in building up ladder networks and other types of classical filter networks.

Stabilized Oscillators

Another device having great potential importance is the surface acoustic wave oscillator, which consists of an amplifier of some standard design whose output is fed back to its input through a surface wave delay line, to form an oscillator whose frequency and frequency stability are determined by the frequency filtering characteristics of the SAW delay line [9]. This oscillator has high frequency and high power capabilities and has the potential for being simpler and cheaper than alternative approaches for producing highly stable signals, such as crystal-controlled multiplier chains.

The key point in these oscillators is that because of the very low propagation velocity of acoustic waves, the delay line can have a very large number of wavelengths between the input and output transducers, measuring up into the thousands. As is well known, the frequency of an oscillator having an external feedback loop adjusts itself such that the phase shift around the loop is $2\pi N$, where N is an integral number. The larger N is, the larger is the short-term stability of the oscillation frequency, the ultimate limit being of the order of one part in N. Both this basic frequency selectivity associated with the length of the delay line path and also the frequency-filtering characteristics of interdigital transducers can be brought into play. The former essentially gives high frequency stability for any of a number of different longitudinal modes (different values of N) having different center frequencies, and the latter are used to select one longitudinal mode from the entire possible comb of modes. Means for correcting for phase shifts within the oscillator itself, resulting from voltage and temperature variation, have been demonstrated, allowing one to approach very closely to the ultimate frequency stability of the delay line itself. Means for voltage tuning of the oscillator frequency, for use in tracking or frequency modulation applications, have also been devised, and possible frequency synthesizers which can be programmed to operate at any of a number of equally spaced frequencies with high short-term stability show promise. Consideration is being given to the use of this oscillator in systems operating at frequencies up into the X band microwave range by multiplying the surface wave oscillator frequency, where it appears that signal purity equal to or better than that achievable by other existing types of signal sources is available in a simpler and less expensive device.

Instrumentation

As stated earlier, dispersive delay lines can be used to extract the complex Fourier transform of an unknown signal. If the ultimate properties of long delay lines can be brought to bear on this problem, there is the prospect of performing 1000- to 10 000-point Fourier transforms with execution times of the order of a couple of milliseconds and accuracy corresponding to 10-bit digital processing, in very compact monolithic devices which might eventually be much less expensive than digital processing devices. Much less ambitious forms of these devices are applicable as portable network analyzers, for impulse testing of UHF devices in the field. Tapped delay lines are applicable as waveform generators. When recursive tap interconnections are employed, they can

be employed as pseudorandom code generators with very long cycle times, for use in secure communication systems or in component testing. The surface acoustic wave controlled oscillator can be applied as a simple, sensitive strain gage, operating through the dependence of total time delay through the delay line on mechanical strain on the delay line surface. These are only a few examples in the instrumentation field, where the small size, high speed, accuracy, and low cost potential of surface wave devices are favorable for a range of applications.

Integration of Surface Acoustic Waves and Microelectronic Components

The integration of surface acoustic wave devices with semiconductor microcircuit elements is an area of very substantial interest. Such integration was first applied to tapped delay lines. The coding of the taps in Figure 5 can be switched electronically by semiconductor switches rather than hard wiring all of the tapping transducers to a fixed sum line. Engineering design work has been done on matching and electrical properties of tapping transducers and semiconductor switches. A further step involves full integration, yielding monolithic devices in which the surface acoustic waves propagate on the same planar surface as is used for the diffusions and depositions of the accompanying electronic elements. Since usual microcircuit substrates are nonpiezoelectric, it has become important to develop means for exciting surface acoustic waves on nonpiezoelectric surfaces. In one approach, a crystalline sapphire substrate wafer has been used, containing side-by-side epitaxial depositions of piezoelectric aluminum nitride for a surface wave tapped delay line and silicon for the semiconductor switching elements [10]. Another approach, which has achieved a substantial amount of success, involves transducers consisting of a sandwich of deposited interdigital electrodes and sputtered zinc oxide piezoeleotric thin films on the surfaces of silicon and other nonpiezoelectric materials [11]. There are optimum geometries and ranges of film thickness which can produce good efficiency and bandwidth. Once the surface acoustic wave is launched on the silicon substrate, various approaches are possible for electronically interacting with the wave. For example, a piezoresistive interaction in the gate region of FET structures can be used as the basis for switchable taps in transversal filters [11]. Another procedure for excitation of surface waves on nonpiezoelectric substrates consists of bonding small segments of piezoelectric crystal onto the extreme ends of nonpiezoelectric delay line plates and using standard interdigital transducers deposited on the piezoelectric regions. With proper attention to the bonds and through the use of various bridging techniques which have been studied, efficient transfer of the wave from the piezoelectric to the nonpiezoelectric areas can be achieved [12].

SURFACE ACOUSTIC WAVE AMPLIFIERS AND CONVOLVERS

Surface acoustic waves which propagate along the surface of a piezoelectric material have an electric field association with them. Thus, it is possible for a surface acoustic wave propagating along a piezoelectric material to interact with the carriers in a semiconductor placed close to the piezoelectric substrate. For this reason, a variety of active devices can be constructed which make use of electron interactions with surface acoustic waves. We will review these acoustoelectric devices in this section and discuss some of their possible future applications, the history of their development, and some of the more interesting research being carried on in this field at the present time.

Amplifiers

When surface acoustic wave devices were first developed, it was realized that the interaction of these waves with drifting carriers in a semiconductor could be useful because it should be possible to construct an amplifier for surface acoustic waves. It had already been shown theoretically and confirmed experimentally in a number of experiments on bulk wave devices at Stanford and elsewhere that when the applied field is large enough so that the carrier velocity exceeds the velocity of the acoustic waves, the carriers can

deliver energy to the acoustic waves. The acoustic wave amplitude increases along its path, while there is attenuation in the reverse direction.

There is a need to obtain internal amplification in acoustic devices because, if the losses in the transducers are high or for that matter the losses in the rest of the system are high, the dynamic range of the operating system is limited but can be increased by the use of internal amplification. External amplifiers can provide only limited relief, because there is a limit to the input power level because of breakdown in the input transducer and the saturation effects in the delay lines.

The first demonstrations of a surface acoustic wave amplifier were made by White [14] at Berkeley, using a piezoelectric semiconductor, cadmium sulfide. Unfortunately, cadmium sulfide has the disadvantage of poor reproducibility and poor semiconductor properties; a large amount of power is needed to make the electrons drift at a high enough velocity to obtain amplification. The group at Stanford realized that it would be necessary to use a semiconductor of better quality, but materials that combine a strong piezoelectric coupling coefficient with good semiconducting properties, however, are not easy to find. They circumvented this difficulty by placing the semiconductor very close to the surface of a piece of lithium niobate along which the Rayleigh wave was passing.

Lakin et al. [15], in their experiments, used spacer rails between the silicon and lithium niobate. These were films of silicon monoxide approximately 500Å thick, deposited on the lithium niobate, as illustrated in Figure 11. A small press was used to push on the semiconductor in order to keep the spacing between the semiconductor and the lithium niobate uniform. The semiconductor consisted of a film of epitaxially deposited silicon about 1 μm thick, deposited on a sapphire wafer. The device produced a very large net amplification, as much as 80 dB/cm over a broad band of frequencies.

Kino et al. [16] developed a theoretical model of the device, which was in excellent agreement with the experimental results, as can be seen from the comparison given in Figure 12. Gains of as high as 60 dB were observed with high attentuation in the reverse direction. Thus, the device had the important property that it was nonreciprocal, so

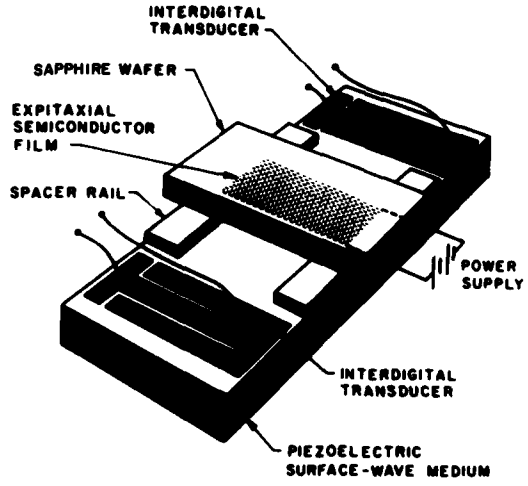


Figure 11—A schematic of an "airgap" amplifier with silicon on sapphire spaced from $LiNbO_3$ by thin SiO rails

that signals reflected from the output transducer could be attenuated, thus tending to eliminate the so-called triple transit echoes.

The device did not receive wide acceptance initially, because of its high dc dissipation, so making it difficult to run on a CW basis and because of the mechanical difficulties of its construction. In further work, Kino et al. made an amplifier with indium antimonide vacuum deposited directly on $LiNbO_3$ [17]. This layer was only 500Å thick, so it provided very little mechanical loading. Such devices were operated at frequencies up to 1.6 GHz. Using a narrow strip of InSb only 25 μm wide, which functioned as a waveguide, it was possible to operate an amplifier on a CW basis, basically because heat spreads sideways as well as down into the substrate [18]. However, the technology proved to be a difficult one, and, although tried in several laboratories throughout the world, has not yet gained wide acceptance.

An alternative route has been to work with another type of vacuum-deposited semiconductor material CdSe, which has a very high resistivity, though a lower mobility than InSb. A particularly interesting version of this device is one constructed by Solie in which the dc potential is applied alternately between an interdigital array of metal fingers laid down on the semiconductor, so the applied dc potential is relatively low [19].

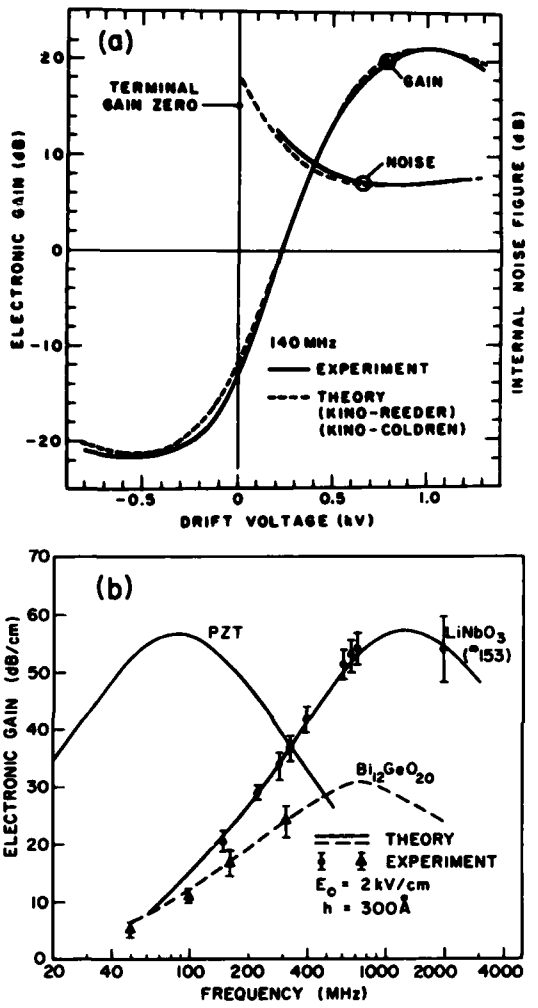


Figure 12—(a) Comparison of theory and experiment for electronic gain and noise figure vs drift voltage in the surface acoustic wave amplifier. (b) Electronic gain vs. frequency in the surface acoustic wave amplifier

This implies that the fields between the fingers are in opposite directions so that the device alternately gives gain and attenuation. However, under certain conditions, the attenuation is less than the gain so that the device becomes a reciprocal amplifier which can just make up for the losses in the system. The nonlinear properties of this device are particularly interesting and have led to a useful new type of efficient and accurate acoustic convolver.

Thus, the applications of acoustic amplifiers still await development of a technology in which CW devices can be made easily and repeatably. Two approaches have been used to improve the technology. One employed by Ralston at Lincoln Labs is an improvement of the original airgap technology [20]. Here a number of posts are etched into the $LiNbO_3$, as illustrated in Figure 13, each post having a diameter of the order of 5 μm. A silicon on sapphire substrate is pushed against these posts. As the posts are so small, they

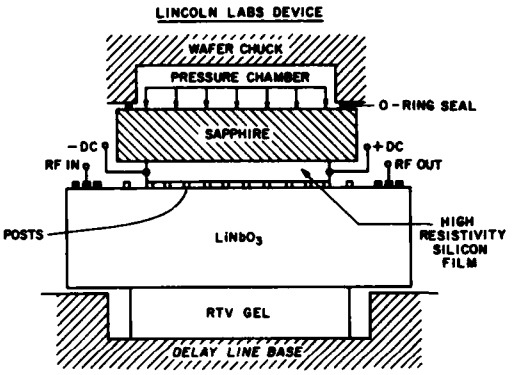


Figure 13—A schematic of Ralston's post-supported "airgap" amplifier. A similar configuration is employed for post-supported convolvers.

provide very little mechanical loading on the surface acoustic wave and do not affect it. Ralston constructed operating CW devices with gains as high as 50 dB between the terminals and demonstrated noise figures of the order of 6–7 dB's, in good agreement with the theory of Coldren and Kino [16]. Theoretically it would be expected that, with a good trap-free material, the noise figure could be reduced to approximately 5 dB at the acoustic input to the amplifier. Nevertheless, despite the improvements, this technology is still an airgap technology. Although a very stable device can be constructed, the precision with which the individual components must be made is severe, so its cost is high.

An approach which would seem more campatible with existing integrated circuit technology is the use of a silicon substrate with ZnO, a piezoelectric material, deposited on it by RF sputtering

techniques. Tarakci and White have demonstrated such a device, using oxide RF sputter deposited on top of a silicon on spinel substrate [21]. This approach should make it possible to construct viable surface acoustic wave amplifiers with desirable characteristics. Further research on this technology remains to be done.

With such technology in hand, switches, amplifiers, mixers, and external storage devices could be combined with SAW devices without the necessity of using a hybrid technology, a requirement which, because it limits the number of interconnections severely, limits the flexibility of SAW signal processing systems severely. As another example, by combining the SAW technology and integrated circuit technology, it becomes possible to make SAW transistor amplifiers in which two acoustic beams are coupled by means of thin metal strips deposited across them, with amplifiers placed in a break in the path of the strips [22]. Another possibility is the use of pn junctions as SAW interdigital transducers, as has been demonstrated by Khuri-Yakub at Stanford [23]. Such transducers can easily be switched because they are sensitive to light and applied dc potentials. There are many other possibilities of this nature which await the full development of the ZnO on Si technology, as well as full use of silicon integrated circuit technology in acoustic wave devices [24].

Acoustic Convolvers, Storage Correlators, and Optical Imaging Devices

We will now review the acoustoelectric interactions associated with nonlinear effects. Because of the highly nonlinear relation between the current and the field in a semiconductor, nonlinear acoustoelectric interactions between an acoustic wave and the semiconductor can be relatively strong. This makes it possible to devise various parametric types of devices. An important class of such devices are the so-called convolvers and correlators; these take the product of two signals and form the convolution or correlation integral of the signals. A recent and perhaps the most important development of this principle is a device which can store signals entering it and take the correlation of the stored signal with a later signal. We will place the main emphasis in this article on the storage correlator, because we believe that it will eventually be the most useful of the convolver type of device, due to its many possible applications to signal processing in radar and sonar systems and because this is the part of the field where considerable research remains to be done.

Another application of acoustoelectric interactions is associated with imaging. As carriers can be generated within a semiconductor when it is exposed to light, the nonlinear interactions of an acoustic wave with a semiconductor can be influenced by the presence of light [22]. By this means, it is possible to utilize an acoustic pulse to scan one line of an optical image formed in a semiconductor. By using more complicated scanning waveforms, it is possible to obtain spatial transforms of the optical images in real time, a process which is difficult to accomplish directly in other types of optical imaging devices. Spatial Fourier transforms of an optical image have been demonstrated, and the inverse transform of this image was also obtained by using surface acoustic wave Fourier transform techniques. The use of this technique has the advantage that, by gating the transform in time, certain spatial frequencies in the image can be eliminated, for instance, background illumination, and, by bandpass filtering the transform, parts of the picture can be eliminated without deteriorating the definition. The technology required for these devices is almost identical to that required for the storage correlators, the only additional requirement being that of transparent electrodes. So, as one device is developed, the performance of the other one improves too. Therefore, we will limit the rest of our detailed discussion to a description of the storage correlator devices [22].

There are closely related devices in which the surface acoustic waves do not interact directly with the semiconductor but instead are sampled by means of taps along the delay line. The signals from these taps are read out into separate diodes or amplifiers, and the basic mixing, integration, storage, or convolution processes that are required can be carried out in external components [22]. One application of these principles is to use the tapped surface acoustic wave delay line as a phase reference, and utilize it for imaging acoustic waves sampled by an array of transducers, one

transducer to each tap. Such devices have been demonstrated in this laboratory to have applications to acoustic imaging for scanned real-time sonar systems, nondestructive testing, and medical diagnostics. The devices were, in fact, first developed with the sonar application in mind and have produced excellent high definition acoustic images [22].

In order to describe the principles of operation of the convolver, we first consider a simple piezoelectric surface wave device in which there is no semiconductor present but in which there can occur a nonlinear interaction between two surface acoustic waves propagating along the surface of the substrate. We suppose, initially, that there are two CW RF signals of frequency ω inserted at each end of the delay line. The acoustic signals at any point z along the device will be of the forms $\exp j\omega(t - z/v)$, and $\exp j\omega(t + z/v)$, respectively, where v is the acoustic velocity. Suppose now that there are nonlinear interactions between the two signals due to the nonlinear properties of the substrate. Then, a second-order product signal will be generated with a variation of the form $\phi(t,z) = exp\ 2j\omega t$. This potential $\phi(t,z)$ does not vary with z, and can be detected between metal films laid down on top and bottom surfaces of the piezoelectric substrate, as illustrated in Figure 14.

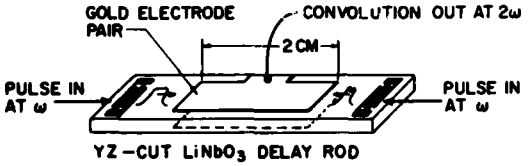


Figure 14—A degenerate convolver with the output transducer consisting of metal films deposited on top and bottom of surfaces of the piezoelectric substrate

When the two input signals are modulated and have the forms $F(t)\ \exp j\omega t$ and $G(t)\ \exp j\omega t$, respectively, the output transducer integrates the induced potential over its length. So, in this case, the convolver can be shown to yield an output of the form

$$V(t) \sim e^{2j\omega t} \int F(\tau)G(2t - \tau)d\tau .$$

This result will be recognized as similar to the convolution of the two input signals, although the output signal is compressed by a factor of 2 in time; this is because the two surface acoustic waves pass by each other at twice the acoustic velocity. It will be recalled that, when a signal is passed into a filter, the output is the convolution of the signal and the impulse response of the filter. In the convolver, because the reference consists of another signal, it is possible to change the reference or the filter response at will. Thus, the convolver is, in principle, an extremely flexible device and may be used to recognize digital codes like Barker codes or pseudorandom codes consisting of long pulse trains or analog codes, such as linear FM chirps. Such demonstrations were made at Stanford in bulk wave devices by Quate [25] and by Shaw [26] and in surface waves devices by Otto [27] and by Kino [26]. The reader is referred to Cafarella [28], Defranould [29], and Solie [19] for some of the more recent results of this type.

The basic problem with the convolver which utilized nonlinear interactions in the substrate material is the weakness of the nonlinear coupling and its low output and dynamic range. A simple and excellent approach to improve this characteristic is to increase the power density and, hence, the acoustic wave amplitude by confining a narrow acoustic beam in a waveguide configuration. This technique has been demonstrated very successfully by Defranould [29] who has obtained a 20 dB increase in convolution efficiency over that of a simple convolver. An alternative approach is to increase the strength of the nonlinear interaction by making use of the nonlinear response of a semiconductor coupled to the RF electric fields of the acoustic waves propagated along the piezoelectric delay line.

The configuration which has received by far the most attention for use as a semiconductor convolver is of the type shown in Figure 15. It will be seen that the basic construction is very similar to that of the acoustic amplifier [22, 26]. However, now the interaction is essentially between the electric field E normal to the surface of the semiconductor and the carriers in the semiconductor; this produces a depletion layer at the surface. Typically, a relatively thick semiconductor layer, thicker than the layer used in the acoustic am-

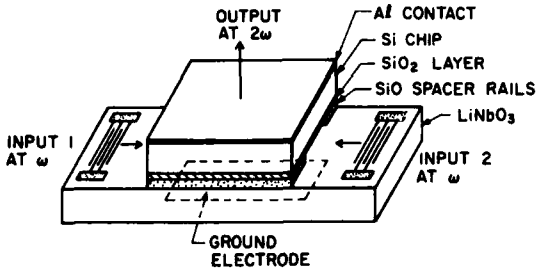


Figure 15—A schematic of an "airgap" silicon convolver spaced by SiO rails

plifier, is employed so that the tangential field component at the surface tends to be shorted out.

Semiconductor depletion layer theory leads to the conclusion that, with a donor density N_d, a potential $\phi = \epsilon E^2/2qN_d$ is developed at the surface of the semiconductor. Thus, the potential formed across the depletion layer at the surface is proportional to the square of the field and varies inversely with the donor density. It is as if the semiconductor behaves as a distributed varactor, with a considerably stronger nonlinearity than can be obtained in the piezoelectric material itself. Normally in this device the potential generated across the depletion layer at any point is proportional to the product of the two input signals. The output is detected between an electrode on the lower surface of the piezoelectric material, capacitively coupled to the surface of the depletion layer and an electrode on the top surface of the semiconductor. Convolvers of this type have been used to take the convolution of Barker codes, pseudorandom codes, and analog codes such as linear FM chirps.

At the present time, the state of technology is such that the airgap devices developed at Lincoln Labs using an improved configuration, like that shown in Figure 13, can operate with input signals at a center frequency of 300 MHz and a bandwidth of 100 MHz and a delay time of 12 μs through the device. This corresponds to a time-bandwidth product of 1200, or the possibility of convolving signals of approximately 1200 bits. The efficiency of these devices is 25–40 dB better than a simple convolver on $LiNbO_3$, corresponding to 60–70 dB's of dynamic range, with maximum input signals of 20 dBm [28, 30].

Storage Correlator

The most recent development in acoustoelectric devices is the storage correlator. This device makes use, in its different forms, of one of several possible storage mechanisms such as storage in surface traps or bulk traps, in diodes, or by charging from an electron beam.

Electron beam storage in SAW devices, of which we shall not describe the details here, was first demonstrated by Bert et al. in France [31] and excited a great deal of interest to devise simpler techniques for the same purpose, using semiconductor technology. At about the same time, Quate at Stanford had demonstrated an optical imaging device which made use of storage in surface states of the semiconductor [32]; he and his coworkers had also demonstrated storage effects in Schottky barriers laid down on the surface of a GaAs semiconductor convolver [33].

Using this work as a basis, almost simultaneously Bers and Cafarella [34] at MIT and Kino and Hayakawa [35] at Stanford were able to demonstrate a new type of device which could store a surface acoustic wave signal in surface states for times up to several milliseconds; this could be read out after a delay of up to several milliseconds, or the correlation of the stored signal could be taken with a later signal read into the device.

The storage in surface states proved to be an unreliable mechanism. So later developments have made use of storage in Schottky diodes or pn diodes laid down on the surface of the semiconductor [36, 37]. Such devices are operated as a convolver in the manner already described, but now the interaction takes place in the buried depletion layers of the diodes, thus eliminating the effect of surface states.

In order to understand the operation of this device, consider the configuration shown in Figure 16 with a row of Schottky barriers or pn diodes laid down in the surface of the semiconductor. Suppose the silicon is pulsed negative with respect to the grounded film underneath the substrate, as illustrated in Figure 17. In this case, the diode would be forward biased and the charge it would receive would be $Q = C_1V$, where C_1 is the capacity of the diode to ground and V the applied potential. If now the pulse were removed, the

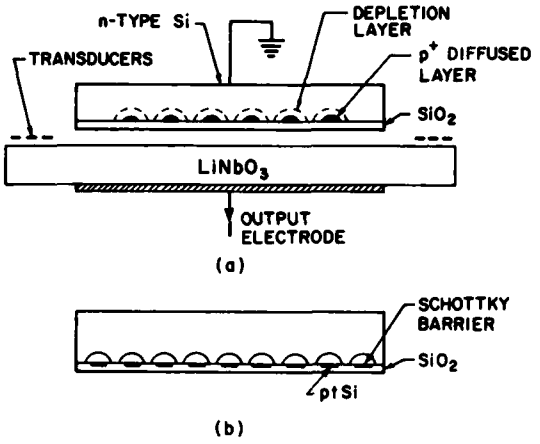


Figure 16—(a) A convolver in which p⁺ layers are diffused into an n-type substrate. The nonlinear interaction occurs in the junction depletion regions. (b) A convolver with Schottky barriers laid down on a semiconductor substrate.

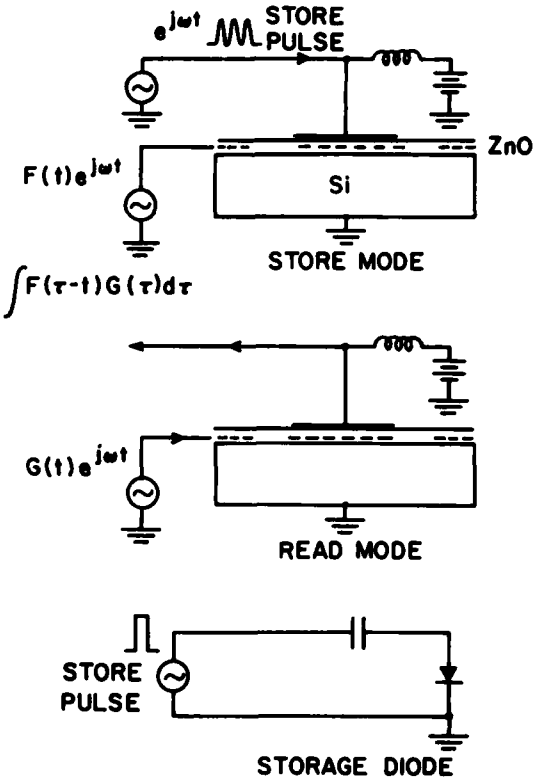


Figure 17—Readin and readout in the storage correlator. Top is readin, accomplished through the nonlinear interaction between the plate signal and the surface wave. The readout is taken from the plate as the nonlinear interaction between the stored charge pattern and the readout surface wave signal. Also shown is a simplified equivalent circuit model of the storage diode.

diode would be reverse biased and the only way the charge could leak away from this capacitor would be through the leakage current of the diode.

More generally, if a pulse has been applied to the device and there is a surface acoustic wave traveling along the surface, the total potential at the diode will depend on the sum of the potentials due to the surface acoustic wave and the applied pulse, as will the stored charge when the device is forward biased. Thus, if a signal $F(t)$ exp $j\omega t$ is inserted into the convolver, it will excite a surface acoustic wave pulse which varies as $F(t - z/v)$ exp $j\omega(t - z/v)$ along the device. At the same time, a short RF pulse of frequency ω is applied between the output plate of the convolver and the semiconductor; this excites an RF field which varies as exp $j\omega t$. The nonlinear interaction between this signal and the surface acoustic wave gives rise to dc terms which vary as $\cos(\omega z/v)$. Thus, a signal of the form $F(t)$ exp $j\omega t$ inserted into the device will give rise to a variation in stored diode charge and, hence, potential across the diodes of the form $F(z/v)$ cos $(\omega z/v)$. The readin time to such a system depends on the time constants for forward biasing the diodes; typically, this is of the order of 1 or 2 ns. The storage time depends on the capacity of the diodes to ground and their leakage currents. Storage times of several seconds have been observed in pn diodes, with storage times of 1–100 ms in Schottky barrier diodes, depending on whether their capacity to ground was increased with the use of electrodes with an excess capacity to the semiconductor or the electrodes were left out.

The stored information in the diodes may be read out by using a reading signal which has the same spatial periodicity. More generally, if a modulated RF signal is applied at one transducer, the output signal obtained from the plate is the correlation of the reading signal $G(t)$ and the original stored signal $F(t)$. If the signal $G(t)$ is read into the other interdigital transducer, the convolution of the stored and reading signal is obtained. It will be noted that, unlike the convolver, the refer-

ence signal does not have to be read in at the same time as the signal to be interrogated. It can be read in within the storage time of the device, which can be in the range of a few microseconds to a few seconds, depending on the design of the device.

An example of the use of such devices could be to employ them in a sonar or radar system. Suppose, for instance, that a coded signal is emitted from the sonar and reflected from an object. The received signal is then stored in the correlator storage device. If a later signal from another part of the object or from a more distant object were received and then correlated with the earlier signal, the distortion due to the errors in the system itself or due to inhomogeneities in the ocean could be removed, for both the reference and the signal of interest would have suffered the same distortion. Such correlations would take place in real time, a considerable advantage in a sophisticated sonar system.

At Stanford, in work partially supported by the Joint Services Program, we have demonstrated this process in a sonar type of system. We used an acoustic transducer in water with a center frequency of 3.5 MHz and a bandwidth of 2.5 MHz excited by a linear FM chirp, as shown in Figure 18 [38]. The acoustic pulse excited by this transducer was reflected from a metal pllate in the water and received at the transducer; after mixing the output up to a frequency of 100 MHz, it was stored in a storage correlator. A later echo, the so-called triple transit echo, was then correlated with the first one. Using a high quality transducer with an almost ideal pulse response, as shown in Figure 19, we obtained a correlation of the type shown in Figure 19. After replacing the transducer with a transducer with a much poorer response, one that rang for several cycles, as shown in Figure 19, we again correlated the reference echo with a later echo. It will be seen that the correlation peak obtained with both transducers had approximately the same width. Thus, the effect of the poor response of the second transducer could be virtually eliminated.

There are many other possible applications of this type of processing. For instance in signal

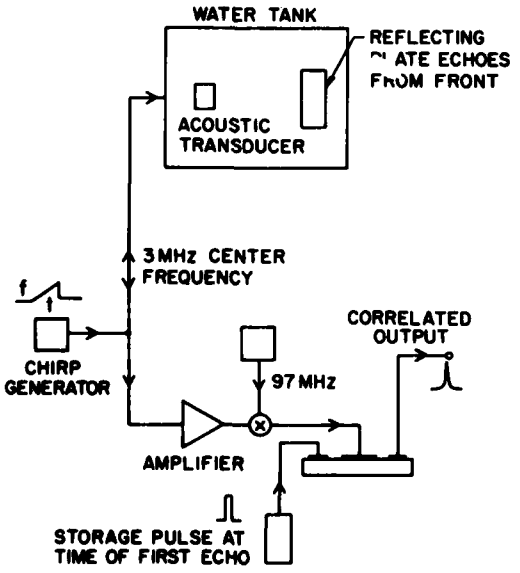


Figure 18—Schematic of the acoustic pulse echo system

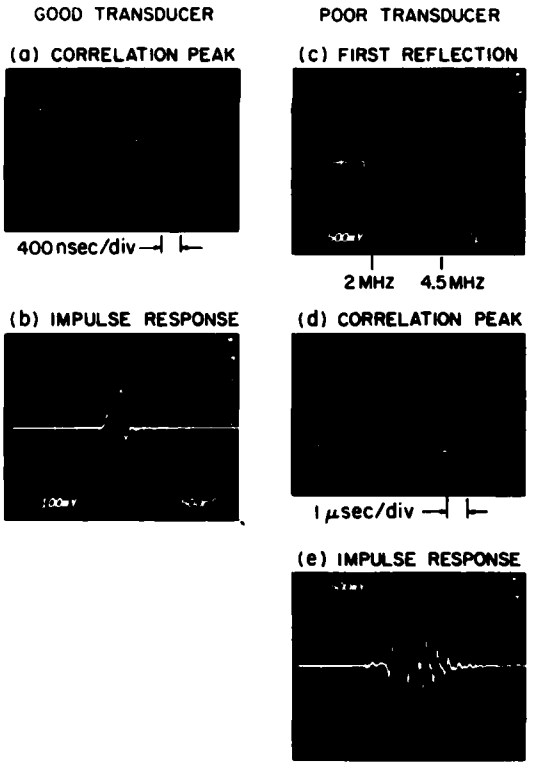


Figure 19—Pulse echo experiment results with both good and poor transducers

processing, if a short pulse were to be transmitted through a distorting path and stored in a correlator, this stored reference could be correlated with an unknown signal and used to remove some of the distortions. Other possibilities involve the correlation of very large time-bandwidth product signals, corresponding to the product of the storage time (0.1–1 s) and bandwidth (10–100 MHz) of these devices, and two-dimensional storage, correlation, and transforms using surface acoustic waves propagated at right angles to each other. Such devices should be able to store signals with very large time-bandwidth products in the 10^5–10^6 range.

At the present time, these devices are still relatively crude. Feedthrough of unwanted signals is a problem, and optimization of the efficiency, dynamic range, and time-bandwidth product another. The theory of operation which has been developed is still extremely crude, although it gives rough qualitative agreement with the experimental results. It is clear that, even though the required theory is highly nonlinear in character, it can be developed and that this should be a primary aim in further research in order to optimize the characteristics of these devices. In the same way, the technology for making arrays of diodes for the storage devices is required. Basically this is an existing technology that is used for construction of vidicon devices. It needs to be adapted, however, to the requirements of one- and two-dimensional convolver configurations and to obtain control over the readin and storage times. It needs also to be adapted to the ZnO on Si technology, to make a useful monolithic device.

As far as the constructional techniques are concerned, the problems are almost identical to those we have already discussed with reference to the acoustic amplifier. The basic configuration of the device is essentially the same as that of the acoustic amplifier, except that bulk silicon, rather than silicon on sapphire, is required, because it is not advantageous to short out the component of field parallel to the surface, which only causes loss. The history of the technology is that simple silicon oxide spacer rails were used initially between lithium niobate and silicon. Because this approach led to nonuniformities in the spacing, the supporting post technique like that employed for the amplifier was developed. More recently in this laboratory we have developed a technique using thin rails 4 μm wide, 150 μm apart sputter etched into the $LiNbO_3$ and alined along the direction of propagation of the acoustic wave. This configuration gives negligible mass loading and is much easier to make than the multiple post device.

In the same way as the amplifier, we believe that it is imperative to construct a monolithic device for eventual use in the field and for compatibility with other integrated circuit components. For this purpose, we have developed and are continuing to develop a zinc oxide on silicon technology. Convolvers made by this technique have performed well, storage has been demonstrated, but complete storage correlators have not yet been constructed with this technology, although it is expected that they will be shortly. The problems with this technology are associated with the influence on the performance of the device by charge stored in traps in the zinc oxide, a long-term storage effect; the lower piezoelectric coupling coefficient than for the lithium niobate devices, and, hence, the lower bandwidth; and the problem of making the semiconductor processing required compatible with surface acoustic wave technology. By using the convolver configuration with buried pn junctions, many of these difficulties seem to be obviated at the expense of a relatively large number of processing steps. Bandwidth can be increased by carefully designed electrical circuit matching and some improvements in the transducers themselves. These approaches are gradually yielding good results, the present bandwidth of the devices being approximately 20%, there being good agreement between the experimental and theoretical performance.

The buried pn junction technique has itself given difficulties in both the airgap and ZnO on Si configuration because of some conduction due to sideways diffusion of carriers between the junctions. This was eliminated in the airgap devices by use of a DMOS constructional technique, which involves etching away the region between the junctions in the form of a triangular groove. Such an approach cannot be used directly under the ZnO because it would interrupt the acoustic path. One method is to place the junctions ouside the acoustic path and connect to them, by means of deposited metal coupling strips, the so-called

strip-coupled convolver. This approach has the advantage that, because the junctions can have a different width from that of the acoustic beam, the nonlinear effects can be stronger and the convolver be made more efficient, a demonstration already made by Kino and Shreve [22] in airgap convolver devices. Initial results in the ZnO on Si configuration are encouraging but still await further development before it can be determined if this is a viable approach.

A simpler technique is to use polysilicon layers deposited on top of the junctions. This produces a potential well which inhibits sideways diffusion of carriers between the junctions and has eliminated the sideways diffusion problem in our airgap convolvers. It should be compatible with the ZnO on Si configuration, so it will shortly be tried in that system.

One great advantage of these monolithic configurations should be that they lend themselves to two-dimensional storage devices and, for that matter, optical imaging devices far better than do the airgap systems. This is because it is possible to divide up the output coupling film into several strips, to which separate connections can be made, if necessary. So individual parts of the acoustic beam can be sampled separately, by dividing up the convolver output electrode.

Conclusions

Surface acoustic wave amplifiers, convolvers, storage correlators, and optical imaging devices have been demonstrated in the laboratory. Convolvers are beginning to be used in radar and communication systems. The simple convolver which employs a piezoelectric material as the nonlinear element gives very high quality output but is somewhat inefficient, although waveguiding techniques have helped in this respect. Semiconductor convolvers are more difficult to construct, but the airgap type is well developed and gives excellent performance. Airgap types of acoustic amplifiers perform well and are just barely operable on a CW basis. Both types of semiconductor devices need developments in the technology to make them in a monolithic form. The most obvious path to this end is the use of ZnO deposited on silicon or on silicon on sapphire. Other approaches involve vacuum deposition of a semiconductor, such as CdSe or InSb on $LiNbO_3$ [17, 19]; the use of a GaAs [33], a piezoelectric semiconductor; or epitaxial deposition of materials like AlN and Si side by side on sapphire with metallic connecting strips laid down across them [39]. In the authors' opinion, the ZnO on Si approach is probably the most promising, basically because it is compatible with other devices such as switches, amplifiers, and mixers which can be constructed on the same substrate by normal integrated circuit techniques. If the technology is developed in this way, it may well lead to new types of acoustic devices such as acoustic amplifiers incorporating transistors and may lead to new concepts such as the marriage of the CCD devices with acoustic techniques. If this were done, acoustic methods might be used, as an example, for nondestructive readout of the CCD registers.

The storage correlator is a relatively new device which shows great promise in its application to radar, sonar, NDT, and a variety of signal-processing systems. The basic technology is that of the convolver, but it has its own special difficulties associated with the use of arrays of pn or Schottky diodes. All the demonstrations of this device have been made so far in airgap convolver configurations, which are now sufficiently developed to be mechanically stable and repeatable. There is good hope of making these devices in the monolithic form, such as with the ZnO on Si configuration, and the necessary technology is being developed for this purpose. There are several possibilities for further developments involving very large time-bandwidth product correlators. The devices are useful for correlating with any reference read into them. Further developments involving already established ROM techniques may lead to semipermanent memories which can be used to correlate a signal permanently stored in the system with arbitrary signals read into it.

We have not dealt in detail in this article with optical imaging devices based on the convolver principle. Such devices, although they may operate as well as a CCD imager eventually, are hardly worth developing to do only the same job. However, they lead to the possibility of perform-

ing functions which are difficult to duplicate by other means, such as taking one- and two-dimensional spatial Fourier transforms of an image in real time. The basic technology required is identical to that of the acoustic storage correlator. Furthermore, in order to carry out the two-dimensional inverse transform it will be necessary to use a two-dimensional storage correlator. So, we have mainly described the present state of the art of the storage correlator and the research which needs to be done in this field. We would expect that, with the development of the storage correlator, good optical imaging transform devices would automatically follow.

Finally, we have given only a short description of convolver types of devices which use external mixers or storage elements connected to taps on a surface acoustic wave delay line. This is a very fruitful approach which has yielded high-quality convolvers and correlators. It has also led to a completely new class of acoustic imaging devices suitable for applications to sonar, nondestructive testing, and medical imaging which can form excellent electronically scanned and focused acoustic images, without the use of physical lenses. Much work remains to be done in this field, and these applications are being rapidly developed at the present time.

REFERENCES

1. H. E. Bömmel and K. Dransfeld, *Phys. Rev.* **117**, 5, 1245–1252 (Mar. 1, 1960).
2. W. Richard Smith, Henry M. Gerard, Jeffrey H. Collins, Thomas M. Reeder, and Herbert J. Shaw, *IEEE Trans. Microwave Theor. Tech.* **MTT-17**, 865–873 (Nov. 1969).
3. J. J. Campbell and W. R. Jones, *IEEE Trans. Sonics Ultrasonics*, **SU-15**, 209–217 (Oct. 1968).
4. This tapped delay line was constructed by T. W. Bristol and colleagues at Hughes Aircraft Co.
5. R. C. Williamson and H. I. Smith, *Electron. Lett.* **8**, 401–402 (Aug. 1972).
6. Larry A. Coldren and Herbert J. Shaw, *Proc. IEEE* **64**, 5, 598–609, and Iain M. Mason, Emmanuel Papadofrangakis, and John Chambers, *Proc. IEEE* **64**, 5, 610–612 (May 1976).
7. T. M. Reeder, H. J. Shaw, and E. M. Westbrook, *Electron. Lett.* **8**, 14, 356–358 (July 13, 1972).
8. Arthur A. Oliner, *Proc. IEEE* **64**, 5, 615–627 (May 1976).
9. M. F. Lewis, *Ultrasonics*, 115–123 (May 1974).
10. P. J. Hagon, F. B. Micheletti, R. N. Seymour, and C. Y. Wrigley, *IEEE Trans. Microwave Theor. Tech.* **MTT-21**, 303 (Apr. 1973).
11. Fred S. Hickernell, *Proc. IEEE* **64**, 5, 631–635, and Gordon S. Kino, *Proc. IEEE* **64**, 5, 724–748 (May 1976).
12. L. T. Claiborne, E. J. Staples, and J. L. Harris, *Appl. Phys. Lett.* **19**, 58–60 (Aug. 1971).
13. M. T. Wauk and R. L. Zimmerman, *Electron. Lett.* **8**, 17, 439 (Aug. 24, 1972).
14. R. M. White, "Surface Elastic Wave Propagation and Amplification," *IEEE Trans. Electron. Devices* **ED-14**, 181–189 (1967).
15. K. M. Lakin and H. J. Shaw, "Surface Wave Delay Line Amplifiers," *IEEE Trans. Microwave Theor. Tech.* **MTT-17**, 912–920 (Nov. 1969).
16. G. S. Kino and T. M. Reeder, "A Normal Mode Theory for Rayleigh Wave Amplifier," *IEEE Trans. Electron. Devices* **ED-18**, 909–920 (Oct. 1971). G. S. Kino and L. A. Coldren, "Noise Figure Calculation for the Rayleigh Wave Amplifier," *Appl. Phys. Lett.* **22**, 50–52 (Jan. 1973).
17. L. A. Coldren and G. S. Kino, "The InSb on a Piezoelectric Rayleigh Wave Amplifier," *IEEE Trans. Electron. Devices* **ED-21**, 421–427 (July 1974).
18. L. A. Coldren and G. S. Kino, "CW Monolithic Acoustic Surface Wave Amplifier Incorporated in a ΔV/V Waveguide," *Appl. Phys. Lett.* **23**, no. 3, 117–118 (Aug. 1973).
19. L. P. Solie, "A New Mode of Operation for the Surface-Wave Convolver," *Proc. IEEE, Special Issue on Surface Acoustic Wave Devices and Applications* **64**, no. 5, 760–763 (May 1976).
20. R. W. Ralston, "Stable CW Operation of Gap-Coupled Silicon-on-Sapphire to $LiNbO_3$ Acoustoelectric Amplifiers," *IEEE Ultrasonics Symp. Proc.*, pp. 217–222, 1975.
21. U. Tarakci and R. M. White, "Layered Media Active Microwave Acoustic Delay Lines," *IEEE Ultrasonics Symp. Proc.*, pp. 440–445, 1972.
22. G. S. Kino, "Acoustoelectric Interactions in Acoustic Surface Wave Devices," invited paper, *Proc. IEEE, Special Issue on Surface Acoustic Wave Devices and Applications* **64**, no. 5, 724–748 (May 1976).
23. B. T. Khuri-Yakub, private communication.

24. F. S. Hickernell, "D-C Triode Supttered Zinc Oxide Surface Elastic Wave Transducers," *J. Appl. Phys.* **44**, 1061–1071 (Mar. 1973).
25. C. F. Quate and R. B. Thompson, "Convolution and Correlation in Real Time with Nonlinear Acoustics," *Appl. Phys. Lett.* **16**, 494–496 (June 15, 1970).
26. G. S. Kino, S. Ludvik, H. J. Shaw, W. R. Shreve, J. M. White, and D. K. Winslow, "Signal Processing by Parametric Interactions in Delay Line Devices," *IEEE Trans. Sonics Ultrasonics* **SU-20**, 162–173 (Apr. 1973).
27. O. W. Otto and N. J. Moll, "A Lithium Niobate Surface Wave Convoluter," *Electron. Lett.* **7**, 696–697 (1971).
28. J. H. Cafarella. W. M. Brown. F. Stern. and J. A Alusow, "Acoustoelectric Convolvers for Programmable Matched Filtering in Spread-Spectrum Systems," *Proc. IEEE, Special Issue on Surface Acoustic Wave Devices and Applications* **64**, no. 5, 756–759 (May 1976).
29. P. Defranould and C. Maerfeld, "A SAW Planar Piezoelectric Convolver," *Proc. IEEE, Special Issue on Surface Acoustic Wave Devices and Applications* **64**, no. 5, 748–751 (May 1976).
30. J. M. Smith, E. Stern, A. Bers, and J. Cafarella, "Surface Acoustoelectric Convolvers," *IEEE Ultrasonics Symp. Proc.*, pp. 142–144, 1973.
31. A. Bert. B. Epstein, and G. Kantorowicz, "Signal Processing by Electron Bean Interaction with Piezoelectric Surface Waves," *IEEE Trans. Sonics Ultrasonics* **SU-20**, 173–181 (Apr. 1973).
32. C. F. Quate, "Optical Image Scanning with Acoustic Surface Waves," *IEEE Trans. Sonics Ultrasonics* **SU-21**, 283–288 (Oct. 1974).
33. T. Grudkowski and C. F. Quate, "Acoustic-Readout of Charge Storage on GaAs," *Appl. Phys. Lett.* **25**, 99–101 (1974).
34. A. Bers and H. J. Cafarella, "Surface State Memory in Surface Acoustoelectric Correlator," *Appl. Phys. Lett.* **25**, 133–135 (1974).
35. H. Hayakawa and G. S. Kino, "Storage of Acoustic Signals in Surface States in Silicon," *Appl. Phys. Lett.* **25**, 178–180 (1974).
36. K. A. Ingebrigtsen, "The Schottky Diode Acoustoelectric Memory and Correlator-A Novel Programmable Signal Processor," *Proc. IEEE, Special Issue on Surface Acoustic Wave Devices and Applications*, **64**, no. 5, 764–769 (May 1976). K. A. Ingebrigtsen and E. Stern, "Coherent Integration and Correlation in a Modified Acoustoelectric Memory Correlator," *Appl. Phys. Lett.* **27**, 170–172 (1975).
37. C. Maerfeld and P. Defranould, "A Surface Wave Memory Device Using p-n Diodes." *IEEE Ultrasonics Symp. Proc.*, pp. 209–211, 1975.
38. P. G. Borden and G. S. Kino, "Correlation With the Storage Convolver," submitted to *Appl. Phys. Lett.*, GL Report 2586.
39. L. R. Adkins, "Strip Coupled AlN and Si Sapphire Convolvers," *IEEE Ultrasonics Symp. Proc.*, pp. 148–151, 1973.

Arthur L. Schawlow has been Professor of Physics at Stanford University since 1961; he was Chairman of the Department of Physics from 1966 to 1970. After two years as a postdoctoral fellow and research associate at Columbia University, he became a research physicist at Bell Telephone Laboratories. In 1960, he was a visiting associate professor at Columbia University. Dr. Schawlow's research has been in the fields of optical and microwave spectroscopy, nuclear quadrupole resonance, superconductivity, and lasers. With C. H. Townes, he is coauthor of the book *Microwave Spectroscopy* and of the first paper describing optical masers, which are now called lasers. For the latter work, Schawlow and Townes were awarded the Stuart Ballentine Medal by the Franklin Institute (1962) and the Thomas Young Medal and Prize of The Physical Society and The Institute of Physics (1963). Dr. Schawlow also received the Morris N. Liebmann Memorial Prize Award from the IEEE (1964). He gave the AAAS Holiday Science Lectures in Philadelphia (1965), Salt Lake City (1966), and Durham (1967) and was the Richtmyer Lecturer of the American Association of Physics Teachers (1970). He received a Senior Postdoctoral Fellowship from the National Science Foundation for 1970–1971; was the Cherwell-Simon Lecturer at Oxford University in 1970; received the Geoffrey Frew Fellowship from the Australian Academy of Science for 1973; and was California Scientist of the Year in 1973. He wrote the introduction for *Scientific American Readings on Lasers and Light* and three articles in that collection. He has appeared on TV programs broadcast on U.S., Canadian, and British networks. Dr. Schawlow was born in Mount Vernon, N.Y. He received the Ph.D. degree from the University of Toronto in 1949 and honorary doctorates from the Universities of Ghent (Belgium), Toronto (Canada), and Bradford (England). He is a Fellow of the American Physical Society (Member of Council, 1966–1969), the Optical Society of America (Director at Large, 1966–1968), the IEEE, the AAAS, and the American Academy of Arts and Sciences and is a member of the National Academy of Sciences. In 1974, he was Chairman of the Division of Electron and Atomic Physics of the American Physical Society and, in 1975, was President of the Optical Society of America.



LASERS

Arthur L. Schawlow

Stanford University
Stanford, Calif.

EXPECTATIONS

Lasers, in the years since they were first proposed in 1958 and demonstrated in 1960, have become ubiquitous tools of science. Nearly every issue of any scientific journal contains at least one report of some research in which a laser was used.

Yet, even now, lasers are not really as commonplace as some enthusiasts had predicted. This is hardly a new situation. Almost as soon as any lasers existed, they were hailed as the realization of the ancient literary dream of an all-destroying energy ray. That concept might be traced to the legends of Archimedes destroying an enemy fleet by focusing reflected sunlight on the sails. In Francis Bacon's *New Atlantis* of 1627, the inhabitants of his utopia intensified light beams and transmitted them over long distances. In H. G. Wells' 1898 novel *War of the Worlds*, the Martians almost conquered Earth with a sword of light. In the 20th century Alexei Tolstoi wrote *The Hyperboloid of Engineer Garin*. Ray guns became standard equipment in futuristic comic strips like Buck Rogers in the 1930s.

Remarkably, as physicists in the 20th century learned more about how light is emitted and absorbed, the less likely did such a device appear. Light is emitted when some atomic system makes a transition from an excited state to a lower one with less stored energy. But the emitted light can be absorbed by atoms of the same kind if they are initially in that lower state. No matter how hot the substance is heated to emit light, in thermal equilibrium there will always be more atoms in the lower state so that absorption will predominate. Thus the radiation emitted by a hot substance is limited to no more than that of a perfect absorber (black body) at that temperature, and amplification of the radiation is not expected. The individual atoms of a thermal radiator emit independently with random phases. Thus, in the 1960 edition of a widely used physics textbook it was stated, with italics for emphasis, that "It is not possible to make two parts of the same light source coherent." But, unknown to the textbook's authors, the theoretical foundation for a device doing just that, the laser, had been laid and several different kinds were operated within that very year.

The way had been led by the microwave molecular amplifier called the maser, which was invented by Charles H. Townes in 1951 and first successfully operated by J. P. Gordon, H. J. Zeiger, and Townes in 1954. This was, incidentally, one of the early important discoveries supported by the Office of Naval Research and other defense agencies, through the joint program of microwave physics research at Columbia University.

After masers of several kinds were developed and in use, Townes and Schawlow gave serious

consideration to extending the maser principles to the much shorter wavelengths of visible light. In so doing, we were trying to continue the historic search for ways of producing shorter radio waves. This was in itself a sufficient incentive. If we thought at all about applications, we had in mind such possible uses as communications, spectroscopy, and photochemistry. In seeking a resonator suitable for optical masers, we realized that two small widely spaced mirrors facing each other would select one or a few of the many modes of oscillation possible in the resonator large enough to contain an adequate number of atoms. With such a resonator, the output of an optical maser (laser) would be a narrow beam.

Thus it was that the first laser used this kind of structure and produced narrow, highly directional beams of light. Moreover, the very first of these, the pink ruby laser built by T. H. Maiman, delivered a peak power of several kilowatts. Even though the duration of the ruby laser's output pulse was less than a millisecond, this was enough to spark a revival of dreams of devastating energy rays.

Thus, from the beginning, lasers were confronted by high expectations which the early ones could not come close to fulfilling. It was soon shown that a pulsed ruby laser could vaporize a sample of even the most refractory substance but only a very small sample and the laser had to be focused onto the target at short range. If the lasers were not all-destroying energy rays, they were equally unsuited for use in other envisioned applications like spectroscopy and photochemistry. Each individual laser had its own characteristic wavelength and could be tuned only over a small fraction of that wavelength. Thus available lasers were of no use for studying the most interesting atoms.

It was not surprising, therefore, that the laser was soon called "a solution in search of a problem." Yet this never was a fair description of what lasers could do. Marvelous as they were, the early lasers were very far from being able to do the important tasks that were waiting. Nor could sustained high power, high efficiency, or tunability be obtained by any amount of refining of the laser designs or scaling them to different sizes. Each kind of laser was individual, with characteristic properties determined by the active material and the method of excitation. To get different wavelengths or higher energy, one had to find different materials. The early lasers were not solutions to the real problems of technology. Rather, they were just a hint as to where solutions might be sought.

Despite enormous progress and great achievments in the intervening years, something of the same basic difficulty remains. Even though there are far more applications of lasers than we can consider here, some reasonable applications are still not matched to suitable lasers. Let us consider a few of these, to see how things are now and how the past history illustrates the prospects.

KINDS OF LASERS

There are already many different types of lasers ranging in size from semiconductor diodes so tiny as to be almost invisible up to giant gas and glass lasers big enough to fill a large building. Wavelengths generated range from the radio frequencies through infrared, visible and ultraviolet, into the vacuum ultraviolet or soft X-ray region near 100 nm. Continuous-wave power outputs of lasers used in research and industry range from microwatts up to about 10 kW. In short pulses, peak powers up to about 10^{13} W are delivered. At less extreme power levels, pulse lengths shorter than 10^{-12} s have been obtained and measured.

This wide range of characteristics reflects an equally wide range of laser types. They may be classified by their method of excitation as optically pumped, gas discharge, gas dynamic, chemical, photochemical, semiconductor diode, and electron beam (acceleration). Any of them has the ability to convert some kind of energy into highly organized, monochromatic, directional light energy. There are in principle no thermodynamic limitations on the conversion from one ordered kind of energy to another, as from electricity to radio waves or to laser light. But most lasers include at least one disordered, thermal stage in the conversion process, so that their efficiency is most often low. A few efficient lasers are known, but they are still exceptional.

The first lasers were optically pumped, and this method of excitation is still used in many lasers.

The working substance may be a solid, liquid, or gas. Indeed so many substances have been made to lase with optical excitation that it has been claimed that anything will lase if you excite it vigorously enough. If the substance hasn't been made to lase, it has not been hit hard enough. That claim is undoubtedly exaggerated, but optically pumped laser action is obtained in a very wide variety of materials.

In any of them, absorption of light, from the pumping source, excites atoms or molecules to energy levels from which emission can be stimulated. Usually the frequency, and so the quantum energy, of the output light is less than that of the pump light. This is inevitably a cause of inefficiency but not necessarily a major one. Typically, both input and output are near the visible region and so this quantum efficiency exceeds 50%.

Much more serious is the inefficiency in generating the pumping light and coupling it to the amplifying medium. Lamps generally produce light with a wide range of wavelengths, only a small fraction of which can be usefully employed in exciting atoms to the desired energy level for laser action. Thus, all present optically pumped lasers have low efficiencies, at most a few percent and often much less.

They could be made efficient by the discovery of efficient light sources suitable for pumping them. For instance, if there were a bright enough light source in the near ultraviolet region it could drive a dye laser at any wavelength in the visible or near infrared region. The overall efficiency could be high if the source is efficient and intense enough. But such sources remain to be discovered despite considerable efforts and ingenuity. Nevertheless, there is reason to hope that they will be found. Some gas discharge lamps, such as those with mercury or sodium, can convert a large fraction of the electrical input to light in a fairly narrow band of wavelengths. Semiconductor light-emitting diodes and diode lasers can be fairly efficient, although they are so far limited to small sizes and comparably modest powers. Perhaps most encouraging is the existence of carbon dioxide lasers which can have overall efficiencies of some tens of percent. But these lasers emit a wavelength far out in the infrared (around 10.6 μm) and so are not immediately suitable for pumping visible lasers. Perhaps some new kind of gas discharge laser will be found which could optically pump other lasers over a wide range of wavelengths.

Efficiency has been emphasized here because it is an essential requirement for any device which is to generate sustained high power. More serious than the cost of the wasted power is the difficulty of dissipating the large amounts of heat produced by energy wasted inside the laser. Even at moderate power levels, thermal distortion of the laser material harms its optical qualities.

Very high peak powers can be obtained in pulsed operation if the pulse is short enough so that not too much energy is involved. As pulse lengths have been shortened from microseconds to nanoseconds and then to picoseconds, peak powers have risen correspondingly from kilowatts to megawatts, gigawatts, and even terawatts. High energy delivered in a very short pulse had been more difficult to achieve. At present it is possible to deliver only a few hundred joules in a pulse of subnanosecond length. Much higher energies are needed for experiments on thermonuclear fusion, and they are being sought by using many laser amplifiers in parallel.

Some problems encountered in very high power lasers arise from the effects of intense light on the laser medium itself. The index of refraction is slightly higher at the places where the intensity is high. Light is refracted toward regions of higher refractive index. Thus any initial nonuniformity of intensity is then enhanced by self-focusing at high power levels, which in turn accentuates the differences in refractive index. Thus, the process can run away until the intensity at the induced foci is great enough to cause permanent damage, as by an electric spark.

So, to keep the intensity down, one must use amplifier stages of large cross-section, usually in the form of large slabs. All of the high peak power lasers use neodymium ions in some glass. The requirements for a glass of high optical quality, low self-focusing, reasonable strength, and thermal conductivity present severe challenges to materials scientists.

Liquid Lasers

Some liquids can be just as transparent as solids. They can contain strongly fluorescing sub-

stances, and they can be used for optically pumped lasers. For high-intensity operation, a liquid has the obvious advantage that any structural damage quickly heals. Liquid flow can permit quick and very effective cooling. However, with few exceptions, liquids show large changes of refractive index for small temperature changes. Unless the heating from the exciting lamps is very uniform, the liquid will thus become optically inhomogeneous and spoil the laser beam quality. Self-focusing and stimulated scattering also can be troublesome for operation at high intensities.

Most liquid lasers make use of dilute solutions of organic dyes. Quite generally these dyes have broad emission bands, so that laser amplification can be obtained over a substantial range of wavelengths, typically a few hundred Angstrom units in the visible region. The dye laser can be made tunable within this region by incorporating a wavelength selector. For instance, one of the mirrors can be replaced by a diffraction grating which acts as a good reflector for only one wavelength at a given angle of incidence. As the grating is rotated, the laser's output wavelength is tuned. In a simple pulsed dye laser, this simple tuning method might produce bandwidths of an Angstrom or more. For narrower lines, other tuning elements can be added until the line width is limited only by the pulse duration.

Pulsed dye lasers can be pumped by flashlamps or by lasers of shorter wavelength. Nitrogen lasers have been widely used for pumping and have produced dye laser action at wavelengths from about 350 nm in the ultraviolet, throughout the visible, up to about 1 μm in the infrared. Continuous-wave tunable lasers using dyes have been pumped by argon or krypton ion lasers through much of the visible region. With care and some refinements, continuous-wave dye lasers can be made extremely monochromatic and stabilized to better than 1 part in 10^9. The most monochromatic lasers are, however, tunable over only a very small wavelength without readjustment.

In the future, we may hope to have widely tunable (perhaps over the entire visible region) but highly monochromatic continuous-wave laser signal generators. Perhaps they may come by refinement of dye lasers, possibly with automatic changing of dye cells for different parts of the spectrum. On the other hand a wholly new approach may be found, as there are many laboratory uses for such sources.

Gas Discharge Lasers

Probably the most widely used lasers up to now have been gas discharge lasers. Low-power helium-neon lasers, emitting in the red at 633 nm, are used everywhere for alinement and surveying. This was one of the earliest lasers to be proposed, having been suggested by Ali Javan in 1959 and first operated in the near infrared by Javan, W. R. Bennett, Jr., and D. R. Herriott in 1960. Visible operation was achieved by A. D. White and J. D. Rigden in 1962.

Since then the design has been simplified and refined to reduce the cost so that simple helium-neon lasers can now be obtained at a retail price around $100. It has been estimated that in mass production they could be made to sell at a price closer to $10. Large-quantity production would require a large market for a single design. It might be warranted by any of such proposed applications as playback devices for commercial video recordings or supermarket checkout scanners. If that happens, the same laser could find many other uses. For instance, it would make an excellent pointer for use when photographic slides are projected. Whenever things need to be placed in a straight line a small laser is a nearly ideal alignment aid, even for mundane tasks like carpentry, masonry, or gardening. For these purposes the laser need not be powerful—indeed it should be weak enough so that it is manifestly safe. It needs only to be a cheap, and reasonably rugged, visible laser.

Very many other gases can be made to lase in an electrical discharge, either continuous or pulsed. In the visible and near ultraviolet, argon and krypton ion lasers are commercially available with power outputs up to 20 W or so. Each of them can be made to oscillate at any of several wavelengths. Krypton, especially, spans the visible. These lasers give only a few particular wavelengths, not complete coverage of the spectrum. For purposes such as Raman spectroscopy or making holograms, that is quite sufficient. Moreover, they can be used to pump tunable dye

lasers, which can be tuned to any wavelength in this spectral region.

At first it may seem ridiculous to use one laser to pump another, because you compound their inefficiencies. Although the argon and krypton lasers indeed have very low efficiency—typically the light output is less than a thousandth of the electrical power input—they can be focused to give the high intensities needed for laser action in the dye. Often the advantage of tunability is important enough to outweigh the low efficiency.

But for some other applications, the low efficiency of ion lasers is a serious disadvantage. For instance, cutting and drilling metals and photochemical processing could use much higher powers, and efficiency is an important consideration.

Really high continuous powers so far are obtainable only from carbon dioxide lasers. Here the conversion efficiency may be about 30%, and thousands of watts of laser power can be generated continuously. However, carbon dioxide lasers emit far in the infrared region, near 10.6 μm. They can be used to pump other lasers for longer wavelengths but not shorter. All of these infrared wavelengths are well absorbed by most insulating substances like wood, cloth, or stone but not by metals. Many carbon dioxide lasers are in use for processing materials.

If the laser's intensity is high enough, even a metal, which absorbs only a few percent of the light, can become very hot. enhanced absorption does occur at high intensity levels, so that there is a power range suitable for metal cutting, welding, or hardening. Nevertheless, the existence of a threshold intensity for absorption makes the conditions for using CO_2 lasers more critical than they would be for a laser of visible or shorter wavelengths. Moreover, the long wavelength cannot be focused as sharply as visible light. For these reasons optically pumped neodymium crystal lasers are also used for cutting and welding, even though they are less efficient than carbon dioxide lasers.

At first it was widely believed that there was little advantage in pulsing a gas laser. In a helium-neon laser the output saturates at fairly low currents. Moreover, the density of atoms is thousands of times less than in many solid laser materials. Such a low-density gas could store only a small amount of energy for release in a short pulse. But for many other gas lasers the situation is quite different, and they can compete with solid lasers even for high peak power outputs. Moreover, some gases can provide laser action only in short pulses.

In pulsed operation, as in the continuous-wave mode, carbon dioxide is one of the most important laser materials. With a transverse discharge, at pressures around atmospheric or even higher, high-power pulses can be generated with lengths from nanoseconds to microseconds.

While some laser gases can be operated at moderately high pressures, others require high pressures. Among these are the excimer lasers. For example, xenon is an inert gas which does not form molecules with other xenon atoms. However, when a xenon atom is raised by electron impact to an excited state, it can bond to a neighboring atom to form the excimer molecule Xe_2^*. It will then radiate spontaneously in the vacuum ultraviolet region around 1700Å, but the ground state of the molecule is not bound and the atoms fly apart. Thus, there are no absorbing molecules in the ground state, and so any excited molecules contribute to the optical amplification by stimulated emission.

A high gas density, typically around 15 times atmospheric, is needed to ensure that an excited xenon atom will find a partner and form a molecule in the brief instant before it loses its excitation. It is difficult to make gas discharges work at such high pressures, and so the energy is supplied by a high-current pulse of fast electrons through a thin metal window.

Excimer laser action is also obtained in xenon and krypton fluoride. Closely related is the process in argon fluoride lasers, where the lower state is bound, but only weakly so that it quickly dissociates. All of these require very high current densities, so that even those which can be run as discharges have been operated only in short pulses.

With all these and other types of gas lasers, there is still none capable of generating sustained high power in the visible region, and none is even in sight. Despite much work in the field, many possible systems remain unexplored. Among these are many of the more refractory metal vapors, largely because they are difficult to handle. There are indeed some metal-atom and metal-ion

lasers, notably cadmium and copper. The latter is reasonably efficient, but it lases in short pulses. Repetition rates as high as 100 000 per second and average powers of some watts have been attained with copper. While copper lasers can be made larger, one must still hope for something more efficient and scalable to large sizes.

Even apart from problems with the laser medium and its excitation, high-power visible lasers are plagued by problems with their end windows and mirrors. Visible or ultraviolet light from the laser beam can produce color centers in most transparent materials, leading to increased absorption in them. More materials research is needed to understand this damage and to find ways to prevent it.

Serious materials problems are also encountered in rapidly modulating or controlling lasers of power output more than a few watts. Laser beams can be deflected or modulated at high speeds by electro-optical cells whose refractive index changes when a voltage is applied. But most electro-optical materials cannot withstand high optical powers. High powers can be controlled, but relatively slowly, by mechanical deflectors or choppers. Intermediate speeds and power capabilities can be obtained with optoacoustic deflection. In this method, an intense sound wave through a liquid produces a density grating that diffracts the light through an angle which depends on the wavelength of the acoustical vibrations.

What might we do when we have efficient, high-power continuous-wave lasers in the visible or ultraviolet region and techniques for fast control of their output? There are evident needs for rapid generation of complex patterns on metals, such as in making printing plates and cylinders. Since one blue or ultraviolet laser can drive dye lasers of several selected colors, it could permit bright, large-screen displays for television or computers.

Some scientific experiments also need such a laser. For example, positronium (the atom made of an electron and a positron) is the simplest of all atoms. Its spectrum should be exactly calculable and so measurements could provide searching tests of quantum electrodynamics. But despite heroic efforts, the wavelength of even the strongest line has been measured only approximately, although its fine structure has been resolved in an ingenious experiment by A. P. Mills, S. Berko, and K. F. Canter. The difficulty with positronium is that the atoms live for only about a hundred nanoseconds after their formation, as the constituent electron and positron can annihilate each other. Thus as positrons are emitted from a radioactive source, find electrons to form positronium, and decay, there is never a time when there are many positronium atoms present. Thus, high laser power is needed to have a good probability of exciting a positronium atom before it disappears. Moreover, a continuous wave is needed, because the positronium atoms are produced at random times as the positrons are emitted. Two-photon excitation of positronium without doppler broadening should be certainly possible when suitably powerful lasers become available at a wavelength of 4860Å.

Other important potential needs are for photochemistry. It has long been realized that lasers could provide a new kind of control over chemical reactions, but these could hardly even be explored with the early primitive lasers. Now lasers can be tuned finely enough to excite a single isotopic species in a mixture of molecules and make it reactive without affecting the other isotopes. Since isotopes are difficult and expensive to separate by any other method, excitation with even the present inefficient lasers may be practical for some substances. Separated uranium and hydrogen isotopes have important uses in nuclear energy generation. So far, economical laser-induced separation of these isotopes appears possible but difficult partly because of low laser efficiencies. However, very simple ways have been found to separate some other isotopes, most notably for chlorine by R. N. Zare. If correspondingly easy ways to separate uranium isotopes were discovered, it might lead to dangerous proliferation of nuclear explosives. Perhaps it is fortunate that the known methods are complex and difficult.

It is also possible that laser light of a particular wavelength may be able to activate a chosen bond within a molecule, so as to cause a reaction at that site and not elsewhere. This is even more difficult than isotope selectivity, because many highly excited molecules very quickly distribute their energy among the many other possible electronic and vibrational modes. Still, it is already known

that lasers can affect reactivity in ways quite different from simple heating, especially in pulsed decomposition. It is intriguing to speculate on how lasers might be able to alter biological molecules and processes, but there is little information even to suggest an appropriate direction to investigate.

Lasers can also be used for a different kind of photochemistry—spatially rather than primarily wavelength selective. That is, one can make a chemical reaction take place where one wants it. For instance, one could make a liquid plastic solidify at selected places, by using laser light to induce polymerization. Moreover, intense light of a longer wavelength can induce this hardening by two-photon absorption. Thus, solidification could occur only where the light is most intense, for instance, at a place where two beams are focused together. By moving the beams under computer control, a three-dimensional object could be constructed. This last process can be thought of as a generalization of photography which would be practical when, through lasers, light is both abundant and cheap.

Chemical Lasers

Even as lasers can be used in chemistry, so chemical reactions can be used in lasers. From antiquity until about 1900, chemical reactions in flames were the main source of light other than the Sun. However, in most flames the pressure is fairly high and reactions occur slowly so that conditions are never far from equilibrium. For laser action, it has been necessary to use rapid reactions so that molecules can be excited to a particular upper level faster than they relax by collisions. Thus, HF can be vibrationally excited as it is produced in a reaction between hydrogen and fluorine initiated by an electron beam or discharge and stimulated to emit near 3 μm in the infrared. Most of the energy comes from the chemical reactants and it can be released in a very intense, short pulse. Powers of the order of 10^9 W have already been reported and it seems possible that the very high powers and energies needed for thermonuclear fusion research may be attained. Continuous-wave action has also been obtained in flowing-gas chemical lasers but so far not at very high power levels. For portalbe applications, chemical lasers can produce large amounts of energy from a moderate weight of fuel.

Gas Dynamic Lasers

Closely related to chemical lasers are gas dynamic lasers. In them, a gas is heated and then allowed to cool by rapid expansion through a nozzle. If the cooling processes are such that some lower state is depopulated faster than an upper state, laser action occurs. Very high continuous-wave powers can be obtained from a carbon dioxide gas dynamic laser. It may be that gas dynamic lasers will be useful for some large-scale industrial applications, although they are more complicated to control than electrical discharge lasers.

Semiconductor Lasers

The smallest lasers are the semiconductor diodes. They can be fairly efficient and by suitable choice of materials can operate over a wide range of wavelengths from the visible far into the infrared. Their brightness is high, but the emission occurs only at the small area of a thin junction between two kinds of semiconductors. Thus the power output is limited in comparison with other lasers that can be scaled to large sizes.

Some degree of tunability can be achieved by applying mechanical stress, temperature changes, or magnetic fields to the diodes. The approximate wavelength is adjustable over wide ranges by varying the composition of the semiconductor materials.

Semiconductor lasers are likely to find very wide application in communications. In addition to their advantages of compactness and efficiency, their output can be modulated rapidly by controlling the current through them. Especially, they will be used to feed low-loss optical fiber communications links. The semiconductor diodes can be coupled directly to the fibers, or they can be used as optical pumps for small neodymium lasers. The latter provide output at a wavelength of 1.06 μm which is a nearly ideal match to the most favorable wavelength for low-loss transmission in quartz fibers.

Optical fibers can provide broadband voice, data, and picture communications over distances up to a mile (1.6 km) directly or as far as desired with the use of semiconductor repeaters. They are very light and compact so that they are well suited for internal communications in ships or airplanes and in densely populated cities. Widespread use of optical fibers for communications seems assured, and it is likely that they will use very large numbers of semiconductor diode lasers.

Electron Beam Lasers

A radically different class of laser, which has promise for producing high power outputs and being very broadly tunable, was proposed by John Madey and is being investigated at Stanford University's High Energy Physics Laboratory. A beam of very fast electrons from a superconducting linear accelerator is passed through a region of transverse magnetic field whose direction is rotated helically around the beam axis. The rapidly moving electrons passing through the magnet experience a strong, high-frequency field which sets them into transverse oscillation so that they radiate an electromagnetic wave. The wave's frequency is determined by the rate at which the electrons traverse the corrugations of the helical field and so by the beam energy. Not so evidently, it can be shown that there is optical amplification which can produce laser action. The electron beam energy must be very sharply defined, as only a superconducting accelerator or a storage ring can provide. Amplification occurs with electrons fast enough so that relativistic effects are important. Typically electron energies are in the range of 10–1000 million eV.

Optical wavelength γ, at which the electron beam emits radiation is given approximately by $\gamma o/\gamma^2$ where γo is the wavelength of the magnet's field alternations (3.2 cm in the present model), and γ is the ratio of the electron's energy to its rest mass (approximately 0.5 MeV). This factor comes from the relativistic length contraction. To the electron, the magnet periodicity appears reduced by a factor γ. The radiation emitted by the moving electron in the forward direction is reduced in wavelength by an additional factor γ. Thus infrared at 10.6 μm is obtained at the electron beam energy at 24 MeV, while 1000Å in the vacuum ultraviolet region would be generated at less than 300 MeV.

For high continuous-wave power, it would probably be best to circulate the fast electrons around the storage ring. Magnets would bend the electrons so that they circulate repeatedly through the wiggler magnet. Buildup of the stored beam can take place slowly, over some minutes, until a current of perhaps 0.5 A of 100 MeV electrons circulates. At each pass, a small fraction, say 0.25% of the beam energy would be extracted as stimulated emission. But since the stored energy is very large, and the electron bunch passes through the wiggler magnet something like 10^7 times per second, the average, quasi-continuous power output would be of the order of 100 kW.

A laser that can be electrically tuned anywhere from the infrared to the short ultraviolet region and give such a large power output is an exciting prospect. To be sure it is still at an early stage and many of the properties remain to be verified experimentally. But the properties of free electrons, even though the theory must be quantum mechanical and relativistic, are more surely calculable than those of any substance.

Lasers in the Extreme Ultraviolet and X-ray Regions

Pulsed gas lasers have been operated throughout the ordinary ultraviolet region and even to wavelengths much shorter than air will transmit. Quite simple repetitively pulsed hydrogen lasers generate wavelengths down to 1200Å. However, they require quite intense excitation with a short pulse of high current density, around 10 000 A/cm^3. As will be discussed later, lifetimes of excited states generally become shorter at shorter wavelengths. For this and other reasons, the required pump power density is expected to rise sharply as the wavelength is decreased.

Any powerful laser can generate optical harmonics in substances whose dielectric constant changes with the electric field strength. Second harmonics, at twice the laser frequency or half the wavelength, are produced in crystals that lack a center of symmetry, like quartz or ADP (ammonium dihydrogen phosphate). However,

nearly all crystals are opaque at the shorter wavelengths and so harmonic generation in crystals has not produced wavelengths shorter than 2000Å.

Gases, however, can generate harmonics of shorter wavelengths. A gas is symmetrical for a reversal of direction and so cannot produce second harmonics, but it can generate third or other odd-order harmonics. Usually, the nonlinear optical coefficients get smaller the higher the order of harmonic. Thus for a given laser power third harmonics tend to be weak compared with second harmonics when both can be generated. But, as Stephen Harris has pointed out, very high focused laser intensities can be used in gases, and near-resonances can enhance the effects. With pulsed operation and multistage harmonic generation, Harris has obtained wavelengths near 800Å. Starting with a xenon laser at 1709Å focused into argon gas, M. H. R. Hutchinson, C. C. Ling, and D. J. Bradley obtained third-harmonic radiation at 570Å. This is well into the middle of the vacuum ultraviolet/soft X-ray region. Further progress by harmonic generation seems possible, even though few substances are at all transparent in this region.

X-Ray Lasers

In extending atomic oscillators from microwave masers to lasers in the visible region, it was easiest to jump over the far infrared where nearly everything absorbs and little was known and go directly to the visible region where there was much more information. Perhaps the same may be true with the extensions of lasers to shorter wavelength, at wavelengths below a few Angstroms, where substances become more transparent again, and we are in the familiar region of ordinary x-rays.

It is tempting, therefore, to speculate that laser action might next be achieved in the true X-ray region. However, the obstacles are formidable enough that we cannot yet see where solutions will be found. For one thing, no substance is even nearly as transparent as glass and air are for visible light. A thickness of 1 mm of tin, an element of medium atomic weight, reduces the intensity of 0.1Å x rays by a factor of 2.4 and of 1Å X-rays by a factor of 10^{27}. So for real transparency we would need to operate at an even shorter wavelength. But no atom or molecule, not even uranium, can emit X-rays shorter than 0.1Å from transitions between bound states. So, if we are to use atoms at all in an X-ray laser, they must be very highly excited to give a large gain per atom. Preferably they should be ionized to remove the extra electrons that do not contribute to the desired radiation but can absorb it.

Apart from any considerations of particular atoms and radiating process, it appears that very intense excitation will be needed for any x-ray laser. In part this is because of the short lifetime of excited S states at short wavelengths, but also more energy must be supplied for excited atoms. Another factor is the increase in line width, whether from doppler broadening or short radiative lifetimes, that causes a reduction in gain per excited atom and a corresponding increase in the number of excited atoms needed. From all of these factors, it appears that the required pumping power density may be expected to rise roughly as (1/wavelength)^5. Thus, reducing the wavelength from the visible around 5000Å to 1Å would require an increase from 1 W/cm^3 which is typical in the visible to $(5000)^5 = 3 \times 10^{18}$ W/cm^3.

This is such a high power density that, when we first thought about lasers, it seemed quite unattainable. However, it is well within the range that can be attained by focusing a high-power pulsed laser, such as those used for nuclear fusion research. Electron beams, ion beams, or intense electric discharges could deliver the very intense excitation needed for X-ray laser action. To calculate whether and how such a burst of energy would be concentrated in a single high-excited state is very complex and difficult. Calculations support the likelihood of laser action in at least the soft X-ray region around a few hundred Angstroms. Most probably laser action will be attained there first and subsequently extended to the ordinary X-ray region.

Since the quest for an X-ray laser has been difficult, one might well ask what uses it would have. Yet, in doing so we must keep in mind that the most important uses will probably not be foreseen in advance. Clearly an X-ray laser will be very different from anything known before. The intuitive feeling for what is possible, on which inventions are usually based, will have to be de-

veloped. Most especially, the uses will depend on what sort of a device it is—how powerful, directional, and monochromatic; whether it is pulsed or continuous; and how short is the output wavelength.

If the wavelength is in the ordinary X-ray region, around 1Å, it could be used to reveal the structure of crystals and molecules. Possibly an X-ray diffraction pattern could be obtained in a nanosecond or less, thus making it possible to study crystal forms created momentarily during shock compression.

It also seems possible that holograms could be made which would directly display the positions of the atoms in complex molecules such as those important in biology. This would not be easy, as each X-ray quantum has enough energy to eject electrons from even the innermost atomic shells and thereby damage the molecule. Moreover, the important light elements, carbon, nitrogen, and especially hydrogen, do not scatter X-rays strongly. Nevertheless, scattered X-rays can be detected with great sensitivity and so it may be possible to get enough coherently scattered X-rays to produce a hologram.

A more modest but perhaps very useful application of coherent X-rays could be for phase-contrast radiography. This might well be a useful way to provide better contrast in X-ray photographs of organisms or human tissues.

The destructive capabilities of X-rays are well known, and so one might consider using X-ray lasers as radiation weapons. There are no really good reflectors known for X-rays, and so it seems impossible to devise reflective shielding for defense against X-rays. However, even at a wavelength as short as 1Å, air is absorptive enough that the range would be only about 10m—roughly the same as a lance! In outer space, however, there would be no such restriction. Moreover, it is theoretically possible that a narrow, intense X-ray beam could bleach a path through the atmosphere.

The disruptive ability of an X-ray laser might be harnessed in other ways. If the laser is finely tunable, it might be able to break chemical bonds selectively and thus alter a chosen part of a molecule. It would surely be interesting to study how a molecule fragments after absorbing X-rays of various wavelengths.

The short wavelength of X-rays, thousands of times less than that of visible light, could permit correspondingly sharper images. X-rays might compete with electron microscopes for studying specimens which would be damaged by being put in a vacuum. Of course that assumes the existence of very high quality lenses or other focusing devices and it is not easy to see how those will be made.

An easier imaging application of an X-ray laser might be for projecting masks onto semiconductors for photoetching of tiny electronic microcircuits. In the complex integrated circuits used for fast computers, it is necessary to minimize the time taken for signals to travel from one circuit element to another. Thus even smaller sizes and finer patterns are needed so that visible light is not short enough to produce them. X-ray lasers could produce very small patterns, but again imaging by electrons is a competitor.

Gamma Ray Lasers

Very short electromagnetic waves are emitted by many radioactive nuclei, both natural and artificial. These gamma rays cover the wavelength range of X-rays and extend beyond it to still shorter wavelengths. As early as 1963 several scientists suggested that, as gamma rays are really the same as X-rays where their wavelengths overlap, their emission could be stimulated. Thus it might be possible to make an X-ray laser by using a supply of radioactive nuclei to provide the excited states. It was soon realized, however, that any excited nuclei which last long enough to be stored, can give very little amplification. This follows because both spontaneous and stimulated emission depend on the strength of coupling between the nucleus and an electromagnetic field and so if one is weak the other is also. Moreover, the nuclei are normally in atoms, whose electrons can absorb the gamma radiation, so that a considerable amplification is needed. There have been a number of studies of this problem, and some ingenious ways have been suggested to suddenly excite a large number of nuclei and get them into the proper configuration. But it is not yet certain how or when or even whether a gamma-ray laser can be built. But it has not been proven to be

impossible. The large number of ingenious ideas already proposed even gives some reason to be hopeful that further progress may lead to gamma-ray lasers.

Tunable Lasers and Spectroscopy

Much of all we know about the nature of matter has come from studying the wavelengths of light absorbed or emitted by various substances—atoms, molecules, solids, nuclei, and plasmas. This is what physicists mean by spectroscopy. To analytical chemists, spectroscopy provides a very sensitive analytical technique for identifying and measuring small amounts of substances through the characteristic absorption, emission, or scattering spectra. Both of these broad aspects of spectroscopy are now being revolutionized by tunable lasers. This revolution has far to go, even though some of the results are already spectacular.

Previously, a spectrograph of some kind was always used to sort out the wavelengths of light emitted or absorbed by the material being studied. But with tunable lasers, as was done earlier with radio frequency and microwave oscillators, we can tune the source of radiation and thus probe at different wavelengths without the need for a spectrograph. As the laser is tuned, we need only record the transmission at the various wavelengths.

In the infrared region, all other sources are so weak that very little radiation can be obtained in a narrow band. The resolution of infrared spectroscopy, that is, its ability to distinguish absorptions differing in wavelength by small amounts, was always limited by the weakness of infrared sources. Even a small, low-powered laser like a semiconductor diode can emit far more radiation within its narrow bandwidth than the hottest thermal source. Diode lasers can be tuned by altering the materials used in their construction by varying the temperature, pressure, external magnetic field, or even the current through the diode. They have been used to resolve fine structures in the spectra of molecules and for detecting pollutant gases in the atmosphere.

Some other widely used tunable infrared lasers make use of spin-flip Raman conversion in a semiconductor, pumped by a gas laser and tuned by a powerful magnetic field. In the far infrared region, broadband laser amplification and tunable oscillation can be obtained from gases such as methyl fluoride pumped by a shorter wavelength infrared laser. For the shorter infrared wavelengths close to the visible region, optical parametric oscillators pumped by fixed wavelength are widely tunable. They even extend into the visible part of the spectrum, overlapping the range of dye lasers. The latter, with various luminescent dyes repetitively pulsed by a powerful source like a nitrogen laser, can generate any wavelength from the near infrared around 1 μm to the near ultraviolet around 3500Å. About half of this range is covered by continuous-wave dye lasers. Still shorter tunable laser wavelengths, approaching 2000Å, can be generated as optical harmonics in suitable crystals.

Thus, tunable-laser absorption spectra can be obtained at nearly any wavelength in the infrared, visible, or ultraviolet portions of the spectrum. But this coverage requires a number of very different devices, which are probably not all to be found in any one laboratory. A universally tunable optical signal generator seems quite remote. However, such a device would be so useful that we can expect the search for new kinds of tunable lasers to continue. Perhaps it will come from a new type of laser, like the electron beam laser, that is inherently tunable. On the other hand, computer control may make practical complex lasers that adjust or interchange many parts as the wavelength is shifted.

But, even now, lasers can do much more than just scan absorption spectra. Lasers are often intense enough that they appreciably alter the properties of a substance that absorbs their light. Whenever a quantum of light is received, the abosrbing atom is raised to an excited state and is momentarily incapable of absorbing any more of the same radiation. Usually, the atom quickly reverts to its original state. With ordinary light only a negligible fraction are excited and so the absorption coefficient is not appreciably altered by the presence of the light. But a laser can saturate a transition so that another beam probing at nearly the same instant may find the substance less absorbing than before.

This ability of a laser to tag those atoms or molecules which have absorbed its light permits

laser spectroscopy to probe more deeply than ordinary light. For instance, laser saturation spectroscopy can eliminate the doppler broadening of spectral lines caused by the thermal motions of atoms or molecules in a gas. In the method introduced by T. W. Hänsch and C. Bordé, the light from a tunable laser is split into two beams which are directed through the gas sample in opposite directions (Figure 1). The stronger beam is interrupted periodically by a mechanical chopper.

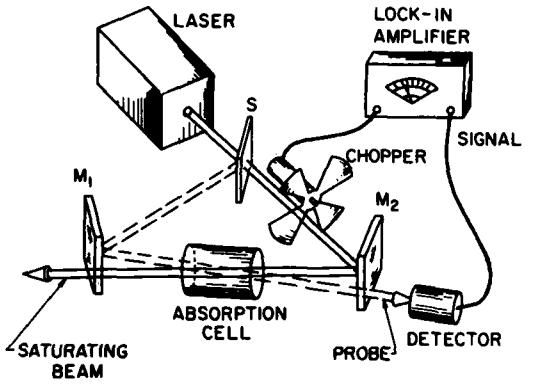


Figure 1—Schematic diagram of laser-saturation method for observing spectra without doppler broadening

Whenever this saturating beam is on, it can bleach a path for the other beam by saturating the absorption. However, this only happens if the two beams interact with the same atoms, which can only be those which are not moving along the line of the beams. To most atoms, which do have a component of velocity along them, the beams appear to have different frequencies because of the doppler shift and those atoms cannot be resonant simultaneously to both beams. Thus, this saturation method picks out those atoms for which the beams have no doppler shift. As the laser is scanned, across a band of wavelengths, fine details of the spectrum are revealed, which would otherwise be obscured by the random doppler shifts in the absorption of light by atoms moving in many different directions. For example, in the spectrum of molecular iodine, hyperfine structures from the interaction of the two iodine nuclei with the molecule were resolved for the first time. Individual spectral lines were found by Hänsch, Levenson, and Schawlow to have as many as 21 components, all clearly resolved, with individual components having line widths less than 1 part in 100 million (Figure 2). In hydrogen the Lamb shift was resolved optically for the first time (Figure 3).

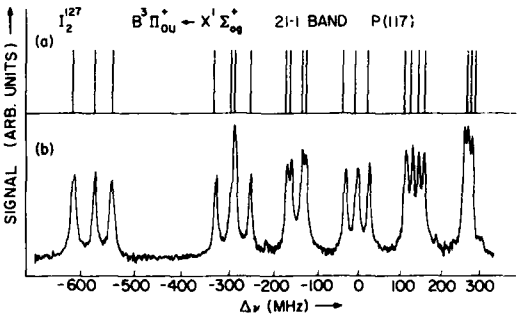


Figure 2—Hyperfine structure of a single line in the visible spectrum of iodine, revealed by saturation spectroscopy. On this scale the visible part of the spectrum would be 18 mi (28.9 km) wide

Two variants of saturation spectroscopy provide spectral lines equally sharp and free from doppler broadening. Moreover, they are more sensitive and can be used for even smaller numbers of atoms and molecules. In the polarization spectroscopic method introduced by Hänsch and C. Wieman, the saturating beam is polarized. Saturation then reduces the absorption and refraction for light of the same polarization but not for the orthogonal polarization. Then if the probe beam is polarized differently, it will be partially depolarized on passing through the medium. An analyzer can be set so that the probe beam is rejected except at those wavelengths where it is depolarized by interacting with the saturated atoms. Thus the signal is seen without a large background and can be observed sensitively at low gas density and relatively low laser power.

Absorption of light often leads to fluorescence from the state excited, and this fluorescence can be used to indicate that abosrption has occurred. When the absorption is saturated, the fluorescence intensity is less than linearly proportional to the laser's intensity. This nonlinearity can be used

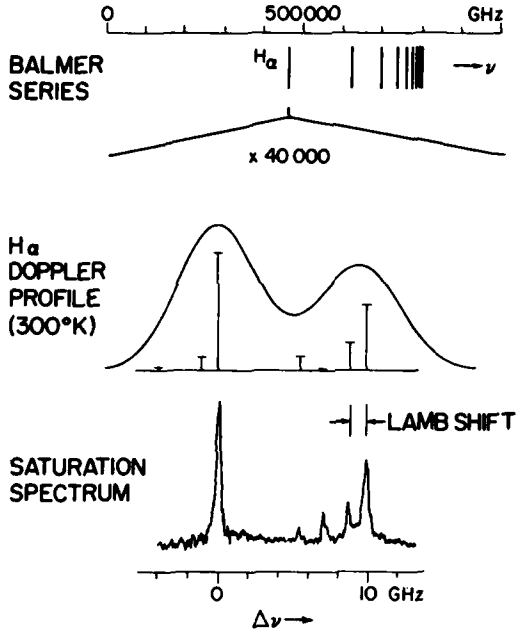


Figure 3—Comparison of ordinary and saturation spectroscopy in observing the spectrum of hydrogen and the fine structure of the H_α line (from T. W. Hänsch and M. H. Nayfeh)

as an indication of when the two beams from opposite directions are tuned so as to work together to saturate the atoms, that is, when they are tuned to the atoms which have zero doppler shift. Then, if the two oppositely directed beams from the tunable laser are chopped at different frequencies, say 1000 and 2000 Hz, the saturated, nonlinear fluorescence shows a component at the sum frequency, 3000 Hz in this case. This intermodulated fluorescence method was used by Sorem and Schawlow to resolve iodine hyperfine structures at a vapor pressure as low as one mTorr, a thousand times lower than could be reached with the saturated-absorption technique.

How few atoms could be seen by the saturated fluorescence technique? Probably very few indeed, although the technique has not been pushed to its limit. Fairbank and Schawlow were able to observe and measure fluorescence of sodium atoms excited by a tunable laser, at temperatures as low as −30°C where the density of atoms is only about 100/cm³. The signal-to-noise ratio would be good enough for spectroscopy at densities nearly as low.

Laser-induced resonance fluorescence could be an extremely sensitive method of detecting small amounts of any element in the vapor phase. However, for most atoms, ultraviolet radiation would be needed. The laser would have to be quickly and accurately tunable to the wavelength for each kind of atom to be analyzed.

Two-Photon Spectroscopy

The high intensity of a laser beam can be used in other ways for high-resolution spectroscopy. An atom or molecule can be put into a condition such that it can absorb another wave. The two beams cooperate to produce a "two-photon transition," in which the quanta of energy absorbed from the two beams add up to the energy needed to raise the atom to a particular excited state. If the two beams come from opposite directions and have the same wavelength, as they would if split off from the same laser, the doppler shifts for any moving atom are always equal and opposite. Thus in the sum of their frequencies, the Doppler shifts cancel out. There is a single, very sharp two-quantum resonance, to which all atoms contribute regardless of their motion. The doppler-free saturated-absorption, polarization, and fluorescence methods, on the other hand, select out just those few molecules which happen to be not moving along the beam direction. However, the methods are complementary, for the spectrum lines studied in two-photon spectroscopy cannot be observed at all in either ordinary or saturated absorption.

One particularly interesting application of two-photon spectroscopy is to study the 1S to 2S transition hydrogen. This atom, the simplest of all stable atoms has for a century served to provide searching tests of atomic theories and to lead the way to improved theories. It is unusual in that the one member of the first group of excited states, 2S, has the same symmetry as the ground state, 1S. For this reason, transitions between them cannot be made by absorbing or emitting a quantum of light. Thus, the 2S state holds its stored excitation for a phenomenally long time. Its excitation lifetime is more than a tenth of s second, a

hundred million times longer than the neighboring 2p state. The 2S state can be reached, however, by a two-photon transition using laser beams in the ultraviolet with wavelength near 2430Å.

With two laser beams of that wavelength, oppositely directed to eliminate the large doppler broadening of these very light atoms, narrow lines have been observed. It happens that the $n = 2$ to $n = 4$ transition, commonly called H_β, in hydrogen is at a wavelength of 4860Å. Thus, a laser of that wavelength was used to scan the H_β line. Part of the 4860Å light was doubled in a crystal to produce 2430Å ultraviolet, which then induced two-photon transitions from the 1S to the 2S state. Since the same laser was the source for both the 1 to 2 and the 2 to 4 transition, these wavelengths could be compared precisely (Figure 4).

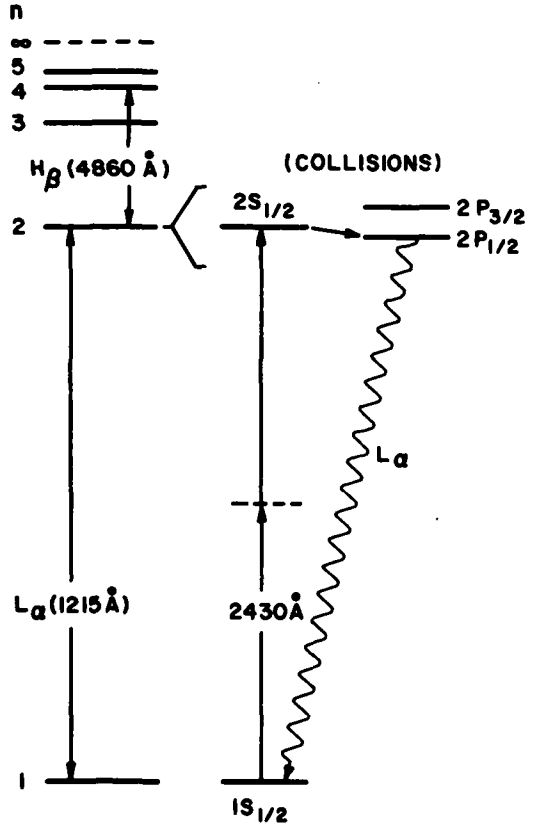


Figure 4—Energy levels of hydrogen and transitions studies in measuring the ratio of energy level spacings 2–1 and 4–2

The first experiment of this kind, by S. A. Lee, R. Wallenstein, and T. W. Hänsch, was good enough to resolve the hyperfine structure from nuclear interaction in the hydrogen 1S state (Figure 5). They were also able to measure the Lamb shift of that state, which appears as a deviation from the exact 2:1 ratio of the wavelengths for the two resonances. The resolution, about 1 part in 10 million, was limited by the laser. Ultimately when all other sources of line broadening, such as those due to pressure or stray electric fields, are removed the line width of the 1S–2S transition should be limited only by the lifetime of the excited state. Because that lifetime is so long, the line width could ultimately be as narrow as a part in 10^{15}. Such a sharply defined wavelength or frequency should be measurable to, say, 1% of the resonance width or to a part in 10^{17}. But nobody measures anything to 1 part in 10^{17}! There are simply no methods or standards of that precision. Attempts to push the accuracy of laser measurements on hydrogen will challenge scientists for many years. As the experimental techniques are improved, the theory will have to be refined to face even more stringent tests. Perhaps there may even be some surprise finding that will force a revision of the basic concepts of physics.

Many other applications of lasers to spectroscopy are worth mentioning, but in this space we

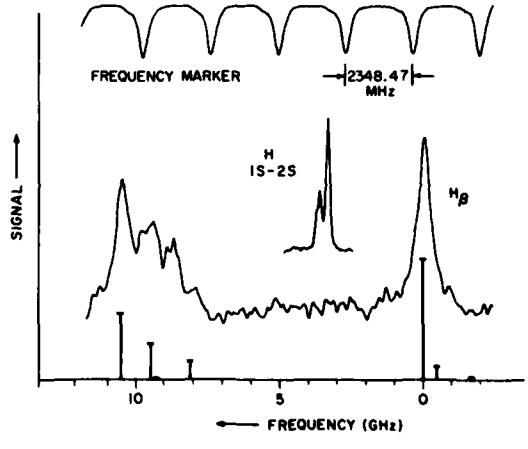


Figure 5—Results of simultaneous doppler-free scans of H_β and 1S → 2S spectral lines

will have to be content with these few. Nevertheless, it is clear that laser spectroscopy is one of the cutting edges of modern science, an active field full of surprises.

LASER CHRONOSCOPY

Very short pulses of laser light can be produced by mode-locking techniques. If the active medium in a laser can amplify a broad band of wavelengths, usually it will oscillate simultaneously in many modes of slightly different wavelengths, usually it will oscillate simultaneously in many modes of slightly different modes to be synchronized. For a very brief instant, all of their peaks are in step, producing a pulse. But since the waves cover a range of different wavelengths, they are quickly out of step, so that the pulse is very short.

Laser light pulses as short as 10^{-13} s have been generated from dye lasers. This pulse is so short that it contains only about 60 cycles of the light wave. In its duration, the light travels a distance of only a few hundredths of a millimeter. The length of such a short pulse is not easy to measure, but one can use techniques for measuring coincidences between two parts of the pulse, one of which has been delayed by traveling a known path distance. Very fast, streaking oscilloscopes using high-speed electrons have also been made to operate in the range of these ultrashort times.

One important application of these short light pulses is in monitoring fast chemical changes, such as those of visual pigments exposed to light. For these studies, the structure of the molecules is inferred from Raman scattering. The Raman spectrum is produced by a second pulse which can follow the first one by a chosen short delay. These studies have shown that the initial effect of the exposure to light is a change of shape of the sensitive molecule. Much more information about fast chemical and biochemical processes will be obtained using ultrashort pulses of laser light.

MEASUREMENTS AND STANDARDS

The new methods of laser spectroscopy have revealed many spectral lines whose wavelength can be defined to a part in 10^{10} or so, and several orders of greater stability can be attained in some cases. But the international standard of length has been a specified line in the spectrum of krypton, and the fractional line width is about one part in a million (10^6). With great care, measurement standards laboratories can locate the center of this standard line to about 1/300 of its width, or 3 parts in 10^9. Good as this is, it is not adequate for the precisely defined wavelengths revealed by laser interaction with atoms and molecules.

Fortunately, standards of frequency are better. In the radio frequency region, cesium and hydrogen standards are reproducible to 1 part in 10^{12} or better. It is not yet possible to measure the frequency of visible light source directly. However, techniques for frequency measurement now span almost the entire infrared, extending to wavelengths as short as 3 μm. The key to these infrared frequency measurements has been the development by Ali Javan and K. Evenson of semiconductor crystal diode harmonic generators and mixers. These point-contact devices are generate harmonics of a microwave frequency standard in the far infrared region. A gas laser, such as HCN, can then be phase-locked to the standard and in turn used as a source of harmonics of precisely known frequency. Using this technique with five successive gas lasers, the frequency of oscillation of a helium-neon laser tuned to a particular resonance in methane gas was found by K. M. Evenson, J. S. Wells, F. R. Petersen, B. L. Danielson, G. W. Day, R. L. Barge, and J. L. Hall to be 8.8 376 181 627 (50) $\times 10^{13}$ Hz. Its wavelength in terms of the krypton standard is 3.392 231 376 (12) μm.

From these the velocity of light,

$$c = (\text{wavelength/frequency}) = 299\ 792\ 456.2(1.1)\ \text{m/s}.$$

This value of c is considerably more accurate than any previous measurement. Moreover, it is limited in accuracy by the krypton length standard. It has been suggested, therefore that the radiation from a single selected source could be simultaneously the standard of both length and time. This

would be equivalent to defining the velocity of light, so that the standard length would be the distance traveled by light in a standard time. This is the way radar distance measurements are made, and their precision already exceeds the accuracy into which time measurements can be converted into distances.

In the future, it will probably be possible to extend frequency measurements into the visible and ultraviolet regions. When that happens, it may well become customary to specify the frequencies rather than the lengths of light waves. However, it will still be necessary to make comparisons between light wavelengths and the dimensions of ordinary objects, at least until laser radar with picosecond chronoscopy becomes easy.

CONCLUSIONS

Lasers and their applications have developed in many different directions, and the end of this proliferation is not in sight. This very diversity has made many things possible but has so far discouraged mass production of any individual kinds of lasers. Although we have been able to reason from past experience to some likely future directions, the possibility of surprises remains very real. Future lasers may be as different from present ones as a transistor is from a vacuum tube. In any branch of the field, the rate of progress will depend on the effort and resources committed to it. Nevertheless, an unexpected idea can still upset all expectations, and it can expose would-be prophets for the fools that we are.

MATHEMATICAL AND INFORMATION SCIENCES



George B. Dantzig has been Professor of Operations Research and Computer Science at Stanford University since 1966. His earlier positions were with the U.S. Bureau of Labor Statistics, USAF Statistical Control, USAF Headquarters, the RAND Corporation, and the University of California, Berkeley. Dr. Dantzig earned an A.B. at the University of Maryland, an M.A. at the University of Michigan, and a Ph.D. at the University of California, Berkeley. He is a Fellow of the Econometric Society, of the Institute of Mathematical Statistics, of the Association for the Advancement of Science, and of the Operations Research Society of America. He has received a number of special honors, including the War Department Exceptional Civilian Service Medal 1944, election to the National Academy of Sciences 1971, the American Academy of Arts and Sciences 1975, the American Academy of Arts and Sciences 1975, the John von Neumann Theory Prize (of the Operations Research Society of America and The Institute of Management Science) 1975, and the National Medal of Science 1975.

LINEAR PROGRAMMING, PAST AND FUTURE*

George B. Dantzig

Stanford University
Stanford, Calif.

The term "programming" is used to refer to planning or scheduling activities of organizations such as factories, airlines, the defense establishment, the national economy, or world trade. (It is not to be confused with "programming" as used for the preparation of a sequence of instructions for a computer.) The goal of programming is to find optimum schedules.

A simple example, "the assignment problem," illustrates the essential difficulty. A factory has 70 men with different qualifications, and it is desirable to assign them to 70 jobs. If a "value" can be attached to assigning a particular man to a particular job, then the problem is to select, from the 70! ("70 factorial," or the product of the integers from 1 to 70) possible ways of permuting the assignments, the one that yields the maximum total value to the factory. Because 70! is approximately 10^{100}, it would take an electric computer executing 1,000,000 operations per second more than 10^{87} years (or many times the projected life of the universe) to examine all the permutations.

Such decision problems are common and have evoked the clever formulation of mathematical models, powerful mathematical methods of solution, and efficient computer algorithms (step-by-step procedures).

One of these methods, linear programming, has come into wide use since its conception in 1947 in connection with military planning. Mathematicians and economists have written books on the subject. Our purpose is to give a brief account of its origins and to point out the influences that brought about its development. Interestingly enough, in spite of its now recognized wide applicability to everyday problems, linear programming was unknown before 1947. Fourier may have been aware of its potential in 1823, and it is true that in 1939 in the U.S.S.R., Kantorovitch made linear programming proposals that were neglected there during a period that witnessed its discovery and rapid development elsewhere.

INFLUENCE OF MILITARY PLANNING

The following statement of M. K. Wood and M. A. Geisler is pertinent:

> It was once possible for a Supreme Commander to plan operations personally. As the planning problem expanded in space, time, and general complexity, however, the inherent limitations in the capacity of

*In developing this paper, I have drawn heavily on the historical material contained in an earlier paper, "Linear Programming and its Progeny" prepared as a *Vicennial Article* of the Naval Research Reviews (June 1966); also on material found in my Encyclopedia Britannica article (joint with R. Cottle).

any one man were encountered. Military histories are filled with instances of commanders who failed because they bogged down in details, not because they could not eventually have mastered the details, but because they could not master all the relevant details in the time available for decision. Gradually, as planning problems became more complex, the Supreme Commander came to be surrounded with a General Staff of specialists which supplemented the Chief in making decisions. The existence of a General Staff permitted the subdivision of the planning process and the assignment of experts to handle each part. The function of the Chief then became one of selecting objectives, coordinating, planning, and resolving conflicts between staff sections.

During World War II, the planning process became so intricate, lengthy, and multipurposed that a "snapshot" of the military staff at any one time showed it to be working on many different programs, some in early phases of development and based on earlier ground rules and facts. To cut the time of the planning process, a patchwork of several of these programs, based on inconsistent facts and rules, was often thrown together. To coordinate this work better, the Air Staff, for example, around 1943, created the program-monitoring function under Professor E. P. Learned of Harvard. This program was started off with a war plan containing the wartime objectives. From this plan, by successive stages, the wartime program specifying unit deployment to combat theaters, training requirements of combat and technical personnel, supply and maintenance, etc., was computed. For consistent programming, the ordering of the steps in the schedule was so arranged that information flowed from echelon to echelon in only one direction, and the timing of information availability was such that the part of the program prepared at each step did not depend on any following step. Even with the most careful scheduling, it took about 7 months to complete the process.

After the war, it became clear that efficient coordination of the energies of whole nations in the event of total war would require scientific programming techniques. Undoubtedly this need has occurred many times in the past, but this time two concurrent developments had a profound influence: (a) the development of large scale electronic computers and (b) the development of the interindustry model proposed by Wassily Leontief. The potential attraction of the input-output model was its simple linear structure. In some ways it was too simple. It was not dynamic; it assumed that each industry had a unique technology that produced only one product. It was not possible with this model to have alternative feasible programs.

It was necessary, therefore, to generalize the interindustry approach. The result was the development of the linear-programming model. Intensive work began in June 1947 in an Air Force group under Comptroller General Ed Rawlings. This effort later was given the title "Project SCOOP" (Scientific Computation of Optimum Programs). Principals in the group included Marshall Wood, Murray Geisler, John Norton, and the author.

The simplex computational method for choosing the optimal feasible program was developed by the end of the summer of 1947. Interest in linear programming began to spread quite rapidly. During this period, the military sponsored work at the Bureau of Standards on electronic computers and on mathematical techniques for solving such models.

Early contacts with Tjalling Koopmans of the Cowles Commission (then at the University of Chicago and now at Yale), Robert Dorfman (then of the Air Force, now at Harvard), and such economists as Paul Samuelson and Kenneth Arrow spurred an intense reexamination of classical economic theory, based on the ideas and results of linear programming.

Early contact with John von Neumann at the Institute for Advanced Study gave fundamental insight into the mathematical theory and sparked the interest of A. W. Tucker of Princeton University and two of his former students, David Gale and Harold Kuhn. With Office of Naval Research support, they attacked problems in linear inequality theory and game theory. Princeton became an academic focal point for these related fields.

The size of the military planning problem made it evident immediately after the war that even the best future computing facilities would not be powerful enough to find an optimal solution to a general detailed military planning model. Accordingly, Project SCOOP modified its approach and

in the spring of 1948 proposed development of special linear-programming models called "triangular models," whose stepwise staff procedure provided feasible but not necessarily optimal solutions.

Since 1948 the military has more and more actively used mechanically computed programs. The triangular models are in constant use for computation of detailed programs, while the general linear-programming models have been applied to smaller systems such as contract bidding; balanced aircraft, crew training, and wing deployment schedules; schedules for maintenance overhaul cycles; personnel assignments; and airlift routing problems.

During the period from 1948 on, granting agencies—particularly the Office of Naval Research—began to support research seeking to develop efficient methods for finding optimal solutions to larger and larger planning systems.

THE INFLUENCE OF ECONOMIC MODELS

The inspirations for the general linear-programming model were the practical planning needs of the military and the possibility of generalizing to this end the simple structure of the Leontief model. From a purely formal standpoint, one could consider the input-output model as a simplification of the Walrasian model. Actually, theoretical economic models were a kind of ivory tower. Leontief stated in the 1930s,

> One hundred and fifty years ago when Quesnay first published his famous schema, his contemporaries and disciples acclaimed it as the greatest discovery since Newton's laws. The idea of general interdependence among the various parts of the economic system has become by now the very foundation of economic analysis. Yet, when it comes to the practical application of this theoretical tool, modern economists must rely exactly as Quesnay did upon fictitious numerical examples.

Leontief's great contribution, in the opinion of the author, was his construction of a quantitative model of the American economy for the purpose of tracing the impact of government policy and consumer trends on a large number of industries imbedded in a highly complex series of interlocking relationships. To appreciate the difference between a purely formal mathematical model and an empirical model, it is well to remember that to acquire data for a real model an organization must work many months, sometimes years. After the model has been put together, a second obstacle looms—solution of a very large system of simultaneous linear equations. In the 1936–1940 period there were no electronic computers; the best that one could hope for in general would be to solve 20 equations in 20 unknowns. After this, there is a third obstacle, the difficulty of "marketing" the results of such studies. From the outset, the undertaking begun by Leontief represented a triple gamble.

As a result of the Great Depression and the advent of the New Deal, the Government made a serious attempt to identify and then support certain activities that it hoped would speed recovery. This brought about more intensive collection of statistics on costs of living, wages, national resources, productivity, etc. There was a need to organize and interpret this data in order to construct a mathematical model describing the economy quantitatively.

From 1936 on, the scope, accuracy, and area of application of Leontief-type models were greatly extended by the Bureau of Labor Statistics. The work there by Duane Evans, Jerome Cornfield, and Marvin Hoffenberg stimulated efforts toward seeking a mathematical generalization suitable for dynamic military applications. Today Leontief models are in wide use—many countries have input-output models of their economies. Leontief received the Nobel Prize for his work in 1974.

In 1947, T. C. Koopmans took the lead in bringing to the attention of economists the potential of linear-programming models. His rapid development of the economic theory of such models was due to the insight he gained during the war with a special class of linear-programming models, called "transportation models," which he applied to Allied shipping problems. In 1949, he organized the historic Cowles Commission conference on linear programming which was attended by young men who since have become well known: K. Arrow, R. Dorfman, L. Hurwica, A. Lerner, J. Marschak, O. Morgenstern, P. Samuelson, and H. Simon; such mathematicians as G. W. Brown, M. M. Flood, D. Gale, H. W. Kuhn,

C. B. Tompkins, and A. W. Tucker. Government statisticians, including W. D. Evans, M. A. Geisler, M. Hoffenberg, and M. K. Wood, also attended. The papers presented there were collected in the book *Activity Analysis of Production and Allocation*. (T. C. Koopmans, ed., Cowles Commission Monograph 13, John Wiley & Sons, Inc., New York, 1951.)

The following quotation from that book's introduction written by Koopmans, serves to characterize the linear-programming model:

> The adjective "linear model" relates only to (a) assumption of proportionality of inputs and outputs in each elementary productive activity and (b) the assumption that the result of simultaneously carrying out two or more activities is the sum of the results of the separate activities. In terms more familiar to the economist, these assumptions imply constant returns to scale in all parts of the technology. They do not imply linearity of the production function. . . . Curvilinear production functions . . . can be obtained from the models . . . by admitting an infinite set of elementary activities. . . .
> Neither should the assumption of constant returns to scale . . . be regarded as essential to the method . . . although new mathematical problems would have to be faced in the attempt to go beyond this assumption. More essential to the present approach is the introduction of . . . the elementary activity, the conceptual atom of technology, into the basic postulates of the analysis. The problem of efficient production then becomes one of finding the proper rules for combining these building blocks. The term "activity analysis" . . . is designed to express this approach.

It is interesting to note that four economists who early in their careers made important contributions to linear programming and its relation to allocation theory have received the Nobel Price: Ragnar Frisch, Paul Samuelson, Kenneth Arrow, and T.C. Koopmans.

MATHEMATICAL HISTORY

The linear-programming model, when translated into purely mathematical terms, requires a method for finding a solution to a system of simultaneous linear equations and linear inequalities that minimizes a linear form. This central mathematical problem was not known to be important until the birth of linear programming in 1947.

We are all familiar with methods for solving linear equation systems, from our first algebra courses. The literature of mathematics contains thousands of papers on techniques for solving linear equation systems with the theory of matrix algebra (an allied topic), with linear approximation methods, and so on. On the other hand, the study of linear inequality systems excited virtually no interest until the advents of game theory in 1944 and linear programming in 1947. For example, T. Motzkin, in 1936, in his doctoral thesis on linear inequalities, was able to cite after diligent research only some 30 references for the period 1900–1936 and about 42 in all. In the 1930s, 4 papers dealt with building a comprehensive theory of linear inequalities and with an appraisal of earlier works. These were by R. W. Stokes, Dines McCoy, H. Weyl, and T. Motzkin. As evidence that mathematicians were unaware of the importance of the problem of a solution to an inequality system that also minimized a linear form, we may note that none of these papers made any mention of such a problem, although there had been earlier instances in the literature.

The famous mathematician Fourier, while he did not go into the subject deeply, appears to have been the first to study linear inequalities systematically and to point out their importance to mechanics and probability theory. He was interested in finding the least-maximum-deviation fit to a system of linear equations. He reduced this to the problem of finding the lowest point of a polyhedral set and suggested a solution by a vertex-to-vertex descent to a minimum, which is the principle behind the simplex method used today. Later, another famous mathematician, de la Vallee Poussin, considered the same problem and proposed a similar solution.

THE WORK OF KANTOROVITCH

The Russian mathematician L. V. Kantorovitch has long been interested in the application of mathematics to programming problems. He published an extensive monograph in 1939 entitled "Mathematical Methods in the Organization and Planning of Production."

In his introduction, Kantorovitch states:

> There are two ways of increasing efficiency of the work of a shop, an enterprise or a whole branch of industry. One way is by various improvements in technology, that is, new attachments for individual machines, changes in technological processes, and the discovery of new, better kinds of raw materials. The other way, thus far much less used, is by improvement in the organization of planning and production. Here are included such questions as the distribution of work among individual machines of the enterprise or among mechanisms, orders among enterprises, and the correct distribution of different kinds of raw materials, fuels and other factors.

Kantorovitch should be credited as the first to recognize that certain important broad classes of production problems had well-defined mathematical structures, which he believed were amenable to practical numerical evaluation and could be numerically solved. If Kantorovitch's earlier efforts had been appreciated when they were first presented, linear programming might be more advanced today. However, his early work in this field remained unknown both in the Soviet Union and elsewhere for nearly 20 years while linear programming became a highly developed art. According to *The New York Times*,

> The scholar, Professor L. V. Kantorovitch, said in a debate in 1959 that Soviet economists had been inspired by a fear of mathematics that left the Soviet Union far behind the United States in applications of mathematics to economic problems. It could have been a decade ahead.

In 1975, Kantorovitch received the Nobel Prize for his contributions.

During the summer of 1947 Leonid Hurwicz, a well-known econometrician associated with the Cowles Commission, worked with the author on techniques for solving linear-programming problems. This effort and some suggestions of T. C. Koopmans resulted in the simplex method. The obvious idea of moving along edges from one vertex of a convex polyhedron to the next (which underlies the simplex method) has been rejected earlier, on intuitive grounds, as inefficient. In a different geometry using a special choice rule it seemed efficient, and so, fortunately, it was tested and is now accepted as the standard procedure.

THE WORK OF VON NEUMANN

Credit for the mathematical foundations of this field goes to John von Neumann more than to anyone else. During his lifetime he was generally regarded as the world's foremost mathematician and played a leading role in many fields. Perhaps in the long run his stimulation of electronic-computer development during World War II will prove his most significant contribution. In 1944, von Neumann and Oskar Morgenstern published their monumental work on the theory of games, a branch of mathematics that aims to analyze problems of conflict by use of models called "games." A theory of games was first broached in 1921 by Emile Borel and was first established in 1928 by von Neumann with his famous "minimax theorem." The significance for us is that game theory, like linear programming, has its mathematical foundation in linear inequality theory.

Von Neuman, at the first meeting with the author in October 1947, was able immediately to translate basic theorems in game theory into their equivalent statements for systems of linear inequalities. He introduced and stressed the fundamental importance of *duality* and conjectured the equivalence of games and linear-programming problems. Later, he made several proposals for the numerical solution of linear-programming and game problems.

ELECTRONIC COMPUTER CODES

New computational techniques and variations of older techniques are continually developed. A number of important variants of the simplex method were proposed by C. Lemke, W. Orchard-Hays, E. M. L. Beale, P. Wolfe, and many others during the 1950s. The well-known econometrician Ragnar Frisch of the University of Oslo did extensive research work on his "multiplex method." Investigations in Great Britain have been spearheaded by S. Vajda and M. Beale.

A special variant of the simplex method, developed for transportation problems, was first coded in 1950 for the National Bureau of Standards SEAC computer. The general simplex method was coded in 1951 under the general direc-

tion of A. Orden of the Air Force and A. J. Hoffman of the Bureau of Standards. In 1952, W. Orchard-Hays of the Rand Corporation worked out a simplex code for the IBM-C.P.C. and later for the IBM 701, 704, etc. His code turned out to be practical for commercial applications. As a result, the use of electronic computers by business and industry grew by leaps and bounds. Many of the digital computer companies provide, as part of their commercial software, codes of the simplex technique. In fact, computer companies are now spending close to a half million dollars on software development for a complete linear-programming system. At one time this was free to their customers, but now such codes are proprietary. The mathematical systems developed for planning in industry and the military are among the largest in the world. Typical problems run from 300 to 800 equations. Some codes are designed to solve practical problems involving as many as 4000 equations. The number of possible activities (variables) can run into the thousands.

MATHEMATICAL PROGRAMMING

If we distinguish, as indeed we must, between those types of generalizations in mathematics that have led to existence proofs and those that have led to constructive solutions of practical problems, then current developments mark the beginning of constructive generalizations of linear-programming concepts to allied fields.

Mathematical programming may be described in terms of its mathematical structure and computational procedures or in terms of the broad class of important decision problems that can be formulated as the minimization (or maximization) of a function of several variables that are subject to a system of side constraints. For example, linear programming is defined as the minimization of a linear "objective" function whose variables satisfy a system of linear inequalities.

In practice, mathematical programming refers to linear programs, the general study of nonlinear programs (those in which either the objective function or at least one of the constraint functions is nonlinear), integer programs (linear programs with the additional restriction that some or all of the variables must be integer valued), stochastic programs (those involving random variables), and network flow theory (dealing with transportation or flow through networks). As such, mathematical programming overlaps with, has contributed to, and has been influenced by operations research, mathematical economics, control theory, dynamic programming, and combinatorial theory.

Nonlinear Programming

A natural extension of linear programming occurs when the linear part of the inequality constraints and the objective are replaced by convex functions. Early work by Barankin and Dorfman centered about a quadratic objective and culminated in an elegant procedure developed independently by Beale, Houthakker, and Wolfe. Wolfe showed how a minor variant of the simplex procedure could be used to solve such problems. Duality concepts first proposed by von Neumann have successfully been extended to certain classes of nonlinear programs. A result of these investigations is a new uniform procedure that solves linear programs, quadratic programs, general matrix games, and fixed-point problems. This is referred to as a complementary pivot theory. The research of C. Lemke and J. T. Howson, Jr., H. Scarf, and H. Kuhn, R. Cottle, and the author should be mentioned in this connection.

Stochastic Programming

It has been pointed out that programming under uncertainty cannot be usefully stated as a single problem. One important class is a multistage one in which technological matrix of input-output coefficients is assumed known and the values of the constant terms uncertain, but the joint probability distribution of their possible values is assumed to be known. Research in this field is still in its infancy, and practical planners will continue to resort for some time to heuristic schemes to cover stochastic events. Stochastic programming methods are being used by A. Manne in developing "robust" policies with regard to the development of nuclear energy in the face of uncertainty about fast breeders and fusion reactors.

Network Theory

A remarkable property of one very special class of linear programs, namely, the transportation, or network flow, problem, is that their solutions are always in integers. This key fact links certain combinatorial problems in mathematical topology with certain continuous problems of network theory. The field has many contributors. Of special mention is the work of Kuhn (for finding a permutation of ones in a matrix composed of zeroes and ones) and the related work of Ford and Fulkerson at RAND (for network flows). Very efficient techniques for solving large-scale networks have recently been developed, based on ideas of D. R. Fulkerson, J. Edmonds, E. Johnson, and others. Network flow theory is now considered part of graph theory. A number of important combinatorial problems, such as covering problems, packing problems, and routing problems, are considered here. An important area is matroid theory.

Integer Programming

Combinatorial problems in general are extremely difficult if not impossible to solve. Important classes of nonlinear, nonconvex, discrete, combinatorial problems can be shown to be formally reducible to linear-programming problems with the additional restriction that some or all of the variables must be interger-valued. The linear-programming approach was used in 1954 by Fulkerson and Johnson and the author to construct an optimal tour for a salesman visiting Washington, D.C., and 48 State capitals of the United States. Our theory was incomplete, however. The foundations for a rigorous theory were first developed by Ralph Gomory in 1958 under an Office of Naval Research contract with Princeton University. Many important planning problems are integer programming problems. The best location of warehouses, optimal routing of a fleet of supply ships, optimal provisioning under space and weight limitations, optimal sequencing of jobs on machines, assignment of crews to meet an airline routing schedule, and optimal ways to cut out patterns from stock materials are some examples. The "cutting-plane" approach of Gomory is often used in conjunction with more heuristic search methods such as "branch and bound." The latter has been very successful in practice.

APPLICATIONS OF LINEAR PROGRAMMING

The history of the first years of linear programming would be incomplete without a brief survey of its use in business and industry. This began in 1951 but has grown so quickly that the commercial offspring has overtaken its military parent.

Linear programming has served industrial users in several ways. It has provided a novel view of operations, it has induced research in mathematical analysis of the structure of industrial systems, and it has become an important tool for business and industrial management to use in improving efficiency. The application of linear programming to a business or industrial problem requires the mathematical formulation of the problem and an explicit statement of the desired objectives. In many cases, such rigorous thinking about business problems has clarified aspects of management decisionmaking that previously had been hidden in a haze of verbal arguments. As a partial consequence, some industrial firms have started educational programs to emphasize to their managerial personnel the importance of defining objectives and of constraints on business policies. Moreover, scheduling of industrial production traditionally has been based on intuition and experience, a few rules, and the use of visual aids, just as in the military. Linear programming has induced extensive research in developing quantitative models of industrial systems for the purpose of scheduling production. Of course, many complicated systems have not as yet been quantified, but sketches of conceptual models have stimulated widespread interest.

The first and most fruitful industrial applications of linear programming have been to the scheduling of petroleum refineries. Charnes, Cooper, and Mellon started their pioneering work in this field in 1951. During the 1950s two books were written on the subject, one by Gifford Symonds and another by Alan Manne. So intense has been the development that a survey by Garvin, Crandall, John, and Spellman in 1957 showed that the oil industry used linear programming in

every phase of its activities from exploration, production, and refining to final distribution and sales.

The food-processing industry is perhaps the second most active user of linear programming. In 1953, a major producer first used it to determine shipping schedules for catsup from 6 plants to 70 warehouses. In 1976, a national company solved a huge linear program by the decomposition method to decide which of their bakeries should fill orders for cookies.

In the iron and steel industry, linear programming has been used for the evaluation of various iron ores and of the pelletization of low-grade ores. Additions to coke ovens and shop loading of rolling mills have provided additional applications. A linear-programming model of an integrated steel mill has been developed. The British steel industry has used linear programming to decide what products their rolling mills should make to maximize profit.

Metalworking industries use linear programming for shop loading and for deciding whether to make a part in a shop or to buy it outside. Paper mills use it to decrease trim losses and to decide which of several mills should respond to a given order.

The optimal routing of messages in a communication network, contract-award determinations, and the routing of aircraft and ships are problems to which application of linear-programming methods was first considered by the military, but they are now significant in industry.

Currently, linear- and nonlinear-programming models are used to assess energy options as a result of the energy crisis. Some examples are the work of W. Hogan at FEA, K. Hoffman at Brookhaven, A. Manne at Harvard, and the PILOT Energy Project at Stanford.

One measure of the use of linear programming and its extensions is the money spent on computer time in the United States. This is known to run in the millions.

LARGE-SCALE SYSTEMS DEVELOPMENT

At the present stage of the computer revolution, there is growing interest on the part of practical users of linear-programming models in solving larger and larger systems. It is difficult to measure the potential of large-scale linear programs and its nonlinear extensions. Certain developing countries appear, according to optimal calculations on simplified models, to be able to grow at the rate of 15% per year; this implies a doubling of their industrial base in 5 years. However, administrators apparently ignore plans and make decisions based on political expediency, which restrict growth to 2% or 3%, or sometimes −2%. Nevertheless, it is my belief that the mechanization of data flow (at least in advanced countries) in the next decade will provide pathways for construction of large models and effective use of the results of optimization. This points up the need to develop efficient tools now for optimizing large-scale linear and nonlinear programs.

In particular let me cite the energy crisis, which will probably go on for some time in the future. Integrated national economy-energy models are being developed and solved using standard linear-programming methods. Already a bottleneck on the size of energy models has been encountered; large-scale solution techniques are not available for practical application.

Of all the progeny of linear programming, perhaps the most fruitful at present are techniques for solving linear programs with special structures. Many groups have been developing linear-programming models of their firms for more than a decade. The trend is toward larger and more comprehensive corporate models that are multistaged and dynamic and exhibit hierarchical structures. Although many proposals have been made, little in the way of practical codes has been developed to handle such problems.

In 1959, Philip Wolfe and the author proposed the decomposition principle, an approach that decomposes a model into smaller parts that can be independently optimized. The solution of each part is treated as a proposal and is modified to be consistent with total system resources and demands. Several companies, such as C.E.I.R., Mathematica, and Bonner and Moore, which specialize in developing computer programs, have written decomposition codes. In the National Biscuit Company's application, a system of half a million variables and 100 equations is solved every 2 weeks.

The power of computing machinery has increased, and the power of methods proposed by

mathematicians has grown. Optimal solutions to large-scale complex planning may someday become achievable.

Can computers be programmed to solve the truly immense systems characteristic of a national economy, particularly dynamic systems involving optimization? Here again we note that by use of the decomposition principle systems of the order of 9×10^4 equations and 5×10^5 variables have already been solved. Even though total system optimization is at present impossible, there are various schemes involving partial aggregation that permit near-optimal solutions.

THE ROLE OF SYSTEMS OPTIMIZATION LABORATORIES

Briefly stated, the objective of a Systems Optimization Laboratory (SOL) is to advance the state of the art of computational mathematical programming and thus expand the ability of this technique to solve important problems of the real world. The importance of this objective would be difficult to overstate. The success of mathematical programming in dealing with the problems of industry and government are well recognized and appreciated. Inevitably, however, ever larger and more complex models have developed to keep pace with the constant advances in science and technology, the apparently ever-increasing complexity of social organizations and responsibilities, the massive volume and detail of data available with modern data base management, and the demand for greater efficiency and cost effectiveness in times of economic stress. The solution of problems of pressing real-world importance is hampered as these models push against and sometimes beyond the capabilities of conventional mathematical programming technology and software.

There is no lack of theoretical proposals for dealing with large-scale mathematical programs; on the contrary, the literature abounds with theoretical and almost always untested algorithms designed to take advantage (at least on paper) of the various structural features of such models. The problem is that, while some of these proposals are clearly impractical, others do indeed show promise. While experience and analysis can make possible some winnowing, the only final criterion can be systematic experimentation with representative models. Clearly, for such experiments to have meaningful and reliable results, implementation must be sophisticated and test problems large enough to give a guide to real problem behavior. Unfortunately, until quite recently virtually the only sophisticated systems capable of handling large problems have been the commercial mathematical programming systesm. These systems have certain limitations; traditionally they have not been designed to be easy to modify or sufficiently modular to use as a collection of subroutines for implementing algorithms that take advantage of such structural features as time-staging and block-angularity. The small body of systems programmers who produce, maintain, and update these systems have little inclination and less time to radically modify the complex and rather rigid systems to experiment with untested algorithms. In general it is only when an algorithm has been proved and its commercial advantages have been demonstrated that it is seriously implemented. Even then the likelihood of even experimental implementation depends very strongly on the degree of system modification required and the eventual marketing potential.

The goal of a Systems Optimization Laboratory is to bridge the gap between theory and practice, thereby to expand the problem-solving power of mathematical programming. Accomplishing this requires several conditions. There must be a flow of practicable algorithmic ideas and developments; in a university, this would be mainly the function of the faculty and doctoral students associated with the laboratory. There should also be interaction between algorithmic and software workers and model developers. The most direct need, however, and the major focus, is for full development of adequate software tools and their use in realistic experiments on representative problems, together with a mechanism for collecting and disseminating these and other results.

The type of software required by a laboratory is rather different from the elaborate commercial systems and the oversimplified programs so often used in small-scale testing. A highly modular system is required, as simple, general, readable, and well documented as possible, and preferably in a

higher level programming language. The system must use mathematical programming technology of a sophistication comparable with existing production programs. The efficiency thus obtained is essential if results are to be useful in comparing new and established techniques.

A Systems Optimization Laboratory has three major functions:

1. Further developing and extending a modular, portable, higher level language mathematical programming system for experimental and real-world problem-solving purposes.
2. Evaluating new algorithmic proposals for large-scale systems in a realistic experimental framework; particular emphasis must be put on multi-time-period models.
3. Extending the Systems Optimization Laboratory's role as a clearing house; this includes compiling a suite of good test problems and a library of programs, as well as disseminating computational information.

A FEW WORDS ABOUT THE FUTURE

One of the most startling recent developments is the penetration of the electronic computer and mathematics into almost every phase of human activity.

If there is a library, then someone is at work representing (in the memory of a computer) the book's number, its title, its shelf location, who has it on loan, the date due, the author, the book's call number, its cross references, its frequency of use, and so on. A library is like a population that does not bury its dead. Out of this straightforward effort to get some of the present information about a library into a more manipulable form will emerge the "information storage and retrieval system" of tomorrow; the old physical book and printed paper page may become as much a relic as an ancient scroll.

Wherever one finds a system for processing insurance premiums, for keeping track of bank deposits and withdrawals, for recording airline reservations, or for any other type of inventory control, someone is at work simulating such a system in an electronic computer and forging the links whereby the real world supplies information to the computers and the orders of the computer are translated into real actions.

It is correct to regard much of what has been done so far as a vast "tooling up," a preparation for new ways to do old tasks. It is the exponential improvement in electronic hardware and the availability of new machine languages and special machine programs that now permit practical implementation of these ideas. We are witnessing an accelerated trend toward automation of simple human control tasks.

Operations research is the science of decision and its application. In its broad sense, the word "cybernetics," the science of control, may be used in its place. This science is directed toward tasks that humans have not yet delegated to machines. Tasks involving human energy and (as we have seen) those involving simple human control already have been conceded to machines even though they have not been taken over fully by them. The automation of higher order human decision processes is the last citadel.

At the lowest level of these higher order tasks is the human ability to recognize patterns in sight, sound, touch, smell, and taste. Although these tasks may elicit simple responses (such as "turn the wheel to the right or left"), human presence is needed because a complex mental recognition process is involed. It is relatively easy to get a machine to mechanically separate returned Coke and Pepsi bottles once it is smart enough to recognize which is which.

At the next level of complexity is the human ability to observe and to adapt to physical movement; for example, to observe a dial or a car's angle to the road direction and to manipulate certain controls to change the physical movement in some preferred way. Here, again, it is easy to get a machine to make the physical movement of the controls if the machine is smart enough to adapt to trends in the observed movements as changes are made in the controls.

Although pattern recognition is by no means a solved problem, banks do have machines that recognize account numbers on checks, and there are machines that give change for a dollar bill but not for a blank piece of paper. Automatic feedback controls in simple situations have been known for a long time. The governor invented by

Watt to control the speed of a steam engine is such a device. Closed-loop controls that rely on computers to analyze input data are now a reality in certain large-scale operations, such as oil refineries, chemical plants, and power-distribution systems.

At a still higher level of complexity are those decision processes that involve many alternative courses of action. An industrial complex may have at its disposal many types of equipment and a variety of raw materials and personnel skills. The complex could manufacture a variety of products by means of alternative process sequences. If the wrong decisions are made in the scheduling of the various processes, labor and machines are idle, throughput is reduced, and in-process inventories are increased. If the wrong decisions are made in raw-material selection, the procedure for manufacture, or the choice of final product, labor and machines are overworked, expensive materials are purchased when cheap ones will do, and unwanted products are dumped on the market.

In the last two decades, great strides have been made in effectively using electronic computers as part of the planning process. As we have noted already, a pioneering effort of this kind was begun by the military around 1947 in Project SCOOP. Part of that project included a 400-sector interindustry model of the national economy. Except for the preparation of input data, the calculations of various planning programs were completely mechanized. The size of systems handled was truly enormous. A program typically stated month by month (for 36 months) the level of each of thousands of types of activity. The balanced flows of tens of thousands of input and output items necessary to support these activities were also given as a function of time. As ground rules, appropriations, or international conditions changed, these programs were recalculated rapidly again and again.

This early pioneering effort at mechanizing the planning process showed that it was possible to describe mathematically the interdependence of various activities, such as training, the work of a combat unit, an engine overhaul, steps in an industrial process, and the shipment of goods from various places of origin to numerous destinations. The approach, as we saw earlier, is to make each activity *elementary enough* so that its inputs and outputs are proportional to the level of the activity. The resulting mathematical system is a system of linear inequalities called a linear program. Use of this mathematical approach relieves planning staffs of much drudgery and enables them to concentrate more and more on overall objectives.

"True optimization" is modern research's revolutionary contribution to decision processes. In the entire history of mankind, a great gulf has always existed between man's aspirations and his actions. He may have wished to state his wants in terms of objectives, but there were so many possible different ways to go about it, each with its own good and bad, that it was impossible to compare them and say which was best. People invariably turned to leaders, managers, governors, or commanding officers, whose experience and mature judgement would point the way. Inevitably, "the way" became the new objective. This *substitution of the means for the objective* is the history of mankind. The slogan "the end justifies the means" perhaps could be better stated as "the end might conceivably justify the means if one could remember what the original end was."

Because man was unable to select the best among infinite alternatives, his planning was characterized by many ground rules and policies, dictated by men of mature judgement. It seemed impossible that planning could ever be done by computer unless the machine was constantly stopped to await decisions by the experts. The habits of centuries are not easily overcome, but planning staffs freed from the drudgery of computing one or possibly two alternatives now are beginning to express themselves in terms of overall objectives and to ask the computers to find from among many alternatives the best.

We are witnessing a computer revolution in which nearly all tasks of man—manual labor or simple control, pattern recognition or complex higher order decisionmaking—are being reduced to mathematical terms and solved by computers. It is in the latter development that linear programming and its extensions play a key role.



Harvey M. Wagner is Dean of the School of Business Administration of the University of North Carolina at Chapel Hill. Earlier, he taught at Yale and Stanford Universities. Dr. Wagner has served as a consultant to the RAND Corporation and for 16 years as a consultant to McKinsey and Co. His book *Principles of Operations Research* (Prentice-Hall) won the ORSA-Lanchester Prize and the AIEE Maynard Award. He has published many articles on logistics, and his book *Statistical Management of Inventory Systems* (Wiley, 1962) is a landmark in that field. Dr. Wagner has been active in operations research professional societies and served as President of the Institute of Management Sciences.

THE NEXT DECADE OF LOGISTICS RESEARCH

Harvey M. Wagner

School of Business Administration
University of North Carolina
Chapel Hill, N.C.

McKinsey and Co.
New York, N.Y.

Abstract: Pathbreaking logistics research over the next 10 years will focus on systems problems. Whereas past research generally has taken a "bottom-up" approach, future investigations are likely to pursue a "top-down" philosophy. Specifically, attention will concentrate on diagnosis of systems' improvement potentials; easy-to-use analytic approaches, inherently approximative, will be devised for quickly ascertaining whether a complex operating system can be substantially and effectively improved. Theories to assist in overall systems design, particularly the setting of boundaries and buffers among systems components, will be developed. At the same time, techniques for accurately forecasting future systems performance will be investigated.

Underlying such research will be efforts to gain better understanding of management information requirements, including approaches for monitoring systems performance and providing early warning detection of systems degradation. Improved management information systems will have to be coupled with appropriate design of managerial organizations and assignment of decisionmaking responsibilities. Important avenues of research will be development of robust approaches, that is, both mathematical techniques and organizational approaches that are not too adversely affected by limited data, a changing environment, and human frailty.

Finally, critical research will be directed at the implementation process, especially the interaction among initiation, design, testing, and ultimate adoption.

This prognosis will explore the above themes in the context of large-scale, complex systems. The decision areas will encompass inventory replenishment, multiechelon hierarchies for stockage and maintenance, procurement, transportation, scheduling, facilities planning, budgeting, reliability, and personnel management.

THE MOMENTUM OF HISTORY

Functional Subdivisions

The logistics functions in commercial and military organizations are so well established that their mission and performance often are taken for granted. Even when an organization undergoes major structural renovation, the logistics functions may escape critical notice. Such activities traditionally are defined to include procurement (including purchasing of raw materials, packaging, product components, subassemblies, maintenance items, and capital equipment); manufacturing administrative processes (including scheduling of machinery, sequencing of work orders, selecting of manufacturing techniques); in-

ventory control (including stocking of raw materials, in-process working inventory, and finished goods); distribution of resources that are held at various storage locations; and transportation (including selection of carriers, scheduling and loading of transportation equipment, negotiation of rates, and movement and deployment of personnel). In some organizations, logistics also encompasses maintenance and repair of equipment, reliability engineering, and facilities planning.

Despite the obvious connections among these functions, many organizations separate the responsibilities for the various logistics activities. As a result, the full economic and service improvement potential that could be realized by a coordinated effort is rarely achieved. Furthermore, logistics managers frequently are postured to have a reactive, rather than initiating, role. More specifically, logistics management is expected to execute requests from other parts of the enterprise, but not to actively suggest how overall integrative systems improvements can be made.

Today the costs of logistics have become sizeable, however, and subject to tighter managerial control, so that large organizations can no longer give short shrift to the logistics functions. To the contrary, many establishments have already made noteworthy improvements by eliminating trouble spots in their logistics functions. As we shall suggest, significant new opportunities can be created by an organization that recognizes and can thus coordinate the linkages among its various separate logistics functions.

Management Science Impact

Early in the evolution of management science and operations research, scientists realized that central logistics issues could be studied and eventually comprehended by means of the developing methods of applied mathematics. In particular, the researchers devoted a staggering amount of effort to formulating scientific models of inventory control; devising scheduling policies for equipment, projects, and production; using mathematical programing in planning analyses; testing operating doctrines for machine maintenance, repair, and replacement; evaluating options for transportation routing; and relieving congestion in queuing systems, to cite only a few of the classic problem areas.

The challenge of these problems has engaged the interest of talented scientists, including several recent Nobel Prize recipients. In addition to the intrinsic fascination of the problems' natural complexities, the research was impelled by the growing availability of large-scale electronic computers that presumably could perform numerous calculations and could store and process the data required to drive the model analyses to usable conclusions.

Without doubt, the degree of increased understanding afforded by the model building of management science and operations research in the past 30 years is impressive. An incredible amount of research has been done in fathoming the nature of logistics processes and their associated decisions, and there is no indication that interest and effort are waning.

Nevertheless, logistics managers are justified in questioning the extent to which the research findings have effected day-to-day decisionmaking. Without denying that model-building research has brought significant systems improvements, such managers may express the wish that they could better use logistics models to help solve the remaining larger issues of the design and operation of entire logistics systems.

The Inward Spiral

As in all branches of applied science, an analytic problem, once defined, takes on a life of its own, regardless of its original source and setting. These problem situations seem to hold endless fascination for succeeding generations of scientists. The result frequently is a steady stream of refinements and extensions of the original formulation and analysis. These additions to knowledge may not be trivial from a technical point of view; their elegance and generality may warrant the intense intellectual effort spent producing them. Whether such progress helps solve the original real-life problem is another matter, however. The nature of model-building analysis is to abstract a piece of a complex problem, which can be subjected to fruitful study. Unfortunately but inevitably, the resulting approximation to reality some-

times misses the target of providing a useful guide to decisionmaking. Ample evidence demonstrates that subsequent research often pushes the formative analysis further from reality—that is, makes progress in areas not pertinent to the critical limitations of the initial approximation.

Thus, despite the current active research in logistics processes, we cannot ensure that significant research breakthroughs will continue if we rely solely on letting past momentum determine the types of problems and the technical approaches of the future. To offset the natural tendency of applied research to spiral inward, logistics managers must energetically make known the problem areas that cry out for new analysis. Constant infusion of reality in logistics research is the best guarantee that the next decade of effort will have a major impact.

A SCORECARD OF RESEARCH PROGRESS

Bottom-up and Top-down Orientation

By and large, logistics models have focused on phenomena at the bottom levels of organizations. For example, the mathematical models derived over the past three decades have dealt with replenishment of individual stock items, initial provisioning of spare parts, sequencing of particular orders, overhaul of particular pieces of equipment, replacement of particular components, and so forth. A corollary is that these models have concentrated on single types of logistics decisions (replenishment, procurement, maintenance, transportation) rather than on systems of decisions. Even the notable exceptions to this generalization, such as in applications of mathematical programing models that deal with the deployment of limited resources, often treat as given certain assumptions that the highest level of management would prefer to consider as variables. To illustrate, in a transportation distribution study using mathematical programing, the analysis typically takes as given the products to be shipped and the customers to be served. Top management may be more interested in whether the products should be manufactured at all, whether certain customers are unprofitable because of the transportation rate structure, and how much service is required by customers. Of course, such issues can be sorted out in part with the aid of models, but in practice the typical study orientation has been to ignore such issues.

Another way of stating the point is to say that most management science and operations research models dealing with logistics have not begun by attacking the questions that would be posed by the topmost level of management. For example, when senior management is asked to approve a systems design effort to tighten inventory control, it wants an estimate of the savings potential of such a new design. When expansion of a factory warehouse is proposed, senior management wants an assessment of the possible share-of-market impact of having more or less stock at the location, which may be geographically removed from the company's customers. When a new product is to be introduced by a computer manufacturer, top management wants to know the economic ramifications of providing for concomitant repair and service, including the cost of parts replenishment. In brief, senior managements typically seek a comprehensive economic analysis of the "big picture."

Management scientists have assumed, almost as an axiom, that to obtain answers to high-level management questions, one must build the analysis from the bottom up. Thus, to predict an inventory system's performance, the researcher has been inclined to add up the performance characteristics of the individual components. Regretably, this bottom-up presumption has not proven itself to be without severe limitations. One difficulty has been the sheer effort involved in ascertaining and then "adding up" the component details. The analytic and data-processing difficulties that arise from starting at the bottom and aggregating up can be severe and can consume much of the analytic staff's time and energy. Ironically, in such instances senior management finds itself funding its own research project to learn whether the organization can benefit from previous logistics research.

To make matters worse, the "adding up" process may amplify rather than dampen the errors in the approximative assumptions of micromodels. When economies or diseconomies of scale, such as occur in the loading and routing of transport vehicles, are present, but virtually ignored by a

microcosmic model, the consequent aggregation of individual calculations can be far off the mark. What appears to be an incidental approximation in the small can turn out to be a gross and misleading oversimplification in the large.

It is becoming clearer that these top management issues ought to be modeled in their own right. The potential advantages include faster and more accurate results. Even more important, perhaps, starting at the top affords a better opportunity to focus on issues, assumptions, and evaluation criteria that are most relevant to senior management.

So that there is no misunderstanding, we hasten to acknowledge that top-down analysis is not yet easy. In fact, we believe that this point of view will be a major focus of research over the next decade. The research tasks certainly will be at least as difficult and challenging as those that have been confronted with the bottom-up approach. Work to date suggests that considerable innovation will be required.

The Narrow End of the Time Tunnel

Logistics models have addressed management decisions that at one extreme pertain to daily phenomena, such as replenishment, scheduling, and repair, and at the other extreme, to long-range commitments, such as plant location, capacity expansion, and development of new products. A common observation is that at the first extreme the mathematical models are simpler to analyze (in the sense that they require less data and computation) but harder to implement (in the sense that they frequently require a sweeping systems design). In contrast, planning models for long-term decisions provide extremely useful information with a reasonable amount of effort, but involve an inordinately heavy use of computers and data manipulation.

Most logistics management functions in large enterprises involve an amalgam of both short- and long-term decisions. An important implication is that management of these enterprises must be prepared to deal with the different organizational stresses that arise from applying management science and operations research efforts at the two ends of the time-horizon spectrum. Research staff thus must include personnel capable of one-time innovative model building and data analysis as well as of designing and implementing operating systems.

Leashing the Crunchers

A curious paradox is connected with the use of large computers. As pointed out previously, advances in computer software and hardware technologies have spurred the development of logistics model building. It is inconceivable that the progress made so far in studying logistics decisions could have taken place if computer developments had leveled off. Furthermore, to the extent that such models have been applied to strategic as well as to operational decisionmaking situations, computers have been essential. Nevertheless, the difficulties in using computers in new model-building situations still are severe. In fact, even in so-called standard applications, such as the development of a new medium or large-scale linear-programing model, the tasks of collecting and analyzing the data, converting the data into model coefficients, obtaining usable optimization results, and providing management with readable analyses are now by no means routine. Admittedly, experienced technical experts now have a much better time of it than do novices. Also, today an organization receives considerably more "computation per buck" than it did a decade ago. Be that as it may, management must not view as insignificant the development and completion effort for a logistics model application. To add to the paradox, those software developments aimed at enhancing the application of a particular class of models, such as mathematical programing, have turned out to increase the learning setup time for beginners.

A related point is that, all of the statisticians' research not withstanding, model-building practitioners often are forced to resort to crude ad hoc data manipulation procedures in order to analyze historical information. Unfortunately, a model builder who has had a standard introduction to regression analysis, for example, is not very well equipped to detect, let alone design, useful data fitting formulas. Part of the difficulty, of course, is inadequate education. However, to offer a com-

parison, a logistics model builder need not be a highly trained technical expert or mathematician to run a standard linear-programing computer routine. Yet the same individual is almost certain to fail in manipulating a set of data on a dependent and several independent variables in trying to obtain a tight regression fit. (The usual approach is to employ standard multiple linear regression and hope that the resulting fit will be fairly good.) Oddly, most high-powered statistical routines now available on computers provide copious statistical tests that seem to make little sense to most users. Hence, data analysis for managerial decisionmaking is a burgeoning field with vast opportunities.

Management scientists and operations researchers are only beginning to come to grips with the intricate data analysis problems that arise in the use of computer simulations of stochastically driven systems. Of course, the complexity of such problems has been recognized for many years, but only recently has there been a better appreciation of how pervasive and knotty these difficulties are. The unsophisticated simulation model builder traditionally has assumed that all such estimation problems could be "bought off" by investing in a sufficiently long simulation history. In a trivial sense, that attitude is correct—but only lately has it become apparent that a sufficiently long history may be far longer than most practitioners would ever have guessed. Computation time is a scarce and costly resource, and the solution to these problems is not to run longer but to run smarter. At last this topic is under active research investigation.

Crossing the Technical Barriers

In the next section of this paper, we suggest several general classes of problems that will challenge future researchers of logistics decisions. Here we note a few of the technical problems that remain and attract the attention of researchers.

In one way or another, all realistic applications of model building to logistics decisions involve dealing with large-scale systems. The source of bigness may be the great detail that must be encompassed, for example, as in implementation of stockage rules for a system of tens of thousands of inventoried items, or the source may be the large number of options to be addressed, as in a multiperiod strategic planning model.

The problems of large-scale applications include both the sheer number of computations required as well as the vast amounts of input data that must be collected and reviewed and the resulting extensive output to be analyzed. Much progress is needed in techniques that help human analysts comprehend large sets of data. (Recent developments in computer graphics are good examples of what can be done to let a human literally see multidimensional phenomena.)

A related problem is the development of methods for testing model assumptions and data error sensitivity. Although many mathematical formulas have been developed to answer specific sensitivity questions about particular model structures (such as those that arise in analysis of linear-programing models), there is still no unifying approach or point of view for ferreting out which of the many parameters are most critical. A higher level of computer-assisted thinking is needed to alert the model builder to the weak points of the model.

Discontinuities, nonconvexities, and combinatorial phenomena are not yet completely under the thumbs of operations research analysts. Although significant progress has been made with such problems in the past 5 years, the halfway mark probably has not been reached.

Interestingly, the applied science community is not complaining that the mathematical problems are too complex to allow continued research progress. Progress seems slow, and the power required certainly is escalating, but there does not appear to be any din of discussion among management scientists and operations researchers centering on the few major unsolved technical problems that persist in defying successful attack. Rather, the lament is that problems currently under study are old-hat and of less intrinsic interest than those addressed in the early days of logistics research.

Without judging the validity or propriety of this lament, we argue in the next section that many important research tasks remain to be faced in the coming decade. As will be apparent from the discussion, the starting point for many of these topics is not the previously made generalizations on the

classic types of logistics models. Rather, the recommended approach is redefinition of the remaining problems, taking into explicit account the pressing needs of logistics managers. We propose a renewed and vigorous look at managers' topical problems rather than previous researchers' leftover problems.

THE CHALLENGES THAT AWAIT

A View to the Practical

In analytic research into logistics decisions, management scientists and operations researchers have been inclined to let the mathematical formulation of a model dictate or suggest the appropriate mode of analysis. For example, when decision problems have been posed in terms of dynamic-programing functional equations, then, generally, researchers have explored mathematical and computational ways to solve the functional equations. In inventory-control models, research has focused on ascertaining the form of an optimal policy and determining the computational implications of exploiting this knowledge of the optimal form. Similar illustrations could be cited for other types of probabilistic applications. Unfortunately, even after an initial mathematical formulation has been simplified by taking account of analytically derived information about the form of the model's solution, the complexity and the computational burden remaining is not trivial. As a result, applications of many such models have been limited, and sometimes even nonexistent.

An alternate approach, which is beginning to have some currency, is to derive simple but close analytic approximations to the original model. These approximations are easier to handle computationally and are therefore much more attractive from an applications point of view. (An example will be provided in the next section.) In most real-life situations the data required by a model are themselves approximate, by the very nature of their historical base. Hence, the degradation of economic performance due to analytic approximation may be negligible. Imperfect information typically overshadows the analytic approximation as a source of model error. Although numerical approximation is a seasoned topic in computer science and, to an extent, in statistics (by way of curve fitting), the subject is relatively new in operations research. It offers considerable promise and may make practical the solution of many models that have been discarded earlier as computationally unwieldy.

A related technique is to derive analytic models with parameter values that are numerically fit from a limited discrete set of optimal points (policies). These fitted relations permit interpolation of intermediate parameter settings. In other words, the researcher starts with a grid of parameter values, performs the detailed model optimizations to derive the best policies for this grid, and then fits an analytic function of the parameter values to the set of numerical policies.

A similar vein of research is to discover the actual sensitivity of optimal policies to various parameters of a model. Evidence is building that many models that appear to involve multivariate optimization can without much loss be factored into separate optimizations, each requiring an easier manipulation of fewer variables.

In summary, considerable future research will be turned to investigating the numerical properties of logistics models, with emphasis on parameter settings that are relevant for actual applications. Such investigations will result in computational models that are simpler to use and thus will enhance the applicability of the models.

Breakdown of the Boundaries

Perhaps the most important of all the new avenues for future research will be modeling efforts that combine heretofore separate investigations of logistics decisions. Examples abound in military logistics systems. There are, for example, significant economic tradeoffs relating to initial procurement, spares provisioning, location of repair facilities, design of component parts, and installation of data collection systems to track weapon-system performance. Similar illustrations are easily cited in commercial organizations. For example, a manufacturing company must balance off considerations of labor stability, the buildup of seasonal inventories, the location of such inventories, the mode of transportation to customers, the frequency of delivery in relation to

the capacities of transport vehicles, and the targeted service performance (that is, availability of stocks and promptness of delivery).

A bottom-up approach for investigating the interactions among logistics functions does not seem as promising or as practical as a top-down approach. In constructing a top-down model, however, a researcher should keep in mind the operating characteristics of low-level logistics models and include these characteristics in the formulation of the high-level model. For example, if a segment of an inventory system has a square-root relational dependency on the annual demand for the encompassed items, then that system's numerical phenomena should be included in the model specification.

Because of the inherent complexity of multifunction models, a successful analytic approach may involve exploring only a set of case studies rather than seeking some sort of global, or even local, optimum. In other words, the model builder may have better success in investigating plausible solutions and, with feedback, refined versions of the alternatives, than in trying to simplify the interconnections in the mathematical structure to permit "automatic" optimization algorithms. The case-study approach to integrative analyses also facilitates the inclusion of discontinuous economic and physical phenomena. After the number of high-level decision options has been narrowed to a select an attractive few, then the now-familiar lower level model-building approaches can be brought into play to refine the analyses if need be.

The Human Side of Systems Design

It is surprising, perhaps shocking, that virtually no research attention has been given to the human factors aspect of modern logistics systems design. If logistics research is to become part of the warp and woof of an organization, attention must be given to the organizational setting, including the assignment of responsibilities. For example, even if model builders succeed in breaking down the boundaries between logistics functions, little benefit will result if there is no corresponding integration of management logistics responsibilities. In a manufacturing company, the links between sales forecasting, production planning, and materials purchasing are critical to the economic functioning of each of these activities. A comprehensive logistics model would combine the three elements, but the model would not produce results unless the three functions were controlled by a consistent corporate-wide logistics management policy.

The organization of most logistics operations in an enterprise is based on historical evolution; changes have taken place, if at all, typically at times of crisis. Yet almost always large improvements can be made as a result of a comprehensive look at the logistics needs of the organization. More often than not, much of the improvement devolves from realinement of responsibilities along with appropriate management review and control, rather than from revision of isolated decisionmaking processes, such as production scheduling. In other words, most separate logistics functions fare pretty well given the organizational constraints under which they operate; any noteworthy improvement comes from breaking down some of the constraints.

Considerable future research effort is required not only in thinking through organizational structure, but in examining effective approaches to personnel motivation, the communication of information for decisionmaking, and management review and control, insofar as these human activities bear on the design of integrative logistics systems. Logistics personnel in most enterprises are prone to a "beat-the-system" attitude; this proclivity should be recognized explicitly and factored into the systems design process.

Finally, even assuming benign attitudes within an organization, researchers must explore ways to improve the interactions between personnel (managerial, staff, and clerical) on the one hand and computer-driven data systems on the other. The notion that a computerized logistics system is conducive to easier decisionmaking is too naive to be of value. In fact, a computerized approach often seems to make some jobs harder and others duller. Rarely does the implementation of such a system result in an upgrading and simplification of jobs throughout. The commonly expressed negative attitudes about computer systems in large organizations are grounded in considerable experience, and the root causes call for careful study.

A Window on the Future

Now that 30 years of logistics research have passed, senior-level management has come to feel that it should be possible to diagnose the need for systems improvement without undertaking a major, lengthy research project. It is incredible to such managers that systems analysts are unable after a brief investigation to at least scope out a reasonable range of improvement potential from contemplated systems revisions. But strange as it may be, management scientists and operations researchers have made little progress in devising powerful diagnostic tools. That should be given priority. The effort will have to be empirically based in part, at least insofar as the suggested approaches should stand the test of actual field validation. The purpose of these diagnostic tools is to provide management with estimates of the future benefits of a commitment to invest in systems revision. A top-down orientation would seem to provide the proper perspective.

A similar, possibly more technical topic is study of methods for predicting systems performance when new decision rules are to be used. In this context, suppose that a proposed design has been worked out in detail, but that some of the parameter settings used in the design remain under investigation. As an example, perhaps the frequency of data revision and file update is in question. Systems performance characteristics often are investigated by means of a simulation. Such simulations usually are computer models themselves, but sometimes, especially in military systems, they are onsite tests. Little scientific research has been done to establish the validity of these predictive approaches. Practical considerations frequently rule out routine application of classical statistical design-of-experiments methods. In the methods commonly used in practice, often a bias exists that makes a proposed system design appear to perform better than it will in fact. The source of the bias is easy to detect, once one is alert to its possible existence, but correcting it may be difficult. In admittedly oversimplified terms, the bias arises because the new design itself has been fashioned according to historical data, and therefore it appears to perform well in historical perspective. The inescapable difficulty is that of necessity many models are driven by historical information that may be so limited as to prohibit using a "split-sample" approach to validation.

A related need is for monitoring devices and early warning controls that automatically determine when a new systems design revision may be warranted. Presumably, if progress is made in fashioning diagnostic and predictive tools, the way will be paved for the devising of continuing controls that automatically determine when a new systems design revision may be warranted. Presumably, if progress is made in fashioning diagnostic and predictive tools, the way will be paved for the devising of continuing controls. Here too, a top-down approach seems appropriate. It may be very difficult to detect any systems' performance degradation by looking at individual components one by one. Sensitive aggregates, if such can be found, are needed.

Disaster Insurance

Mathematical programers have learned an important lesson that should be noted by all model builders. A single-criterion optimization model typically pushes to the greatest extent possible each simplifying assumption in a model. For example, if a nonlinearity has been approximated, the optimization process will find how to exploit the *approximation*. As a result, the solution may strain the assumptions beyond credibility and usability.

To the extent that logistics research model building will break down the barriers between functions, as proposed earlier, care will have to be taken that the resulting solutions are not "too tightly tuned." The organization must be able easily to buffer unexpected (unmodelled) events. It is likely that second-best (less-than-first-best) strategies may be preferred if they do not force the organization into assuming a confining posture. Observers of real organizations recognize that most managements, usually with good reasons, shy away from strategies that have serious downside risks. Aside from recognizing the existence of multicriteria problems, management scientists and operations researchers have not made much progress in discovering the sensitivity of strategies to criteria that recognize and avoid downside risks.

The goal-establishment problem is not solely technical; it also concerns the organizational issues mentioned above. The enterprise must build in buffers, by a careful structuring of the organization, to absorb unplanned-for shocks. To illustrate, the production management component of a system may need to have a backlog of maintenance projects to fill up slack time that may arise when the marketing organization has been overoptimistic in its forecasts of sales.

To the extent that approximate models will be devised, care must be taken that the recommended decisions do not degrade too badly when the model's assumptions become invalid. For example, even though there may be very little lost in the original optimization model when a parameter is misspecified, the same need not be true in the approximative version. The chief source of misspecification in real applications is the uncertainty about future demand, failure rates, procurement costs, transport reliability, and so forth.

Getting the Job Done

The process of systems implementation deserves attention in its own right. It has become apparent that the full process of implementation has many components, some of which concern the nature of the decision problem, some the organizational setting, and some the support systems design. It is important that a framework of analysis be established to piece together the essential components, namely the decisions affected, the targeted benefits, the downside risks, the assignment of responsibilities, the development of the systems approach, the education of managers and support staff, the inherent life cycle of the application, the specific systems design the required data, and the model's validation.

In addition, it would be helpful to examine managers' psychology with regard to systems' development authorization—for example, how do they view associated career development hazards, assess the reasonableness of a project's timetable, decide whether the design will be useful, and avoid being embarrassed by an unsuccessful outcome.

The proper methodology for studying implementation is itself a research issue. The term "implementation" actually presents a problem of definition and, in any event, implies a value connotation in that agreeing to implement is normally presumed to be good and failing to implement to be bad. To make sense out of implementation processes, researchers must establish standards of comparison that are legitimate within a single organization as well as across organizations.

Summary

This section has touched on a number of avenues of research in logistics systems design that could have significant impact if successfully pursued. In looking back over the list, it is clear that the suggestions are not aimed at particular types of logistics decisions. They are aimed, rather, at a type of approach that cuts across individual logistics decision areas. Hopefully, the list makes clear those challenges that stem from recognition of organizational and managerial needs in relation to unsolved and mind-boggling technical puzzles. Assuredly, the suggested research areas are replete with tough analytic tasks, and the technical inspiration required will not derive solely or even mainly from the methods of past applied logistics research.

A GLIMPSE AT THE POSSIBLE

Strategy for Research

A rich variety of applied mathematics approaches has become standard in management science and operations research studies of logistics processes. They include mathematical programing optimization, dynamic programing, Markovian analysis, and computer simulation, to name only the more prominent. The primary role of computers has been to perform algorithmic computations on particularized versions of mathematical programing models and to provide simulated results for (typically) stochastic systems run with special settings of the underlying model's parameters.

Interestingly, the computer has seldom been used to ferret out the *qualitative* properties of models, to provide the analog of the physical sci-

entist's experimental laboratory. We believe that substantial breakthroughs are possible in many logistics research problems that are now deemed intractible because the standard applied mathematical approaches have been pushed to their limit. We suggest and illustrate in this section how computers can be used to provide new analytic models capable of solving some currently unanswered high-level management questions.

A Case in Points

Take as an example the subject of inventory control. Over the past two decades, mathematical analysis of inventory stockage models has made great progress, and real-life implementation of inventory systems, based at least in part on the results of this modern research, has taken place. Nevertheless, when an organization considers the possibility of designing and installing a new replenishment system, senior management typically finds it arduous and time-consuming to obtain reliable answers to questions such as

- What are the effects of consolidating demands from several different warehouses into a single central warehouse?
- If system-wide demand increases (through, for example, an enlarged share of the market), what are the resulting cost and service implications?
- How much is it worth to obtain quicker delivery of replenishment orders?
- By how much will costs rise if service is increased?
- How will costs be affected by less frequent updating of information?

For some of these questions, no easy-to-use analytic formulas have been devised. For others, an answer is forthcoming only if the analyst painstakingly uses a bottom-up approach, that is, makes the calculations for each of a number of individual stockage items and then aggregates the results.

Recently an alternative analytic approach has been investigated by the author and his associates, Alastair MacCormick, Richard Ehrhardt, Ronald Kaufman, Arthur Estey, and John Klincewicz. A capsule view is provided below to indicate the nature of the research strategy.

Systems Design Scenario

Consider an inventory manager who must design a system of replenishment rules for the stockage of possibly thousands of items. Assume that the manager can specify a criterion function to determine whether one system design is better than another. Suppose that the manager has elected to use so-called (s,S) policies: when inventory on hand and on order falls below s, place an order so that, as a consequence, inventory on hand and on order equals S. It is necessary to compute numerical values for the pair (s, S) for each item to be stocked. Under widely applicable conditions, it is possible to employ an algorithmic approach that provides optimal values for (s,S), but the computations are numerous and make application to a large-scale system prohibitive. Further, the optimizing algorithm assumes that the demand distribution for each item is known exactly; this is virtually never true in practice. The manager inevitably must use past data to *estimate* the demand distribution.

The systems designer's tasks then include selecting in concert the number of historical observations to use, the frequency for repeating the reestimation process, the form of the replenishment rule, the statistical estimators to produce the demand parameters required by the rule, and the design parameters of the rule, namely, the values of s and S in our illustration. Typically the manager makes all of these choices, at least in part, according to simulations of how the proposed system would have performed in the past. In doing so, the manager typically uses the same limited data for both estimating the demand parameters and predicting systems performance.

Recognizing and Attacking the Issues

Eventually, inventory managers will have to provide the answers to the questions posed by senior management. But even before attacking top management's questions, the designer must find a practical approach to the mundane issues of calculating the rule values themselves and discovering how accurate the retrospective predictions are likely to be. Regretably, these tasks are mathematically so complex that they do not ap-

pear tractible by known methods of applied analysis.

It is possible to make considerable headway, however, by devising an experimental design approach with the further help of a computer, first postulating a set of parameter values that encompasses most of the cases likely to be encountered. For the sake of definiteness, suppose that the parameter values are given as in Table 1.

We examine a full-factorial representation of all levels of these parameters in combination with each other, yielding a total of 288 settings. Using exact computations, we find the corresponding 288 optimal (s,S) policies. Next, using standard curve-fitting techniques on these 288 pairs (s,S), we obtain numerical approximations for the quantities D = S - s and s. Specifically, we derive the equations

$$D = (1.463)\,\mu^{.364}\,(K/h)^{.498} \times [(L+1)\,\sigma^2]^{.0691}$$

and

$$s = (L+1)\,\mu + [(L+1)\,\mu]^{.416} \times (\sigma^2/\mu)^{.603}\,U(z),$$

where

$$U(z) = .182/z + 1.142 - 3.466z$$

$$z = \left\{ \frac{\mu^{.364}\,(K/h)^{.498}}{\left(1+\frac{p}{h}\right)\,[(L+1)\,\sigma^2]^{.431}} \right\}^{1/2}.$$

To test whether this approximation is close enough (near optimal), we derive the 288 approximate (s,S) pairs, calculate their corresponding expected cost using *exact* formulas, and compare the associated cost with the original optimal cost. In this design, 95% of the 288 cases are within 1% of optimal. Then we examine the robustness of the approximation by trying a number of interpolated and extrapolated sets of parameter values. (In such tests, we had equally good results.)

Thus the curve-fitting exercise provides the system's designer with an easily computed replenishment rule that depends on the economic parameters and only the mean and variance of demand. But since the mean and variance are not known in real-life applications, the next step is to ascertain how well the approximation works in a statistical environment.

Presumably, in an actual situation the mean and variance of demand for each item would be estimated by the usual statistical techniques, that is,

Table 1

System Parameters

Factor	*Levels*	*Number of Levels*
Demand Distribution	Poisson ($\sigma^2/\mu = 1$) Negative Binomial ($\sigma^2/\mu = 3$) Negative Binomial ($\sigma^2/\mu = 9$)	3
Mean Demand μ	2, 4, 8, 16	4
Replenishment Leadtime L	0, 2,4	3
Replenishment Setup Cost K	32, 64	2
Unit Penalty Cost p	4, 9, 24, 99	4
Unit Holding Cost h	1	1

by computing a sample mean and variance, from a limited history of data, and substituting these values into the approximation formulas. Again for the sake of definiteness, suppose that the designer wishes to investigate three possibilities: updating s and S (by recomputing the historical mean and variance of demand) every 13 weeks, or every 26 weeks, or every 52 weeks.

We can test the performance of the approximation rule under these different circumstances by running a computer simulation for each possibility. In particular, we again can choose a factorial design for the parameter settings, simulate the use of the rule for a sufficiently long history and for each of the three revision possibilities, and at the same time simulate the retrospective approach to predicting the future performance of the rule. In summary, we found that systems costs increase, on the average, by 20% above the optimal with complete information when only 13 weeks of data are used and variance/mean = 9; by 11.5% when 26 weeks are used; and by 6.3% when 52 weeks are used. For these same three cases, the forecast of systems cost performance are, respectively, 25.1%, 17.1%, and 10.7% *under* the actual values; interestingly though, most of the underestimation comes from the service (stockout cost) component, and the separate predictions of inventory and replenishment costs are typically less than 2% under the actual values.

Finding Systems Response Functions

Next we are ready to obtain simple-to-use analytic expressions for the total costs of using the approximate policies. We again employ for this purpose a curve-fitting approach. For the situation in which the mean and variance can be exactly specified, we derive

Total Cost $\cong 5.663h\mu^{.4495}(L+1)^{.2528}(p/h)^{-.9230/p + .1357}(K/h)^{.208}$, assuming that variance/mean = 9. Similarly, when 26 weeks of data are used to estimate the mean and variance, we find

Total Cost $\cong 3.798h\mu^{.4309}(L+1)^{.3024}(p/h)^{.2550}(K/h)^{.1917}$.

These cost functions provide the needed wedge into the problem of answering senior management's questions about forecasts. To illustrate, if mean demand doubles, total cost will increase by 20% in both cases. If, for example, the demands from eight independent and identical warehouses are consolidated in a single central warehouse, total cost will be reduced by about 68% in both cases. If service protection is increased from 0.9 in-stock probability to 0.95, it can be demonstrated that total cost will rise by 25% in the statistical environment. If leadtime is cut in half at the expense of doubling setup cost, then total cost in a statistical environment is reduced by 7% (*after* the higher setup costs are paid). If the system is updated only half as often, total costs may be reduced substantially; for example, if inventory costs are charged on end-of-the-review-period levels (as is frequently done for the property tax valuation component), the cost reduction is near 40%.

The above discussion has focused on total costs, but similar systems-wide approximations have been derived for each of the components of total cost and other operating characteristics.

Summary

What this abbreviated survey of recent inventory research advances has demonstrated is the way in which seemingly intractible mathematical problems can be solved by empirical and statistical investigation. Like any experimental approach, the suggested research strategy requires careful prior planning and sufficient completion time. The impressive tightness of the approximations, however, is encouraging.

EXPECTATIONS FOR THE FUTURE

Perceiving the Sector Factor

Unquestionably there are important differences between the private and public sectors in solving real logistics problems. The obvious differences are related to the sheer possibility of truly integrating separate logistics functions, the limited budgetary and personnel resources for systems redesign, and the fiscal constraints on any implied multiyear spending. Beyond these are differences in the basic missions of the logistics function. In a commercial enterprise, the logistics

decisions support the buying, making, and selling functions and rather clearly lead to an eventual profit-and-loss impact. But in a military environment, the logistics mission is highly intertwined with the critical notion of combat readiness, which in the final analysis is only rarely tested and then under crisis circumstances. Perhaps ironically, it is in a military setting that the top-down approach to logistics is most essential, because very large sums of dollars are committed by the logistics decisions, and these must be balanced off against dollars spent on other military readiness functions.

Watching the Sign Posts

A truly telltale criticism of past management science and operations research investigations into logistics functions is that they rarely reflect timely economic issues. To illustrate, one is hard pressed to find in the applied-mathematics-oriented logistics research literature a careful discussion of the impact of inflation, the limited availability of fuels and other strategic resources, or the rate of technological change. However, actual logistics managers are painfully aware of these environmental changes and their impact on logistics decisions. Logistics research will only stay vital if it pays heed to the changing world.

Generating Viable Options

It is virtually a tautology to say that a formal logistics decision model encompasses a static universe of options. The solution drawn from this universe by the model may or may not yield a recommendation that can be implemented, but if the solution is unacceptable the analyst always can go back to the drawing board, revise the model, and try again. What is more important to the search for significant progress in logistics decisionmaking is to concentrate on discovering truly new options. Without sinking into a philosophical quagmire of subtle distinctions, we suggest that analysts pay more attention to relieving constraints, finding new conceptions and criteria, combining separate processes, and so forth than to searching for the very best answer within a well-established framework of concepts, laid down constraints, and circumscribed functions.

Substitution at the Margins

A related topic is the necessity that a wide view be taken of the important substitution possibilities. For example, there are tradeoffs between computer information systems and skilled labor, between large stocks of disposable spares and limited stocks of high-technology components, between fast modes of transport and extensive amounts of inventory, and between rapid communications systems and multiple pipelines, to name a few. The point is so obvious that it may not seem worth making, except that most logistics research takes place in a very limited context. The analyst may be either proscribed from examining such tradeoffs or ignorant of their existence and feasibility. Thus, one function of senior management is to encourage logistics staffs not to be too circumspect in considering possibilities. An ancillary observation is that a logistics organization making such investigations must have access to a broad spectrum of skills and knowledge.

Next Up

In summary, this survey has attempted to realistically assess both the strengths and the limitations of logistics research to date and to generate excitement and enthusiasm for the worthwhile but difficult tasks ahead. Our prognosis is that substantial advancements will be made in the coming decade by researchers who focus on problems at the traditional boundaries of the logistics functions, who keep abreast of the changing outside environment, and who break away from sole reliance on the well-worn applied mathematics techniques that have already run their courses with regard to many now-classic logistics problems. None of our exhortations is meant to detract, however, from the unassailable value of building on past research momentum. We have tried, rather, to indicate where we think some of the still-buried great treasures are to be found in the next 10 years of logistics research.



Marvin Minsky has been Donner Professor of Science at the Massachusetts Institute of Technology since 1974. Dr. Minsky has also had appointments at MIT as Professor of Mathematics and Professor of Electrical Engineering. In 1959, he was cofounder of the Artificial Intelligence Project at MIT. In 1964, the project became the Artificial Intelligence Laboratory, and he served as codirector from 1964 to 1973. Dr. Minsky earned a B.A. from Harvard and a Ph.D. from Princeton. He is a Fellow of the Harvard Society of Fellows, of the Institute of Electrical and Electronics Engineers, of the American Academy of Arts and Sciences, and of the New York Academy of Sciences. He is a member of the National Academy of Sciences.

AUTOMATION AND ARTIFICIAL INTELLIGENCE

Marvin Minsky

Massachusetts Institute of Technology
Cambridge, Mass.

The uses of robots and machine intelligence have long been popular subjects of futuristic literature. This paper explores applications of computer science to real-world problems of this type.

I will not discuss the broader consequences of building intelligent machines. This would be too difficult, too speculative, and—frankly—too scary. Instead, I will focus on more conventional prospects of automatic machinery in industry and in everyday life, and argue that while advanced automation is still very primitive, it contains the seeds of several more industrial revolutions.

ONR has had a substantial role in the history of this area. There is no need to recapitulate its central role in the emergence of modern Mathematics in several countries, but perhaps not so well known is ONR's imaginative support of early cybernetic and computational theories. Along with that of a few others, notably the Air Force Office of Scientific Research (AFOSR), this agency's work was critical when discriminating and sensitive understanding was most important.

ADVANCED AUTOMATION

Everyone knows that "automation" means: using machinery to "automate" jobs done by people. However, there doesn't seem to be a word for using machines on jobs that weren't done before at all. This paper will discuss several of these.

We can envision automation in various ways. For our purposes here, it is natural to think about the extent to which the machine incorporates "intellectual processes." From this viewpoint, one sees several stages, with increasing technical problems.

Stage 1: Remote Manipulators—Direct Augmentation of Human Control

The most primitive form of automation is of course the handtool, which augments a person's strength, speed, or precision. Handtools require the operator to be close by; modern servomechanisms allow the operator to be far away. This is the "teleoperator" concept, which plays a large role in this essay. Its prototype is the remote manipulator, a device that senses human arm and hand motions and duplicates them at a remote location. The motives for its development were the problems of handling radioactive materials. The first such systems were mechanical pantograph linkages; later improved (notably at Argonne) with "force-reflecting" servomechanisms that allow the operator to "feel" what happens at the remote hand.

Teleoperators surpass handtools in separating the operator from hostile or inaccessible environments. Remote manipulators are of enormous and immediate potential value, but they have not been adequately developed in recent years. A few million dollars spent here would soon return billions in energy-related industries. Below we will discuss some of the scientific and engineering problems in improving teleoperators.

Stage 2: Supervisory Control

By attaching a computer to a remote manipulator, we can give the human operator a less direct, more supervisory role. Rather than carrying out each motion in detail, he supervises the process by indicating goals and trajectories. Perhaps by moving his hand toward an object in a certain way, he indicates to the machine that the object should be grasped and lifted. This vastly improves the application potential:

The operator can do more work, making fewer specifications.

The system can exploit special knowledge stored in a data base. For example, the machine might know specific details of how a particular object should be handled.

Such a scheme can overcome some delay/bandwidth problems, e.g., control of an effector at satellite-communication or lunar distances.

Supervisory control requires some computer intelligence:

The ability to recognize and orient target objects

The ability to interpret the input "intention" language

Enough "problem-solving" ability to anticipate and cope with changing spatial relations, inertial phenomena, gravity, etc., without concerning the operator.

The problems of making a computer deal with "commonsense" physical knowledge about such things as spatial relations, support, trajectories without interference, etc., are more difficult than they seem at first, and they have been major concerns of "artificial intelligence" projects at MIT, Stanford, SRI, and other centers. The theory of supervisory control servomechanism processes has seen much development.

Stage 3: Autonomous Robots

Finally, there are machines without human supervision. In a sense, conventional assembly line machines are autonomous already. However, we are concerned with a qualitatively larger range of responsibility and flexibility:

Versatility. Assembly lines require separate machines for almost every operation. General-purpose manipulators could reduce costs and time by doing many jobs at each place.

Tolerance. Modern production systems are based on uniformity of components and placement, which can be costly. More intelligent assembly machines could reduce some of these costs. and also perform new inspection and quality control services.

Tailoring. Mass-Production imposes annoying uniformity constraints on goods. Indeed, the word "mechanization" today usually implies undesired constraint. With the Industrial Revolution, many more people could obtain tolerable clothing, for example, but all but the wealthiest lost access to individually tailored clothes. Now we are ready for a restoration. A computer should have no trouble remembering personal measurements and calculating optimal seams and darts, and mechanical hands could be made to sew clothes that fit. The same potential for marrying automation with the craftman's skill at personalization exists in the furniture industry as well.

Planetary exploration. The landing of the Viking planetary probe was autonomous; the time-delay prohibited real-time control from Earth. Subsequent scientific operations involved only slightly less autonomous operations. Supervisory control will not be suitable for the much broader Martian explorations that must be made.

Transportation. The ecologically motivated pressure for mass transit threatens to become as constraining as the early products of "mass production." It would be tragic to repeat the same mistake. The automatic automobile could become practical in a decade or two. It could be far more efficient than any system that moves masses of people to places they don't really want to go. Only "transit-tailoring" will solve other problems in transporting children, the elderly, and the handicapped. We will return to this later.

PROBLEMS AND PROSPECTS

Artificial Intelligence

Mechanical autonomy needs artificial intelligence ("AI" for short), the name for scientific study of theories of the nature of intelligence. "Cognitive psychology" is another name for the same thing in the specifically human context. The difference is one of orientation and application. AI points mainly to intelligent machines, while cognitive psychology points to questions of how human "computers" work. There is no room to review here the status of these fields, but see Winston [1975] and Minsky [1977]. I can only discuss a few fantastic, yet practical, applications that seem almost within reach.

What does "practical" mean? Is a device practical if no one cares to build it? The question is whether economic demands will focus on the kinds of technology proposed here, which promise new approaches to individualization (and thus a higher "quality of life"); efficient use of energy and materials; and sources of knowledge, energy, and materials.

However, these goals will not be realized unless more people and agencies understand their importance and value.

Teleoperators

Already there are many machine-controlled "mechanical arms" on the industrial market. Typically, such machines have two to six degrees of freedom, strength larger than human speed comparable to that of a human operator, and precision in the millimeter range. They cost about $10,000 or more. These "industrial robots" are seeing increasing areas of factory use, but they are not quite yet a major factor in modern industry because it is not easy to apply them to most jobs. The commercially available "hands" are very crude; production engineers who buy the robots usually must design new hands for them. Predetermined motions require predetermined locations; the available systems permit only very limited options in response to contingencies. Little or no touch and force feedback is provided; production engineers must provide their own instrumentation, then design a computer system to use that information. Because the machines lack force-controlled feedback, they are generally unsuitable for delicate assemblies.

On the other hand, production engineers and "assembly-machine" designers are very good at jigging together actuators, clamps, and linkages to perform preprogramed operations for assembly lines. They do not find it cost effective to adapt a general-purpose industrial robot to this; it is more difficult and expensive than using their current bags of tricks. To develop more generally useful automatons, we need better technology, whether the machines are used for human amplification, supervisory control, or autonomous operation.

Input Sensors—Present remote manipulators are usually controlled through something like a scissor- or pistol-grip device. There is a need for a more sensitive, versatile way to sense more of the operator's hand motions, pressures, and tensions. The same input device must also signal back to the operator what is happening at the working end. For indirect control through a computer, the requirements for tactile feedback can be dropped, but the computer must have feedback information in some other form.

Output Sensors—the sensors at the output should be able to sense touch, pressure, textures, and vibrations and transmit them back to the operator control input device. This is profound problem in engineering. I am convinced that some sort of superglove that could do this without discomfort and clumsiness is a realistic possibility.

Output Motor Control—These problems are somewhat better understood, but no one yet knows how to build a motor hand with anything

like human dexterity and articulation, although R. Mosher's work at GE went far toward this goal. Elaborate computations are necessary for converting the pulls of gravity and inertia in a massive hand into signals that resemble those from a human hand, and this complicates force-sensing at the working surfaces. But we still need a deeper "control theory" of stability for servo-systems with a skeleton-like number of degrees of freedom. I don't believe that current knowledge is adequate to stabilize such mechanisms.

To break the logjam, industrial robots must be made much more "general." They need powerful computer programing systems and much more versatile sensors and actuators. They also need, to be economically practical, the benefits of mass production. The most important breakthrough, though, will come from the use of artificial intelligence programs to help users develop the sophisticated computer programs that a "robot with common sense" must have.

UNDERSEA TECHNOLOGY

Most of our planet is ocean, but we know little about it. The hazards to human life in the depths are (and will remain) so intense, and the importance of learning more is so great, that this should become an outstanding area for robotic development.

Continental-shelf Drilling

In the next decade, we will need a better technology for exploiting continental shelf oil. The expense and danger of ecological accidents at present inhibits this, and quite properly so. On the other hand, such exploitation is demanded by the worldwide energy crisis.

Undersea oil spills are dangerous, expensive, and wasteful and must be corrected quickly. There is no effective way to seal off an undersea fault; seepage can rarely be stopped quickly with relief wells, and sometimes relief wells cannot stop seepage at all.

With teleoperators, the equivalent of an "undersea construction crew" could be developed, with experts in comfortable offices working through remote devices just as if they were a conventional ground crew.

With such technology the costs of site preparation and maintenance could be far lower than the costs of the current weather-troubled ships and expensive fixed tower structures.

Perhaps the best way to approach this might be to develop an approximately humanoid robot, controlled by an instrumented wet-suit, to make control as natural and comfortable as possible. In some ways, the undersea problem might be simpler than its terrestrial counterpart, because bouyancy can be used to neutralize weight-compensation problems. For undersea work, sophisticated teleoperators should be adequate. In most cases, supervisory control is probably not necessary, except perhaps where visual bandwidth problems become serious.

Undersea Exploration

Commercial site-preparation technology should lead to more mobile exploratory facilities for better understanding of the sea. Many feel this will be the key to understanding the planet in general. Such experimental vehicles as the ONR's ALVIN have made large contributions, some of which can be credited to its teleoperator arms and hands.

Manned exploration of the depths is technically as difficult as exploring space. However, those complex and courageous expeditions in which men descend thousands of fathoms, insulated by massive mechanical shells—bathyspheres, bathyscaphes, and bathyboxes—resemble Apollo more in its weaknesses than in its strengths. There have been no "moonwalks" at a thousand fathoms: Manned pelagic exploration is harder than manned lunar exploration, and the super-submarine does not solve the problem.

Undersea Mining and Industry

A versatile, mobile pelagic exploratory laboratory will surely uncover new resources, many at greater than continental-shelf depths. Perhaps there are chemical syntheses or material

fabrication processes that would proceed more economically at pelagic pressures; factories might be situated in the deepest places.

Hydrothermal Energy

Hydrothermal energy is the largest terrestrial energy pathway; most of the earth's solar energy is "processed" by the sea. If exploitation of thermal gradients becomes important, undersea robots will surely play an important role. The proposed vertical heat-cycle engines, for example, have problems with fouling of intake and circulation exchange systems. Chemical remedies on a large scale have ecological problems, and the solution might well involve robot maintenance. Indeed, if biological fouling is really a major problem, it should be possible to exploit it as a by-product.

I don't know if anyone has considered mechanical exploitation of the deep and slow, but vast, ocean currents by such means as undersea "windmills" in the Gulf Stream. Robot maintenance might be the key to making such a system practical.

Aquaculture

Mechanical cultivation could yield vast vegetable and animal crops in ocean areas. A side effect of deep hydrothermal plants could be fertilization of the surface milieu by nutrients moved from deeper strata. In any case, mechanical aquaculture using teleoperators and, eventually, autonomous "farmer robots" would seen an important area for research.

Rescue

Submarine rescue is notoriously difficult; so is retrieving nuclear materials from misplaced weapon systems. This is an obviously cost-effective use for even expensive teleoperators. To be sure, such systems already exist, but my impression is that they are too clumsy.

ROBOTS IN INDUSTRIAL PRODUCTION

The Assembly Line

Modern assembly line production is like a tree; finished material flows toward the root, and parts are combined at branch points. The factory itself, however, need not be so organized in space. If working sites had more general-purpose automata, then more steps could be done at each location. Dextrous robots could throw and catch materials and so break free of conventional layout constraints.

An intelligent general-purpose robot could, for example, assemble a telephone or a typewriter from a kit of parts, testing subassemblies as they are completed. Prototype systems of this sort already exist. The problem is that for very large volume production runs of a uniform item, it would be hard to compete with special-purpose factories like the plants that today mass-produce items like telephones and typewriters. For other purposes, though, we could have in a generation or so an assembly robot that would observe assembly once or twice, try to do it itself, perhaps ask a few questions, and be then ready to go into moderate-volume production.

THE NUCLEAR INDUSTRY

Reactor Maintenance and Safety

The problems of dealing with radioactive materials daily become more critical as we grow more dependent on them and as the quantities involved grow more massive. Nor does fusion power promise "clean" energy in the next era. Most of us already know about the dreadful combination of circumstances that make each problem worse than the others. Problems of radiation-shielding and disposal of waste materials are extremely serious. Very high temperatures weaken structural materials. In fact, they exclude most materials entirely. The high flow rates needed to transport the heat impose substantial forces on the weakend structures. Radiation causes cumulative structural damage, leading to interior flaws, surface corrosion, and the like. Onsite inspection for these is difficult and hazardous.

The aircraft industry has achieved an outstanding safety record by adopting an expensive and meticulous schedule for frequent inspection of critical components. The powerplant is disassembled and inspected regularly, yet this seems to be cost effective.

In the nuclear industry, no such frequent shutdown, disassembly, and inspection of each reactor is now envisioned. It would take an extraordinarily long time using the teleoperators available today. In fact a "spill" that would be considered minor in a chemical plant could cause a shutdown of many months in a reactor.

Fuel Reprocessing

There are similar problems in connection with fuel reprocessing and effluent extraction and treatment. At this writing, there is an increasing shortage of such facilities, with no prospect of relief for at least a decade! I feel certain that the unavailability of a new generation of versatile teleoperators is in large part responsible for the reluctance of industry to even try to build such facilities. At the moment, no one wants very much to do it.

The problem can thus be seen in terms of two opposing forces: (a) Long component life and high reliability are required because normal, routine maintenance is out of the questions; (b) The extraordinary materials problems make it too hard to achieve long component life and reliability. Our inadequate tools add another cost. The mechanical design of nuclear equipment is constrained by the requirement that it be serviceable, to whatever extent possible, by the available teleoperator claws.

These inspection and maintenance problems, in my view, could be greatly alleviated by better teleoperators. Ironically, most early development of teleoperators was done by workers in this area. But research support for this dwindled in the 1960s despite forecasts of mounting problems.

SPACE

The success of the Viking landers shows how much can be done with autonomous control. Transmission delays on the order of an hour, round-trip, prevented direct teleoperator control of landing. Later operations were more supervisory in character, performed via hour-long "move, wait, and see" cycles.

Near-Space Exploration

Everyone surely realizes by now how much we could have learned about the Moon if one of the Apollo missions could have left even a simple remote vehicle in operation. The Earth-Moon transmission delay is small enough to make direct teleoperator control effective, and relatively primitive equipment could have been used. (It is curious that the successful early Soviet Missions using this idea were not followed by more.) Remote vehicle walks of the order of a kilometer per day would be feasible; by now we could have surveyed a substantial part of the surface.

The advantages of space technology using teleoperators were suggested in Robert Heinlein's prophetic "Waldo" (1940). The use of such devices for industrial fabrication and for prosthetic use are also predicted in this novel.

It was often pointed out during the Apollo era that machines could not replace men for all purposes. Most of those arguments were quite weak. So far as lunar exploration was concerned; teleoperators could have done quite well. (I have no quarrel, however, with the nonscientific motivations of manned space exploration.) The miraculous completion of the flawed Apollo 13 mission, however, must be credited mainly to the teleoperation of the ground crew. The internal instrumentation was inadequate for the flight crew even to find out how serious the damage was.

Space Stations

There are many reasons why substantial space stations in earth orbit would be useful; well-known is the proposal of Gerald O'Neil to build colonies of about 10 000 persons to operate and maintain solar power stations and factories. Regrettably, the economics of this seem implausible; teleoperators might reduce the costs by a huge

factor, eliminating the vast life-support requirement.

Nonetheless, large near-space capabilities might indeed be profitable in the energy and fabrication fields and would open the way toward more thorough exploration of the solar system. Until the development of true artificial intelligence, exploration of the planets might best be done by using large manned orbiting spaceships controlling ground-based teleoperators.

As for interstellar exploration, the alternatives are self-contained colony ships or autonomous explorers using artificial intelligence. Both options could be available in a century or less; the colonies pose massive engineering, social, and psychological problems. The scientific problems of artificial intelligence cannot yet be fully anticipated. Even if we thought we could build a suitably intelligent computer, we would have real problems in "validating" it, and no one would want to entrust a billion-dollar interstellar craft to a potential "HAL."

DOMESTIC AND REAL-LIFE APPLICATIONS

Home

This is clearly one of the largest "markets." Housecleaning and household management are, perhaps the largest scale unproductive human activities. But an unintelligent helper is usually worse than none. It replaces physical effort by administrative effort, and the latter is (at least to some people) even more burdensome.

The mythical housecleaning robot poses, in fact, higher technological requirements than most industrial, military, and scientific applications! Nevertheless, I believe the next generation will see the beginning of moderate-cost machines that are able to:

See enough to recognize objects and configurations usually found in a household. They should also "know what they don't know" so as not to damage unfamiliar structures.

Handle objects with dexterity. Progress in industrial effectors should make possible mass-production of low-cost "pairs of hands" adequate for most household jobs.

Understand what they see and feel. The household robot will need software based on commonsense algorithms. Every normal person has huge files and procedures in his head that tell—for each common object—something about where it belongs, how it may be handled, how to tell (to a degree) when its present context should not be disturbed, how to clean it, and even how to maintain it if it needs regular attention.

Why not begin with things we know how to build, such as automatic lawnmowers that follow preprogramed paths or buried wires, and floor cleaners that do simple tasks like dusting and vacuuming? The answer is that all those separate appliances would leave most of the work undone. Eventually the high-technology, general-purpose, computer-based robot must cost less. Once the intelligence, sensors, and dexterity are here, the rest is software. And while that software may be enormously expensive to create, it can be duplicated indefinitely without cost, save for the memory cost that we all expect to decay exponentially for a century!

Besides mowing lawns and sweeping floors, such robots could be taught to sew and cook, to file and keep accounts, to sort waste and reclaim much that is now wasted. They could increase our effective individual wealths by making possible sharing of goods among households, reducing everyone's capital and materials investments in the things that can be so shared.

Entertainment

This is surely the next largest "market" of human activities. Computer games are replacing pinball machines, computer animation is infiltrating the film industry. The computer itself is becoming the basis of a new, highly developed hobby field. Through "networking," I expect to see a whole spectrum of social activities take shape. They will engage the handicapped for the first time, and they will cross language boundaries with machine translation. While the precise shape of this future cannot be foreseen, most humans will surely continue to spend most of their

energy and resources outside spheres considered directly productive.

Office

The administrative environments are already changing. Most "computer-aided" services are less help than one has a right to expect. But, as artificial intelligence develops toward common-sense responsiveness, personal files will begin to understand what is wanted of them; networks of these will be shared among people with common interests, and the physical forms of offices and places of employment will mutate beyond recognition. A million handicapped persons will return to our work society, while countless others will choose to withdraw into more remote activities.

Transportation

We have noted that individual automated transport could become available to children, the elderly, and the handicapped, while many other transportation needs will diminish. The possibility for white-collar people to stay at home if they desire is obvious once interactive network systems are available. The ability of production workers and engineering professionals to be where they choose also grows as teleoperators improve. Even now, there are relatively few actual persons in mines, and there will be fewer in the future.

Prof. John McCarthy of Stanford University has convinced me that the automatic car should be our goal for the next century. We have already invested on the order of a trillion dollars in the roads and other facilities that make possible individual transport. A competitive investment in "mass transit" would be an economic disaster to a generation that doesn't use it.

Individual automatic cars will require some artificial intelligence to be sure. A foolproof vision or other sensor system to ancicipate accidents will be necessary. Modified military sensor systems will make pedestrian-detection standards higher than those of the best human drivers, and obviously far better than those of average and incompetent drivers.

A computer network capable of efficient routing and scheduling, with a thorough understanding of potentially dangerous configurations, will prevent waiting at intersections by grouping traffic into suitable packets.

Flexible sharing of the vehicles will make the individual capital investment modest—no more than at present, say—while permitting much higher investment in the safety and efficiency of individual vehicles.

For those who enjoy driving as sport or entertainment, all is not lost. Manual operation with emergency computer takeover could be available for those who will pay the slight extra computation fee. Once the system becomes foolproof, the speeds and accelerations possible could be paced to make racing drivers prefer walking.

MICROAUTOMATION

Robots could work with very large and very small things, as well as "normal" sizes. The large is already familiar; the steam-shovel is a teleoperator, and the construction crane is very like a giant arm. Biologists, at least, have long had micromanipulators, "miniteleoperators," but we do not have much general technology in the micro domain. Indeed, I have an uncomfortable impression that high technology, rather than advancing microdexterity, is bypassing it; the mechanical calculator was on the road toward microscopic clockwork, but was short-circuited by the new optical electronic fabrication methods.

So many things can be done in such small sizes with microelectronics that it is difficult to think where micromechanical systems are really needed. (Contemporary solid-state electronics do not work in high temperatures or high radiation fields, but that is another matter.) The clearest applications, perhaps, are in surgery and biology. Surgeons today can suture millimeter blood vessels, under ideal conditions. Conditions, unfortunately, are not ideal inside the brain, where the need is perhaps greatest. Access is often impossible even when the repair itself is possible. There is no reasonably versatile, touch-reflecting microhand with enough dexterity to perform such repairs; microvascular surgery needs a small teleoperator that can work along narrow passages. In

both heart and brain vessels, supervisory control would be desirable to permit ultrafast repairs and thus reduce the anoxic periods; these often preclude conventional methods that are otherwise technically feasible. Many other surgical repairs would be made simpler with a minihand that could make and enter a small incision, then traverse natural pathways. Even today stones and emboli are removed, pacemakers implanted, and viscera inspected with simple probes. The complexity of such repairs is limited, on the whole, to a narrow spectrum of cutting, crushing, and stretching operations.

BIBLIOGRAPHY

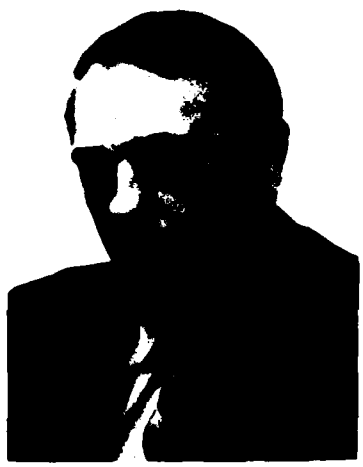
W. R. Ferrell and T. B. Sheridan, "Supervisory Control of Remote Manipulation," *IEEE Spectrum*, pp. 81–88, (Oct. 1967).

W. M. Whitney, "Processing and Storing Information," in Part 3, Management of Information, of *A Forecast of Space Technology 1980-2000*, NASA ST 387, Jan. 1976.

Robert Heinlein, *Waldo and Magic, Inc.*, Doubleday, New York, 1940, Signet, (1970).

P. H. Winston, *The Psychology of Robot Vision*, McGraw-Hill Book Co., New York, 1975.

M. Minsky, *Computer Science and the Representation of Knowledge* in The Future of Computers and Information Processing. M. L. Dertouzos and J. Moses (eds.) MIT Press, Cambridge, Mass. (in preparation).



Herbert Solomon has been Professor of Statistics at Stanford University since 1959. His earlier positions have been with the U.S. Air Force, the Office of Naval Research, George Washington University, and Columbia University. Dr. Solomon earned a B.S. at City University of New York, an M.S. in Mathematics at Columbia University, and a Ph.D. at Stanford University. He was awarded the S.S. Wilks Memorial Medal of the American Statistical Association and is a John Simon Guggenheim Fellow. He is a Fellow of the Institute of Mathematical Statistics, of the American Statistical Association, and of the International Statistical Institute.

APPLIED STATISTICS

Herbert Solomon

Stanford University
Stanford, Calif.

INTRODUCTION

In looking ahead at applied statistics, one is bound somewhat by past history and by present activity. Statistics is not an old discipline. It had its origins in the second half of the 19th century, and, of course, most of its development in this century. Two British savants of the late 19th and early 20th century, Francis Galton and Karl Pearson, stand out in their contributions and the British school of statisticians continued this preeminence. Ronald Aylmer Fisher dominated the scene in statistical inference for about 40 years in this century. Through an overlapping period with Fisher in England and until this day, Jerzy Neyman is another towering figure in statistical inference. While both engaged in an historical dispute for many years on tests of statistical hypotheses based on sample data, much of this work and their independent work on a number of other topics were directly related to and motivated by applied statistics.

Neyman received his formal training and early experience in Russia and Poland, but his major statistical contributions stem from his efforts while in England and subsequently in this country. Statistics is essentially an Anglo-American activity with important developments also coming from the Indian school, drawing on the British connection, and the Scandinavian school through their work in risk and insurance analysis. Curiously, great scientific centers in mathematics, such as those in France, Germany and Russia, have not joined the main stream of activity in statistics. This is obviously a manifestation of the culture and national ethos of these countries, and I leave analyses of this situation to those who investigate the history of science.

Statistical thinking is pervasive in many subjects currently under study. Very little is excluded from its onslaught. It has a rich tradition in the social and behavioral sciences, a recent history in public health and medicine, and is seeping into the humanities, including the law. Strangely enough, physics and chemistry are somewhat resistant to it, and its affiliation with biology is quite unusual and unsettled. One should add that its association with engineering, especially in the modern sense, is quite thick, especially in such topics as quality control, reliability, inventory control, systems analysis and operations research. There is much ferment going on in statistics in a number of these fields. One of the biggest catalysts for this is the computer.

There is no doubt that the computer has revolutionized statistical thinking and methodology in the last 25 years and especially so in the last 15 years. The kind of material published 25 years ago in statistical modeling and methodology were elegant attempts to get at approaches and solu-

tions to problems by ingenious mathematics because the modern computer was not available. This was especially true in multivariate data analysis, and some scholars of the profession, such as R. A. Fisher, P. C. Mahalanobis, Harold Hotelling, Abraham Wald, and Samuel S. Wilks, gave their efforts to this important subject. Much of this work was motivated by rather specific problems arising especially in the biological sciences, physical anthropology, and psychology

Because the latter half of the 19th century saw much data collection in these disciplines, it was only natural that investigators would try to construct parsimonious models to account for the data. This led to factor analysis models and a host of other multivariate models in regression analysis and correlation. Once models were developed, questions of goodness of fit arose and extensive efforts were given to the question of associating data with models.

Closely allied with development of models and prior to goodness of fit is the question of estimating the parameters of models. This led to estimation procedures and obvious queries as to which estimation procedures might be best in some sense. Some of R. A. Fisher's works on parameter estimation stem from queries raised by astronomers in connection with data they were analyzing. Another kind of estimation problem was that faced by insurance companies, who obviously would stratify a population by intensity of risk and base premiums on the risk in each category, but then would modify this by shrinking all estimates toward the mean. Obviously the premium associated with the best risk and the premium associated with the poorest risk were not feasible for marketing and administrative reasons, and so the estimates in each category took into account values ov observations in all categories. In this intuitive way, simultaneous estimation of parameters was accomplished.

The design of experiments permeates much of statistics. Certainly, investigators in applied fields would find the concepts and methodology involved to be of paramount importance in their everyday work. Originally it was associated with rather formal and elegant designs that were motivated by experimentation either in a laboratory or in agricultural settings. R. A. Fisher was responsible for encouraging these developments, which included Latin squares, Greco-Latin squares, factorial and confounded experiments. An important tool in the analysis of the resulting data is the concept of randomization, and this motivated the exact distribution theory of statistical tests. The analysis of variance is, of course, the methodological tool for deciding whether these experiments indicate effect. Ranking of effects, if they are found after experimentation, and selection procedures, also have a large literature. Regression analysis is an analogous technique when the predictor variables are measurable rather than categorized variables. Modern developments include much activity in directly finding a minimum or maximum value of a criterion, say cost, or yield of a chemical process, by sequentially selecting appropriate levels of experimental variables. The selection of the best subset or variables to be used in an experiment or in regression methods has interested a number of authors. Optimal designs are always under study. However, on the practical side, one settles for what one can do. In public health questions and drug effectiveness studies large scale clinical trials are usually required. These should also be categorized under experimental design. This vast field has such an eclectic role in applied statistics that it is difficult to report on its development in less than a monograph.

We will pay attention, in looking ahead, to such topics as simultaneous parameter estimation, goodness of fit testing, multivariate data analysis, and several other subjects in applied statistics in some detail. Other topics in applied statistics will receive brief mention. All this will be preceded by some introductory remarks on statistical inference. The choice of how much attention each topic receives obviously mirrors the author's mind and interests at the present time.

The Department of Defense, like other institutions in society, is an avid consumer of statistical thinking. Because of the broad sweep of its programs, statistical thinking enters in many ways. Design and alalysis of experimental data for weapon selection, recruitment policy, classification of individuals in the services, reliability and maintainability of weapons systems, behavioral studies of diverse groups in military specialties, inspection procedures for the acceptance of military items, and so forth, all show the pervasiveness of statistical thinking in defense programs.

Through their support of research and development in applied statistics, the research units in the services have made possible a large body of results of use to them and to all other elements of society. Likewise, statistical results stemming from other sources have found their way into the service programs. The Office of Naval Research in its first thirty years has supported a number of successful efforts in applied statistics. The results are available in many statistical journals and books. Together with its counterparts, the Army Research Office and the Air Force Office of Scientific Research, the research arms of the three services have aided immeasurably in the development of scientific methodology through their support and encouragement of contributions in applied statistics.

Inference

The Past. As we have indicated, much of the early work of statisticians was directed at reducing large quantities of data to summary statistics from which patterns would be deduced from observation without any precise mathematics. However, some was done in a scientific context, and this led naturally to statistical testing, and to estimating parameters. A scientist often has a theory which leads to one or more hypotheses which can be precisely formulated; and it was natural to think how a statistic can be found to test the hypothesis. Early tests like Student's t and χ^2 (chi-square) were probably devised with this type of application in mind. Similarly, in science, there will be important parameters which it is desired to estimate as accurately as possible from experiments.

Jerzy Neyman and Egon Pearson (son of Karl Pearson) began to set tests in a modern mathematical and conceptual framework. They introduced the idea of a best test against a specified alternative to the hypothesis tested. It was clear that tests and estimation procedures form a duality; the best of each nearly always depends on the same statistic. The idea of a confidence interval for estimating a parameter, that is, a range estimate for a parameter, is also a natural development, when the distribution of the appropriate statistic is known.

With confidence intervals, however, came difficulties. Although one could emphasize to a non-statistical research worker that he was given a random interval which included a parameter Θ with 95% probability, he would soon turn this into a statement about Θ, as though it had a probability distribution. It was probably an attempt to put structure behind this practice which led Fisher to introduce his controversial "fiducial probability." Unfortunately, both with confidence intervals and fiducial probability it was possible to construct examples of data which would give absurd or paradoxical answers. Often such data did not in fact appear to fit the model being used, but there was less emphasis at the time on testing this.

Another aspect of testing which did not always appeal was the arbitrariness of the statistical significance level, say at the now classical .05 or .01 level. The testing situation in any case is not always nearly so clear when one is faced with medical data, industrial data, or data from the life sciences. An attempt to put more structure into the process of decision-making through statistical testing led to Decision Theory, much developed, especially by Wald, after World War II. This has weaknesses such as how to decide on appropriate loss functions and mathematical difficulties in constructing estimators and tests of hypotheses.

Out of the search for answers to some of these problems came important ideas like sufficiency, the importance of the likelihood function, sequential methods, and robustness. The latter takes on importance when faced with data which is of bad quality, or which otherwise does not appear to fit a tractable statistical model. Robustness is a prime topic of study these days in the still burgeoning field of non-parametric inference, that is, data analysis with few or no assumptions about an underlying model.

In parallel, and somewhat orthogonal to those who wished to do independent experiments and let the data give all the answers, there has been a Bayesian school whose members would give a prior distribution to a parameter and allow this *and* the data to give a posterior distribution for the parameter after the experiment. They would find little need for formal tests or confidence intervals, but can be criticized both because prior distributions can lead also to paradoxes, and also because such a strong element of personal judgment can

enter to influence the scientific results. The same experiment can lead to different answers. Bayesian techniques have considerable appeal in some situations, for example, when one wants to bring knowledge of another experiment, say, done in another scientific center, to add to one's own results, or when it does seem reasonable that, say, past history gives a reasonable feel to where a parameter should lie.

The construction of paradoxes and counterexamples has become a flourishing industry among theoretical statisticians. How to resolve them has generated much heat, some of it, especially in the early days, apparently based as much on personality conflicts as scientific ones. Nevertheless, it seems strange that there has not been a more concerted mathematical attack on the problem of the conditions under which certain systems of inference will work well. Only Donald Fraser, it seems, has put much effort into these questions. It should be emphasized that the practical statistician rarely felt obliged to follow slavishly one or another of the schools, and his practical decisions would rarely, if ever, have been changed by their different procedures. He relied still on the big techniques: normal theory tests, ANOVA, contingency tables, regression, followed by interpretation.

The Future. It may be that time has caught up with the controversies. Since the arrival of computers, there has been a new interest in data analysis; i.e., starting with a quantity of data, much of which may be of indifferent quality, and "digging into it" to see what can be found. This is especially appealing when the data comes from evaluation attempts of large government social programs, clinical trials studies in public health and similar large-scale efforts in data collection. In these situations, problems of accuracy in measurement, reporting, and so on are huge, and the tight models behind so many of the schools of inference seem not applicable. Moreover data bases are very large and the notion of a statistical significance level becomes moot. With computers available, it is not difficult to throw out suspected methods and repeat a test; or to plot graphs of data in various ways; or to look for clusters; or to do regression or other multivariate techniques which once would have taken months by desk calculators. Influential encouragement of this approach ("seek and ye shall find") has come from a number of investigators. Tied to a great deal of data collection in social programs, marketing, attitude measurement, etc. is the field of sample survey design. This hearty field of applied statistics is ever increasing in usage but does not receive much formal attention at the large graduate statistical centers.

In much of this work, it is difficult to include much mathematical structure. For example, after much manipulation the final significance level of a conclusion would be impossible to know. The basic idea is to explore, sort out data with the aid of computers, allow the model to vary to see what happens if it does, and at the end allow the investigator to come to common sense conclusions about what the data says, based on all the evidence he then has. In spirit, this returns us to the earliest days of statistics, but with much more powerful tools. We can expect enormous developments along these lines. There should also be some attempt to put structure behind the procedures, to decide the error probabilities of decisions and the consequences they could have, and so forth. Here the modern interest in robustness is important. Practical men will not worry about schools of inference if they can be satisfied that their basic techniques will lead to correct decisions on the whole, despite loose specifications in the model they use. There is much work to be done on these lines, and in model building itself. Here, too, the computer will be pervasive, especially in handling mathematical intractability through Monte Carlo methods.

Simultaneous Parameter Estimation

In the last few years Bradley Efron and Carl Morris and other authors have developed and applied a method of estimation due to Charles Stein. The method represents a significant advance in the theory and practice of simultaneous estimation of several parameters, a situation that occurs often in present day data analysis. Briefly, this procedure suggests that estimation of parameters from each of three or more categories or populations can profit by using the sample data from all categories rather than employing sample data from the cth category to estimate parameters

in the cth category. Some immediate applications are risk categories for insurance and population proportions in strata in sample survey situations.

To motivate and describe the approach, and to extract a philosophy of estimation which will hopefully be applicable to new situations and models, we will discuss a specific case in some detail. This case is due to Stein. Some technical language will have to be employed to maintain the flavor of the approach.

Suppose $X_1, \ldots, X_K$ are independent normally distributed random variables with unknown means $\Theta_1, \ldots, \Theta_K$ and common variance 1. We wish to estimate the vector $\underline{\Theta} = (\Theta_1, \ldots, \Theta_K)$ where the estimation error is governed by total square error loss. The maximum likelihood estimator, which is also the best unbiased estimator, is $\underline{X} = (X_1, \ldots, X_K)$ itself. For $K = 1$ and 2, $\underline{X}$ is an admissible estimator. However, for $K \geq 3$ the James-Stein estimator $\hat{\underline{\Theta}} = [1 - (K-2)/(\Sigma_1^K X_i^2)]\underline{X}$ dominates $\underline{X}$. Note that an estimate for the ith category, $\hat{\Theta}_i$ for X_i, employs all the sample information through $\Sigma_1^K X_i^2$. In fact,

$$E_{\underline{\theta}}\|\hat{\underline{\theta}} - \underline{\theta}\|^2 = E_{\underline{\theta}}\|\underline{X} - \underline{\theta}\|^2 - E_{\underline{\theta}}\left(\frac{(K-2)^2}{\sum_1^K X_i^2}\right)$$

$$= K - E_{\underline{\theta}}\left(\frac{(K-2)^2}{\sum_1^K X_i^2}\right).$$

The quantity $E_{\underline{\theta}}((K-2)^2/(\Sigma_1^K X_i^2))$ represents the savings in risk gained by using $\hat{\underline{\Theta}}$ rather than $\underline{X}$. This savings attains a maximum value of $K-2$ for $\underline{\Theta} = 0$, then decreases to zero as $||\underline{\Theta}||$ increases, but always remains positive. This result is very surprising. The X_i's are independent and no relationship among the Θ_i's is assumed. It does not at first seem plausible that any observation other than X_i should be used to estimate Θ_i. Further thought, however, does reveal the plausibility of the James-Stein estimator. Below are three perspectives which motivate the estimator:

(i) Estimate the best weighted average of $\underline{0}$ and $\underline{X}$

Suppose we guess a priori that $\underline{\Theta} = \underline{0}$. We decide that our estimator should be a weighted average of $\underline{0}$, our prior guess, and $\underline{X}$, the maximum likelihood estimator. We consider estimates of the form $\lambda \cdot \underline{0} + (1-\lambda)\underline{X}$. Now if $\underline{\Theta}$ is the true parameter value the risk of this estimator is $\lambda^2||\underline{\Theta}||^2 + (1-\lambda)^2 K$. The risk is minimized at $\lambda = K/(K+||\underline{\Theta}||^2)$. The proportional savings in risk in using the optimal λ rather than $\underline{X}$ is $K/(K+||\underline{\Theta}||^2)$. The proportion of risk saved is 1 at $\underline{\Theta} = 0$, decreases to zero as $||\underline{\Theta}|| \to \infty$, and is always positive. The above estimator cannot be used because $||\underline{\Theta}||^2$ is unknown. However, $||\underline{\Theta}||^2$ can be estimated from the data. The problem of estimating the scalar parameter $||\underline{\Theta}||^2$ is considerably less difficult than that of estimating the vector parameter $\underline{\Theta}$. Intuitively we feel that if the optimal λ is small (and thus the savings in risk over $\underline{X}$ is large) we should be able to detect it from the data, to fairly accurately estimate the best weighted average of $\underline{0}$ and $\underline{X}$, and thus to secure a good share of the savings in risk. Now $E_{\underline{\Theta}}(\Sigma_1^K X_i^2) = K + ||\underline{\Theta}||^2$. This suggests estimating the optimal λ by $K/(\Sigma_1^K X_i^2)$. For technical reasons the estimate $(K-2)/(\Sigma_1^K X_i^2)$ is preferable, and this leads to the James-Stein estimate

$$\frac{K-2}{\sum_1^K X_i^2}\,\underline{0} + \left(1 - \frac{K-2}{\sum_1^K X_i^2}\right)\underline{X} = \hat{\underline{\theta}}.$$

The savings in risk is, of course, smaller for the estimated optimal weighted average than for the actual optimal weighted average, but not a great deal smaller. Note that when the data supports the guess $\underline{\Theta} = 0$, our estimated λ is large, and we thus give substantial weight to the guess. If the data tells us that $\underline{\Theta} = 0$ is an obviously bad guess, then the estimated λ is small and we essentially ignore the guess. Thus we capitalize on a successful prior guess but pay no penalty for a bad guess.

(ii) *Preliminary test estimation*

Suppose prior to estimation we perform a test of an hypothesis. We test $\underline{\Theta} = 0$ vs. $\underline{\Theta} \neq 0$. Our test rejects $\underline{\Theta} = 0$ if $\Sigma_1^K X_i^2$ is large, and accepts $\underline{\Theta} = 0$ is

$\Sigma_1^K X_i^2$ is small. If we reject $\underline{\Theta} = 0$ then we use the estimator $\underline{X}$; if we accept we use the estimator $\underline{0}$. This type of approach is known as preliminary test estimation. One way of viewing this procedure is that based on the data we choose λ either equal to zero or 1, then use the estimator $\lambda \cdot \underline{0} + (1-\lambda)\underline{X}$. Rather than allowing only zero or one, it makes better sense to have λ assume all values between 0 and 1, depending on the credibility of the hypothesis $\underline{\Theta} = 0$. This suggests having λ be monotone decreasing in the test statistic $\Sigma_1^K X_i^2$. The value of λ, $(K-2)/(\Sigma_1^K X_i^2)$, in the James-Stein estimator has this property.

(iii) *Empirical Bayes*

Suppose we have a prior distribution on $\underline{\Theta}$ under which the Θ_i's are independent, normally distributed with mean zero and variance A. The Bayes estimator of Θ is given by $\underline{\Theta}_B = A/(A+1)\,\underline{X} = (1 - [1/(A+1)])\underline{X}$. Suppose now that A is unknown. We follow the approach of estimating the Bayes estimator from the data. This is known as empirical Bayes estimation. The statistic $\Sigma_1^K X_i^2$ is distributed as $(A+1)$ times a chi-square statistic with K degrees of freedom. It follows that $(K-2)/(\Sigma_1^K X_i^2)$ is an unbiased estimator of $1/(A+1)$, suggesting the estimator $[1 - (K-2)/(\Sigma_1^K X_i^2)]\underline{X}$ as the estimated Bayes or empirical Bayes estimator. But this is precisely the James-Stein estimator.

At first one may suspect that the above example is uniquely contrived to yield a mathematical curiosity, with no great relevance to statistics. Perhaps there is something unique about the loss function, or the normal distribution, the equality of variances, or the special role played by the $\underline{0}$ vector. Perhaps the savings in risk will be negligible in most applications. Time and a lot of good work by many people have shown that the above suspicions are basically groundless. The phenomena illustrated above holds for very general loss functions, for normal distributions with very general covariance matrices, and for several non-normal families of distributions. The special role played by the $\underline{0}$ vector can be replaced by an arbitrary prior model of parameter structure which places the mean in a lower dimensional subspace. The savings in risk can be substantial when the parameter structure which we hypothesize turns out to be reasonable. For example, in the case we considered above, suppose we feel a priori that it might be reasonable that the Θ_i's are approximately equal. If the Θ_i's were equal we would estimate $\underline{\Theta}$ by the vector $\overline{\underline{X}} = (1/K\,\Sigma_1^K X_i, 1/K\,\Sigma_1^K X_i, \ldots, 1/K\,\Sigma_1^K X_i)$. In the absense of known relationships between parameters we might try the maximum likelihood estimator $\underline{X} = (X_1, \ldots, X_K)$. Motivated by the arguments given above we now would combine $\overline{\underline{X}}$ and $\underline{X}$ by

$$\left(1 - \frac{K-3}{\sum_1^K (X_i - \bar{X})^2}\right)\overline{\underline{X}} + \left(\frac{K-3}{\sum_1^K (X_i - \bar{X})^2}\right)\underline{X}.$$

The savings in risk would be given by

$$E_{\underline{\theta}}\left(\frac{(K-3)^2}{\sum_1^K (X_i - \bar{X})^2}\right),$$

which exceeds

$$\frac{\frac{(K-3)^2}{K-1}}{1 + \left(\frac{1}{K-1}\sum_1^K (\theta_i - \bar{\theta})^2\right)}.$$

If the true $\underline{\Theta}$ has approximately equal components so that $(1/[K-1])\Sigma_1^K(\Theta - \bar{\Theta})^2$ is small, then the savings will be substantial, especially for large K.

When faced with a multiple parameter estimation problem the statistician should be aware that the Stein approach may be helpful. It should be part of his arsenal, along with the more standard estimation approaches. In carrying out the estimation, he should think about the parameters and decide what sort of relationship between parameters might be reasonable. The resulting Stein estimator (which has not yet become automatic to construct) will save significantly on risk if the data supports the relationship, and will essentially ignore the hypothesized relationship if the data firmly rejects it.

Recent Results. Efron and Morris have constructed Stein type estimators which limit the amount of shift that any one component can undergo. This is highly desirable in practical situations. They show that most of the savings in risk in the James-Stein estimator is salvaged. Stein has considered an alternative approach to prevent extreme shifting. They have also extended the James-Stein estimator to the case where each observation is vector valued and have contributed to the important problem of whether to combine possibly related estimation problems or to treat them separately.

Stein has constructed a rich class of estimators which dominate the maximum likelihood estimator in the normal case with independent components, equal variances, and square error loss, and Stein, Joshi, and Faith, in separate papers, have considered the problem of constructing confidence sets from Stein estimators. The theory has not yet been completely developed. Efron, Morris, and Stein have considered the problem of estimation of a covariance matrix in the normal case. Their improved estimate of the covariance matrix leads to improved estimators of the means in the case of independent normal random vectors with unknown means and common unknown covariance matrix. This is of great importance in a number of applications.

Clevenson and Zidek, and Pong, in separate papers, have studied the case of estimation of several Poisson parameters. They construct estimators which dominate the maximum likelihood estimator under two common loss functions. Hudson has studied the case of one parameter exponential families and has extended some of the normal theory results. Fienberg and Holland have constructed Stein type estimators for the multinomial case. Here the usual estimator is admissible because it does well at extreme points, but the authors show that their estimator has lower risk for most of the parameter space. In a sense that they make precise, their estimator asymptotically dominates the usual estimator.

Brown has demonstrated the inadmissibility of the maximum likelihood estimator in the normal case under an extremely wide class of loss functions. Strawderman has constructed admissible estimators which dominate the maximum likelihood estimator in the normal case with equal variances, independent observations, and square error loss. Berger has extended the theory to normal distributions with arbitrary positive definite covariance matrixes and arbitrary positive definite quadratic loss. He has also obtained results for normal distributions with random scale parameter. Fienberg and Holland, Peng, and Hudson, in separate papers, have applied Stein estimators to contingency table estimation.

Future Directions. In many situations, including contingency table analysis and analysis of variance, one is faced with several possibilities of relationships among the parameters. A priori it is hard to say which relationships are reasonable. If we follow the basic approach we have been discussing, we would have to pick out one and only one such relationship. We would then benefit if the relationship was confirmed by the data, but would gain little, if anything, if the relationship was rejected. What we need is an estimator sensitive to several different hypothesized relationships which will capitalize on those which are supported by the data. The current practice in contingency table analysis and analysis of variance is to take a nested set of models and perform a series of hypothesis tests to find the acceptable model with fewest parameters. Once the model is chosen, the maximum likelihood estimator is used. The estimator is thus a preliminary test estimator. Extrapolating from our knowledge of Stein estimators, it would seem preferable to take a weighted average of estimates from the different models. The theory of such an approach has to be worked out.

Stein estimators will undoubtedly have many applications in reliability. For example, it should be possible to simultaneously estimate the failure rates of several different components, obtaining a reduction in the total mean square error over the estimator which treats each component separately. In sample surveys we often try to simultaneously estimate several probabilities. It is clear that a Stein type approach will lead to estimators considerably more accurate than the raw frequency estimates. A natural problem for the Stein approach is the estimation of high order transition probabilities in discrete stationary systems. The problem is very much related to that of contingency table analysis. The problem of obtaining confidence sets based on Stein estimators

is quite important. We generally do not want a point estimate alone, but also a confidence set for the parameters. This is still somewhat intractable.

In estimating a cumulative distribution function (cdf) from a sample we might a priori expect that the underlying distribution is normal, exponential, or belongs to some other parametric family of distributions. If we have a particular parametric family in mind, we might estimate the cdf by estimating the unknown parameters and then substituting into the parametric form for the cdf of the family. On the other hand, if we were not prepared to assume a parametric family, we would estimate the cdf by the sample distribution function. The Stein approach suggests taking a weighted average of the two above estimators, basing the weight given to the parametric estimator on a goodness of fit statistic which tests whether the data supports the parametric model. Such an estimator should give a good practical improvement over either of the two separate estimators, or over a preliminary test estimator.

Goodness-of-Fit Testing

The Chi-Square Test. For many years the only well-known goodness-of-fit statistic was the χ^2 (chi-square) test introduced by Karl Pearson. The test was naturally attuned to testing for a discrete distribution. If a continuous distribution were to be tested, the distribution had to be divided into cells, and the probabilities of falling in the cells calculated; the numbers of observations in the different cells were then counted and treated as though they came from a discrete distribution. A common problem is to test that a sample comes from a given distributional form, with, however, one parameter or more unknown; e.g., to test that the observations are Poisson with unknown λ, or normal and unknown μ and σ^2. A great advantage of the χ^2 test is that it can readily be adapted for testing with such unknown parameters.

As with so many of the earlier procedures, precise theory was given only much later, and it is worth remarking that the standard techniques of estimation usually followed are not the correct ones to give the (asymptotic) distribution usually used; namely, a χ^2 distribution with a reduction of degrees of freedom equal to the number of parameters estimated. However, the test is easily understood and quickly became an important tool in applied statistics. Among the many important problems connected with the test, much research was done, instigated notably by Mann and Wald some 30 years ago, on choosing the best way to divide a continuous distribution to maximize the statistical power of the test.

EDF Statistics. Another approach to goodness-of-fit testing for a continuous distribution was taken by Kolmogorov in the early 1930's and later by other authors. This was to draw the empirical distribution function (EDF) of the data (i.e., plot $F_n(x)$, the number of observations less than or equal to x), and then to base a test on some measure of the discrepancy between $F_n(x)$ and the hypothesized distribution $F(x)$. Kolmogorov chose D, the supremum of the absolute difference, as x varied over its range. When $F(x)$ is known completely, i.e., no unknown parameters are present, Kolmogorov's test statistic has a distribution which does not depend on $F(x)$, i.e., on the distribution tested; such a test statistic is called distribution-free. Harald Cramér and Richard von Mises later suggested another test statistic, W^2, based on $\{F_n(x) - F(x)\}^2$, integrated over the range with a weight factor to give a distribution-free test; subsequent authors have incorporated other weight factors to give prominence to the tails, for example the statistic A^2 proposed by Anderson and Darling, or have adapted better types of statistics for use on a circle, the statistics V and U^2 by Kuiper and Watson.

An enormous literature has developed on these statistics, particularly the Kolmogorov D, discussing methods of obtaining the small-sample distributions, possible variations of the statistics, for example, to give one-sided tests, and giving power comparisons with χ^2. It is hard to make comparisons because of the broad nature of possible alternatives to $F(x)$, but in a general way it seems to be clear that D, and even more so, W^2 and its variants, are more powerful than χ^2 over a wide range of alternatives for this case where $F(x)$ is continuous and completely specified. This is to be expected since there is a loss of information where the measured observations are grouped into cells for the χ^2 test.

Presence of Unknown Parameters. Until recently, the EDF statistics have not been able to be

used if unknown parameters were present in the distribution tested. However, it was known that, if the parameters were location and scale, then the null distributions would depend on the type of distribution tested, but not on the true values of the parameters. Further, it had been shown how at least the asymptotic distributions of the Cramér-von Mises family (W^2, A^2 and U^2) could in principle be found for these situations. In recent years further work has been done to provide significance points for certain important families of distributions, particularly the normal, exponential, and Gamma. Asymptotic theory has been developed by Michael Stephens and other authors, and for finite sample size several authors have provided significance points.

Closely related work has been done by Durbin and Knott and Stephens. For F(x) specified they expand $\sqrt{n}\,(F_n(x) - F(x))$ as a Fourier series and base tests on the coefficients of the terms. The statistics W^2, U^2 and A^2 can be expressed in terms of these, and in some cases the early coefficients will be more powerful than the entire statistic. Asymptotic distributions of W^2, U^2 and A^2 can also be found this way, the asymptotic power studies can be provided. This is a valuable addition to the Monte Carlo studies for finite sample size on which judgments must usually be based. Durbin and collaborators have extended this work to the case where F(x) is the normal or exponential distribution, with parameters unknown.

Regression Methods. Another important approach to goodness-of-fit was introduced by Shapiro and Wilk about ten years ago. The technique is useful when unknown parameters and those for location and scale. Suppose F(y) is the parent population with standard variate y; i.e., in general, the distribution to be tested is $F((x-\alpha)/\beta)$ where α and β are the unknown location and scale parameters. Suppose then a sample of size n is taken from the distributions F(y), i.e., $\alpha = 0, \beta = 1$, arranged in ascending order, and let m_i be the expected value of the ith order statistic. The ith order statistic is the value of the observation with rank i when the observations of the sample of size n are placed in ascending order. For example, the first order statistic is the smallest value, and the nth order statistic is the largest value.

For a sample of values from the more general population, let x_i be the ith order statistic; then we have

$$\text{(i)} \qquad E(x_i) = \alpha + \beta m_i.$$

If the hypothesized distribution is correct, a plot of x_i against the known m_i should produce a straight line, similar to that in simple regression. The Shapiro-Wilk method consists of estimating the parameters α and β be generalized least squares (the x_i's are correlated) and then devising a test statistic which in some way compares these with other estimates, say those given by maximum likelihood. Thus in the case where the distribution tested is normal, α is μ and β is σ; if the least squares estimate of β (i.e. σ) is $\hat{\sigma}$, and the usual estimate is s, the Shapiro-Wilk statistic W is a multiple of $\hat{\sigma}^2/s^2$.

A disadvantage of the statistic W is that little is known about the null distribution, even asymptotically, so that all significance points are based on Monte Carlo results; also W is calculated from a linear combination $\Sigma a_i x_i$, and the coefficients a_i differ for every n. These coefficients were provided by the authors for n up to 20, and approximate values for n up to 50. Beyond n = 50, a modification of the statistic has been suggested. For tests of normality W seems to be slightly superior to the best EDF statistics, though these are easier to calculate. Attempts to extend the basic technique to other distributions, e.g. the exponential, suggest that the superiority over other statistics is not so marked.

Tests for Special Distributions. Among much older statistics which have been advocated for tests of normality are b_1 and b_2; b_1^2 is m_3^2/m_2^3, and $b_2 = m_4/m_2^2$, where m_j is the jth sample moment about the sample mean; b_1 takes the same sign as m_3. These are sample equivalents of the population parameters β_1 and β_2 devised to measure skewness and kurtosis respectively. Over many years tests for normality were suggested, based on using b_1 and b_2 to test that $\beta_1 = 0$ and $\beta_2 = 3$. A major difficulty was that the exact distributions of b_1 and b_2 are intractable. Many attempts at approximation were made, and finally some very extensive Monte Carlo tables have recently been produced. An interesting recent development has been introduced in which b_1 and b_2 are recorded

on a chart, using usual rectangular axes, and contours are given beyond which the hypothesis of normality will be rejected. Thus in effect two statistics are being used to make the test.

Tests for the Exponential Distribution. After the normal distribution, the exponential distribution receives most attention; in some important applied fields, for example, reliability occupies the central position. The exponential distribution is closely associated, in various ways, with the uniform distribution, and tests for exponentiality can sometimes be turned into tests for uniformity. Further, test statistics with certain optimal properties can be devised against specific alternatives, such as the Weibull and Gamma distributions. All these distributions are important in reliability models. Thus again many statistics have been proposed to test for exponentiality and in some cases they are complicated and distribution theory is difficult. Much work needs to be done to sort out the merits of the different procedures.

The Future. This is an appropriate point to consider the future in goodness-of-fit work. In what follows, we make a number of connected points.

(a) In the past, classical goodness-of-fit statistics (those introduced, let us say, before the arrival of electronic computers) inspired an enormous literature because they posed interesting mathematical problems; though frequently the papers written were not very useful to the practitioners in deciding which way to proceed. With this in mind, let us consider the probable needs of an applied statistician making a test of fit:

(b) First, the practicing scientist will surely not want to see his data transformed too much. Given a set of values, the histogram or the EDF gives him a good picture of his sample distribution, and he will not want to get too far away from this. The probability integral transformation, which takes his x-values and returns a new set of values which ought to appear uniformly distributed between 0 and 1, is probably as far as he would want to go in this line. This is loosely equivalent to the use of probability plotting paper. Similarly he or she will be happy with the graphical approach implicit in the regression tests discussed above.

(c) Even with the present availability of computers, there is considerable advantage to easily calculated techniques which give good power. The basic EDF and regression methods are in this category and they will probably gain ground on the chi-square test, which is usually inferior in terms of power.

(d) In the light of the above general comments, work needs to be done on EDF statistics in providing points for other distributions with location and scale parameters, and in examining what can be done when parameters are not of this type. As to the regression techniques, it would certainly be desirable, for rounding off the mathematical aspects, if more distribution theory could be provided for the Shapiro-Wilk technique and other techniques. It would be valuable, for all these statistics (EDF and regression) to have calculations of relative efficiency and asymptotic power.

(e) From the mathematical point of view, some of the newer statistics leave much to be desired. Distribution theory is often lacking. Asymptotic theory, percentage points for testing, and the power studies to support their use, must be provided by Monte Carlo methods. There often seems to be no coherent philosophy or basic principle behind the *ad hoc* introduction of these tests. The appeal for statistics for which some mathematical results can be supplied is made not just for the sake of elegance. Nearly always such statistics can be examined more critically, and measures of efficiency can be found which give an overall picture of where the statistic fits in with its rivals. Statistics which extend one of the basic techniques will probably be of greater interest than those which are simply slight modifications of older statistics. Such extensions show signs of giving good overall power and research will have to be done on their distributions, power properties, and relative utility.

(f) A more constructive use of the computer is to make good use of several test statistics to decide the goodness-of-fit of a sample. Here we can hope to exploit the fact that the machine will calculate any number of statistics, and it seems correct not to try to reduce a test for a distribution to the calculation of only one number. In the case of tests for normality, the values of b_1 and b_2 are "interpretable" statistics, in terms of concepts (skewness and peakedness) readily accessible to the applied worker. Here the computer makes it possible to calculate these otherwise difficult statistics, and the use of both together, especially

graphical use, should give a new lease of life to these older statistics.

(g) The general problem of how various test results can all be exploited to give an accurate picture of the parent distribution is greatly in the spirit of today's interest in data analysis. Certainly if we know how test statistics will behave when the parent population is not the one which is tested, we can look at several such statistics to indicate the nature of the departure, if any, from the proposes parent population. But if one wants to use these several values to make an overall test procedure, there is much work to be done. The background question is the general one of how to combine several test statistics, a subject which has had a long history. Often the statistics to be combined are independent, perhaps from a number of independent samples. In the present instance, statistics for one sample will not be independent, often not even asymptotically. But the correlation will not be known and Monte Carlo work will be necessary (as has been done in the case of normality) to evaluate test combinations. It is to be hoped that the great deal of work involved will be expended only on procedures which will have a real practical appeal.

(h) There is also a wide open field in the provision of appropriate tests for multivariate distributions. There are no extensions yet devised for EDF statistics in two dimensions, for example; or methods of procedure to test if a distribution is multivariate normal. Here again we run into the problem of how much the data should be condensed before a decision is made. Sequential goodness-of-fit testing has also received little attention, and it would seem that this area would have considerable potential if developed. It must be useful to decide, as observations come in, whether these appear, say, normal, to decide how best to analyze them next.

Data Analysis

It may look out of place to initiate a section labeled "data analysis" when in effect the previous sections have discussed this in the context of simultaneous parameter estimation and goodness of fit. There is a wide variety of techniques that are more centrally fundamental in that without or with limited assumptions they try to achieve a parsimonious view of the data on hand. We list and discuss these data dependent techniques which are already receiving much attention and will continue in a more intensive way in the foreseeable future. Quite often the data is multidimensional, the structure is not known, and the data analyst wishes to make some sense out of it for the investigator. Another type of observation is one where direction as well as the magnitude of the observation is central and the direction can be viewed in two, three or even higher dimensions. This will be reviewed at the end of this section.

For the multi-dimensional non-directional variety there is now a grab-bag of techniques such as classification, discrimination, clustering, scaling, multi-dimensional contingency table analysis, and, where the data warrants it, seriation procedures. The latter arises when arahaeological, epigraphical, and intelligence data is to be analyzed and we will return to this subsequently. The other techniques mentioned will now receive attention.

Classification and Clustering. Data analysis has undergone a resurgence in the last two decades. In the main, this is due to the advent and development of the electronic computer and its extraordinary capacity to ingest data and spew out its product in accordance with instructions supplied by the appropriate algorithm. The eager and voluminous collection of data in the nineteenth century, especially by the British school of scholars, was denied the additional analysis it merited by the lack of a computer technology. In a very specific and substantive way, the desire to do data analysis and some of the frustrations encountered led to mathematical modeling and the modern school of statistics.

Scientists and scholars have long been concerned with "sorting things into groups" and numerical taxonomy either does this directly or serves to guide those who make such decisions. Under numerical taxonomy, we can list two categories: i) clustering of data, ii) classification of data. The latter can be viewed as a subset of the former. In the former category, we require the data to produce both the number of groupings or clusters and the assignment of each element or individual to these groupings. In the latter category, the number of groups or clusters is predeter-

mined, each group is labeled, and rules are desired on the basis of which an assignment of each element is made to one of the fixed groups. Classification procedures may also be termed assignment procedures.

It is not prudent to convey a sharp distinction between clustering and classification in an operational sense. If a classification procedure is not producing meaningful groups through the assignments that are made, then changes are called for—namely, revising the pre-determined groupings either in number or in shape, or both, on the basis of the new information. This sequential revision of groups on the basis of the data available at any one time suggests that one is indirectly engaging in clustering procedures. On the other hand, it is wise to keep in mind these conceptual differences when attempts at clustering and attempts at classification are made.

Data Summarization and Representations. There are several ways to begin the data summarization. All give a picture of data interrelationship but each has special reasons for its employment by an investigator. One representation is that of the scatter matrix. Here we portray the total scatter or dispersion displayed by n individuals or elements each measured on p variables (n points in a p-dimensional space).

If each element in the scatter matrix T is divided by n, the resulting matrix is the covariance matrix. Now if we also divide each element by the appropriate standard deviations, the resulting element is the correlation coefficient and the matrix is now labeled the correlation matrix.

An important advantage of T is the manner in which it can be decomposed into two matrices that are especially pertinent in clustering and classification studies. In a classification study, the n elements will be assigned to k pre-determined groups. Each group with, say, n_i elements, can be viewed as a universe with its own scatter matrix formed as before and labeled W_i. If we sum all the W_i scatter matrices, we get

$$W = \sum_{i=1}^{k} W_i$$

and let this represent the within scatter or homogeneity of the groupings. Likewise, if for each of the k groups, we compute the group mean, we can obtain a (p × p) matrix that we label B, for it expresses a measure of the "betweenness" or heterogeneity of the k groups. The central point in this development is the existence of the fundamental matrix equation

$$T = W + B.$$

This result suggests immediately an index by which classification (pre-determined number of groups) can be evaluated and by extension how clustering can be terminated at some cluster size. For any given data set T is fixed. Thus measures of "groupiness" or "clusteriness" as functions of W and B are thrust forth for examination.

For p = 1, the matrix equation reduces to an equation about scalars. Thus a good grouping index is one which minimizes W or equivalently maximizes B. We may also consider maximizing the ratio B/W or T/W = 1 + B/W. An added benefit is that this ratio is invariant under linear transformations of the data. Statisticians have long exploited this fact for B/W multiplied by an appropriate constant is the familiar F ratio in the analysis of variance.

When the number of measurements per element is two or more (p > 1), grouping criteria are not so straightforward. Several possibilities suggest themselves and have been developed and studied by investigators. One criterion suggested by several authors that is a quite natural index is the minimization of the trace of W (sum of all elements in the main diagonal of the matrix) over all possible partitions into k groups. This is equivalent to maximizing trace B because

$$\text{Trace } T = \text{Trace } W + \text{Trace } B.$$

However Trace W is invariant only under an orthogonal transformation and not under nonsingular linear transformations. The trace is the sum of all the elements in the main diagonal of the matrix.

Another criterion that may be employed for p > 1 is the ratio of the determinants

$$\frac{|T|}{|W|} = |I + W^{-1}B|.$$

We can use $|T|/|W|$ as a criterion for grouping and select that grouping for which this index is

maximized, or equivalently $|W|$ is minimized. Also we may employ $\log(|T|/|W|)$ since it is a monotonic function.

Another criterion for grouping is the trace of $W^{-1}B$ and we select that grouping that maximizes this index. This index has been used as a test statistic in multivariate statistical analysis as has the ratio $|W|/|T|$. The latter was employed by S. S. Wilks to test whether groups differ in mean values.

Both Trace $(W^{-1}B)$ and $|T|/|W|$ may be expressed in terms of the eigenvalues, λ_i, of the matrix $W^{-1}B$. We write

$$|T|/|W| = \prod_{i=1}^{p} (1 + \lambda_i)$$

and

$$\text{Trace } W^{-1}B = \sum_{i=1}^{p} \lambda_i$$

where λ_i are the roots of the determinantal equation, $|B - \lambda W| = 0$. The characterization of these ratios in terms of eigenvalues is helpful in data representation, especially when the effects of some reduction in dimensionality is desired. All the eigenvalues of this equation are invariant under non-singular linear transformations of the data. It can be proved that these eigenvalues are the only invariants of W and B under non-singular linear transformations.

Distance Matrix, Thus far we have discussed some summarization of multivariate data in matrix form, either T (scatter), covariance, and correlation and the kinds of grouping criteria that are suggested by the T format. Intuitively, we see that any grouping criterion is a function of homogeneity within groups and heterogeneity between groups and the indexes already described are specific quantities embodying these notions. For the correlation coefficient index, large values indicate homogeneity; small values indicate heterogeneity.

Another method of summarizing data that is more appropriate on many occasions is to find the distance between each pair of the n points in p-dimensional space. This leads to a representation in matrix form of an $n \times n$ matrix where each element, in the ith row and the jth column. say d_{ij}, is the distance in the p-dimensional space between the ith element or individual and the jth element or individual. All the elements in the main diagonal are zero. The distance matrix is akin to the correlation matrix in that both may be viewed as similarity matrices—the jumping off place for clustering and classification attempts.

The decision as to whether correlation matrices or distance matrices are to be employed is usually determined by the problem at hand. If n individuals or n elements are to be grouped on the basis of p measurements on each, then the $n \times n$ distance matrix is the natural summarization; if the p measurement variables are to be grouped on the basis of the measurements on n individuals or n elements, then the $p \times p$ correlation matrix is the natural summarization of the data. This latter matrix is the natural point in factor analysis where parsimony in the number of latent measurement variables is the desired goal.

However, we are now at a juncture where a large number of clustering techniques have been developed and promulgated. The major activity along these lines has taken place in the last 15 years or so. In fact, we are now at the pretentious stage of thinking about the clustering of "clustering techniques." First a word about some specific clustering techniques. Here are some of the more popular varieties with brief comments about each.

1) Q-Factor Analysis: Factor analysis of elements rather than variables, number of clusters defined by factors and entry into cluster determined by highest factor loading.
2) Single Linkage (Nearest Neighbor): Groups initially consisting of single individuals are fused according to the distance between their nearest neighbors, groups with smallest distance being fused. Each fusion decreases by one the number of groups. Distance between groups is defined as distance between their closest members. This leads to "serpentine" or "chained" clusters.
3) Complete Linkage (Furthest Neighbor): Distance between groups is now defined as distance between their most remote pair of individuals. Distance between merging clusters is the diameter of the smallest sphere which can enclose them. It yields tight, hyperspherical clusters that join others only

with difficulty. Each fusion decreases by one the number of groups.

4) Average Linkage (King's Method): Distance between groups is judged by their centroids and closest centroids are fused. Each fusion decreases by one the number of groups.
5) k-means: Start with k-clusters (e.g., first k points); use minimum intracluster distance around the mean as criterion. As an element enters the cluster, the mean is updated and this continues until all points are placed.
6) ISODATA: Start with k-clusters and assign all elements by intracluster minimum criterion. After all elements have been assigned, update the means and do again until no gain occurs in intracluster minimum criterion.
7) Covariance Criterion Optimization: Place points in k-clusters and reassign according to a variance-covariance criterion; that is, maximize the determinantal ratio $|T|/|W|$.

Note that clustering techniques (2), (3), (4) are hierarchical grouping procedures, i.e., one begins *with n clusters each containing one element* and each fusion reduces the number of clusters by one until only one cluster containing all points is achieved. In the k-means technique and the ISODATA technique, it appears that one deals only with k-clusters. This is not so, because the k-cluster configuration can be reduced or enlarged in number of clusters if the intracluster distances are either too small or too large for the finally selected k-clusters. This characteristic is true also for the covariance criterion optimization procedure.

There are some data representation and graphical techniques that are sometimes mistaken as clustering procedures. These techniques make visual clustering feasible by reducing the p-dimensional space to two dimensions or changing the multidimensional vector to a human face or some other analogue representation. The latter represents an interesting device developed by Chernoff, who translates the multidimensional data vector into a face and then judges are assigned to group the faces. Judges are also involved in Kruskal's multidimensional scaling technique, which reduces the dimensionality of the p-dimensional space. Regular factor analysis can also be employed to reduce the p-dimensional space to two dimensions when measurement variables rather than elements are being clustered. Each variable is then a point in two dimensions, can be plotted, and the n points grouped by eye.

Since a number of clustering techniques are available, some evaluation of these techniques becomes a necessity. In order to accomplish this, some evaluation indexes are required. They are: a measure of external criterion validity, a measure of internal criterion validity, and a measure of replicability. External criterion validity is obtained by computing the percentage of concordance of expert assessments and the results of the clustering procedure. This can be accomplished by the use of a contingency table. Briefly, in this situation, the expert or the consumer decides how the actual clustering developed by the clustering procedure relates to the substance of the problem that produced the data in the first place.

The measure of internal criterion validity is the cophenetic correlation coefficient introduced by Sokal and Rohlf. This is the ordinary product moment correlation coefficient between corresponding cell entries of the similarity matrix derived from the cluster configuration and the initial similarity matrix employed to initiate the clustering. The measure of replicability or stability is also essentially a correlation coefficient. In this situation the data base is divided at random into two equal data sets. For each of the two data sets, a clustering configuration is derived by the clustering procedure in question. From each of the two clustering configurations, a derived similarity matrix can be constructed. Accordingly, a correlation coefficient can then be computed over corresponding cells in each of the two similarity matrices, each of which is developed from the clustering configuration produced by the clustering technique. In this way, each of the clustering techniques previously described briefly can be evaluated; that is, in terms of each of the three measures: external criterion validity, internal criterion validity, and replicability or stability. The future will see much effort in producing and evaluating clustering techniques.

Seriation. As a broad definition, seriation consists of arranging a set of collected items so as to infer ordering in some dimension such as time or space. It is a frequently occurring problem in archaeology and probably in intelligence settings.

Often under the umbrella of "seriation" we find two additional terms, "sequencing" and "scaling". Sequencing denotes the attempt to order the collections nonmetrically, i.e., to rank them on a one dimensional scale. Scaling attempts to do more by assigning a numerical value to each collection so that not only is order achieved but also some quantitative measure of relative closeness is computed.

Problems of seriation arise in various fields of research. We have already mentioned archaeology and intelligence settings. An issue in political thought is the ordering of a group of individuals on a scale from 'Liberal' to Conservative' on the basis of their responses to political questionaires. An example of a psychological application is the attempt to order a group of children on an intelligence scale through IQ test scores.

By far the most common application of seriation methodology is the case where the dimension not directly observable is time. This takes us immediately into the realm of archaeology where the term seriation is most frequently used and where such methodology is employed for inferring relative chronology.

Archaeologists have been somewhat reluctant to employ formal mathematical techniques which often do no more than make explicit the implicit mathematical reasoning they already use, but in the area of seriation the gap has been rather successfully bridged. The archaeological literature contains numerous applications of seriation methodology to such diverse sets of objects as grave sites, sediment deposits, manuscripts, incriptions and statuary.

To identify studies in seriation we must enumerate the three stages by which a seriation is achieved and indicate the critical questions within each stage. The stages are

(i) establishing which of the attributes of the objects are to be used in attempting to order them,

(ii) formulating a notion of "closeness" or "distance" between pairs of objects,

(iii) accomplishing the seriation based on these "distances".

With respect to (i) our problem involves specifying characteristics of the objects which provide information on the relative positions of these objects on the scale of interest. Let us confine ourselves to an archaeological setting where we would be interested in chronological sequencing. The objects usually consist of collections of items. The key issue to be dealt with in this connection is what data pertaining to a given set of archaeological material will permit a reconstruction of its relative temporality. In this vein we note that the single occurrence of one artifact may be much more "significant" than the tenfold occurrence of another. We must examine both incidence and abundance. The result of stage (i) will ideally be an attribute vector with weights attached to each component which can be measured for each object. All the components, quantitative or qualitative, should be order-related.

We next turn to stage (ii). We would now need to develop comparisons between objects using their associated attribute vectors. Similarities are usually measured between pairs of objects. Numerous indices have been suggested. Some have been employed successfully but to date there is no widely accepted "similarity function." We may be able to say meaningfully that one pair of objects is more similar than another pair (including the possibility that one object repeats in the pairs) but questions of quantification of the similarity remain.

We are thus led directly to stage (iii) which is the most crucial and likely the most fertile with problems. Given similarities (or relative similarities) between all pairs of objects, how do we reconstruct a 'good' estimated serial order? An important point to note is that through similarities the best we can hope to do is obtain an estimated order up to reversibility. This is clear since an estimated order and its reverse have the objects in the same relative order and it is up to us to orient the direction of the underlying scale for each particular problem. Realistically this should present no difficulty for expertise ought to be able to distinguish the earliest from the most recent.

Solutions thus far fall into two categories—"quick and dirty" (usually hand-calculable procedures) and computer oriented search procedures. The latter approaches will typically establish some criterion for searching through the various permuted object orders and selecting the optimum one. The criterion would likely arise from modeling presumptions which may be artificial and insensitive to variations in the model. Moreover as the set of objects to be ordered in-

creases in size the number of permutations to be searched becomes astronomically large. Procedures involving restricted searching (local searches, random searches, etc.) are useful in polishing rough permutations found by other methods but do not obviate the above problem. Alternative computer approaches have been suggested. One method notes the relation between the seriation problem and the famous traveling salesman problem and searches for the linear order of the objects having minimum sum of "distances" between points (equivalent maximizing the sum of "similarities" between objects).

No unique computer procedure has emerged as the most effective. Visual efforts based on large scale graphs or on mechanical constructions have been suggested particularly when undertaken by researchers possessing considerable experience and insight in the particular field and with the data itself. Such attempts, by employing crucial subjective judgments on the part of the scientist, may prove more successful than the most sophisticated mathematical procedures. Ultimately the best solution may be a blend of both methodologies. Perhaps a mathematical approach might be employed to obtain a rough order and then expertise used to refine it.

Seriation, as a data analysis technique, has captured the imagination of just a few investigators. However, it is tied to important problems in scientific and military settings and will receive additional efforts in the near future.

The Analysis of Categorical or Count Data

An important class of categorical or count data is that of contingency tables. Further studies encompassing the practical side and the theoretical side are desirable.

On the practical side we mention the following. There are a number of computer programs in use which carry out various aspects of appropriate analyses. Since these have been developed over a number of years they do not represent an integrated set of programs. It would be useful to develop a second generation set integrating them in the sense of a common nomenclature, similar output and common standard statistical results. In addition, although the minimum discrimination information estimation approach is essentially dimension free, current algorithms impose limits on the size (number of cells) of data sets that can be analyzed because of the programming. By that is meant that minimum use of tapes and discs is made and the core memory is overtaxed. It would be desirable as part of the task of second generation programs to incorporate greater use of disc and tape memories, and thereby make available the possibility of analysis of data sets of many cells.

On the theoretical side the analysis of data sets with some variables nested within others seems to be an increasing problem and merits careful study. Since one is dealing with discrete data, the statistical analysis when the null hypothesis is not satisfied is rather difficult. This is an important problem to the experimenter, particularly when the observations are expensive to collect. What is the relation between the differences that can be detected and the number of observations? This is a question meriting further detailed examination. Asymptotic results are clearly not accurate for smaller numbers of observations. Corrections to the asymptotic distributions taking into account variations due to sample size would be useful. Such studies should be undertaken. Generally small sample properties need investigation.

Directional Data and Spatial Variation

There are many statistical problems in which it is natural or convenient to represent the data as points on the circumference of a circle or the surface of a sphere. An important case is when the data represents directions. A direction is represented by a unit vector in two or three dimensions, from the center 0 of a circle or sphere of radius 1 to a point P on its surface, and a sample of n directions is represented by vectors OP_i, $i = 1, \ldots, n$, or by the points P_i themselves.

In two dimensions the directions may represent directions of flights of migratory birds, or of prevailing winds, or geographical or geological data on the earth's surface. In three dimensions the vectors can denote direction of magnetization of rocks.

Another important application in two dimensions is when the circle is used to represent a time

period, e.g., a 24-hour clock, and the data are times of events (e.g., road accidents, or robberies) during the day.

In recent years much work has been done on analyzing such data. Pioneering papers are by Richard von Mises, R. A. Fisher, and Watson and Williams. An important distribution, introduced by von Mises for the circle and extended by Fisher to the sphere, has the probability density P per unit of surface area proportional to exp (k cos Θ), where Θ is the angle between OP and a given *modal vector* OA, and k is a positive constant. This distribution describes a unimodal distribution with mode at A, symmetric about OA, and with concentration around OA increasing with increasing k. A special case is when $k = 0$ and the density is uniform over the sphere or circle.

Historically the uniform distribution over the circumference of a circle or surface of a sphere has been examined in many varied applications, including the early theory of Brownian motion. The exact distribution of the resultant vector $\underline{R}$, for example, has been examined by many authors. The von Mises-Fisher distribution has been much used where the data has appeared to be clustered around a central mode. The distribution theory of useful descriptive statistics (such as the resultant $\underline{R}$ of a set of vectors $\underline{OP}_i$, or the component X of $\underline{R}$ on $\underline{OA}$, the modal vector) has been extended by M. A. Stephens and percentage points were given by him for statistical tests in a series of papers.

Another useful distribution is one where the density is clustered equally around opposite modes; several suggestions have been made, and the one which seems to lend itself best to statistical analysis is G. Watson's paper about ten years ago, in which the density per unit area is proportional to exp ($k \cos^2\Theta$). This is especially useful for *axial* data, i.e. data where the direction is known as a vector with its sense not important: so if OP is a vector, and POQ a diameter of the circle or sphere, either OP or OQ will equally well represent the data. Such data arises, for example, when planes are determined by their normals. Genuinely bimodal data, with opposite modes but of unequal strength, also arises in e.g. biology, where birds or animals sometimes are found to have a sense of home direction but some are unable to distinguish forward from backward. A distribution useful for this type of data, for both two or three directions, has been provided by M. A. Stephens.

A section on the von Mises and Fisher distributions has been included in Volume II of *Biometrika Tables for Statisticians*, and there are many tables to facilitate statistical analysis. Recently a book has appeared by Mardia which is entirely devoted to this field.

Vectorial Data. There will be other ways in which k-dimensional vectors can be used to record data, and for which they might be a useful tool in data analysis. The extension of the von Mises and Fisher distributions to higher dimensions would permit their use in much more general situations, where the vectors represent data not directly physical. The components could be related to proportions, for example, where a typical vector represents the composition of a chemical or geophysical material (e.g., rocks at sites where oil or minerals are being sought). The theory of this has been substantially worked out by M. A Stephens, but needs much implementation.

Spatial Variation. In a more general context, there has been a growth of interest in general problems of spatial variation—for example, variation over the earth's surface of (a) population density, or other demographical statistics, (b) positions of plants or trees, (c) incidence of certain diseases. The subject is often closely related to problems of clustering, since the presence or otherwise of clusters is often of interest to the investigator. There are also many applications in geology, especially with reference to mining: from the given drill-holes and the ore quality therein, one wishes to know where to site a mine for optimum returns. Techniques, fairly primitive but effective, were developed, specifically for mining, by Krige and co-workers in South Africa, and "Kriging" has entered the jargon. These techniques involve a spatial correlogram similar to the one dimensional correlogram used in time series analysis. In recent years a French school, headed by Matheron, has developed a more mathematical theory of spatial correlograms; though characteristically, almost every technical word that could be changed, has been changed, so that interconnections between this work and preceding work are sometimes hard to make precise.

The Future. There has been a great growth of interest in directional data in recent years. Most

of the work done so far has been on the unimodal von Mises and Fisher distributions, and it is clear that there is a need for more research on many problems in this area. Specifically:

(a) Work is needed on a more general distribution, to describe data which is not symmetrical. Bingham has discussed a useful distribution, on the sphere, but statistical tools are not sufficiently developed so far. It is also important, on both the circle and the sphere, to use a distribution with modes which are not opposite; for example, when using a circle to describe the 24-hour day, road accidents may occur with several peaks, not necessarily 12 hours apart. It will be possible to use superimposed von Mises distributions, but the analysis will be difficult to apply in practice.

(b) The general adaptation of "directional" results to data on the circle of the road accidents type, where the circle is used to represent a period and points on the circle will represent periodic data, should be an important technique for the future. It may well be found valuable as a data analysis tool in general time series analysis of periodic data.

(c) Some of the ideas which are very important in discussing linear observations need to be introduced to this area. For example, one needs to have a theory of correlation between sets of vectors, e.g. if vectors denoting magnetization directions are correlated before and after (say) heat treatment in a laboratory. Various definitions of correlation have been proposed, but much more needs to be done.

(d) Vectorial data in higher dimensions require additional analysis. There should be considerable potential applications of the idea expressed above of allowing proportions to be treated as components of unit vectors, so that an ore composition, say, is represented by a point on a k-sphere. The directional techniques can then be used to analyze these points. The procedure is easy to comprehend visually and this should give it some advantage as a tool of data analysis. Much implementation work should be done to see whether the results would compare with more traditional multivariate techniques; also, robustness of the methods will need to be examined, including much computer work.

(e) Spatial variation presents interesting problems. There would seem to be a vast area of potential application of the techniques of Matheron and others, to general problems in which spatial variation (usually two-dimensional, but not, in principle, confined to this) is of importance. The stumbling block has been the extending of the correlogram to more than one dimension; many mathematical problems remain and there is room for much in this area.

On the practical side, the idealized mathematical models will often fail to reflect the true physical situations and users should try to exploit the basic techniques, and to examine their properties in practice. This is already being done in geological and mining contexts. It will be important to know properties of estimators of the spatial correlograms, and the robustness of estimation and other techniques.

Acknowledgments

This look into the future would not have been possible if there had not been a good start. I would like to acknowledge the stimulation my early teachers provided. They are John Firestone and Selby Robinson at City College, New York City; and Harold Hotelling and Abraham Wald at Columbia University. A number of colleagues visiting Stanford in the summer of 1976 have discussed with me the topics presented in this report, and I would like to thank particularly Mark Brown, Alan Gelfand, Soloman Kullback, and Michael Stephens. My thanks go also to my regular colleagues in the Statistics Department at Stanford for the many informal chats in the corridors of Sequoia Hall.

The narrative style of this chapter and the theme of the volume have precluded the traditional listing of references in the text. A number of names and topics have emerged in this essay without recourse to their specific origins. For the convenience of the reader there is presented below a bibliography that gives papers and books by subject matter. Within each subject authors are listed alphabetically.

BIBLIOGRAPHY

Simultaneous Parameter Estimation

Berger, James (1975). "Minimax estimation of location vectors for a wide class of densities," *Annals of Statistics*, 3, 1318-1328.

Berger, James (1976). "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss," *Annals of Statistics*, 4, 223-226.

Brown, L. (1971). "Admissible estimators, recurrent diffusions, and insoluble boundary value problems," *Annals of Mathematical Statistics*, 42, 855-903.

Clevenson, M. L. and Zidek, J. V. (1975). "Simultaneous estimation of the means of independent Poisson laws," *Journal of the American Statistical Association*, 66 807-815.

Efron, B. and Morris, C. (1972). "Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case," *Journal of the American Statistical Association*, 67, 130-139.

Efron, B. and Morris, C. (1972). "Empirical Bayes on vector observations: An extension of Stein's method," *Biometrika*, 59, 335-347.

Efron, B. and Morris, C. (1973). "Stein's estimation rule and its competitors . . . An empirical Bayes approach," *Journal of the American Statistical Association*, 68, 117-130.

Efron, B. and Morris, C. (1975). "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311-319.

Efron, B. and Morris, C. (1973). "Combining the possibly related estimation problems (with discussion)," *Journal of the Royal Statistical Society*, B, 35.

Faith, R. (1976). "Minimax Bayes set and point estimators of a multivariate normal mean," Technical Report No. 66, Department of Statistics, University of Michigan.

Fienberg, S. E. and Holland, P. W. (1973). "Simultaneous estimation of multinomial cell probabilities," *Journal of the American Statistical Association*, 68, 683-691.

Hudson, H. M. (1974). "Empirical Bayes estimation," Technical Report No. 58, Department of Statistics, Stanford University.

James, W. and Stein, C. (1961). "Estimation with quadratic loss," *Proceedings of the Fourth Berkely Symposium*, Vol. I, University of California Press, Berkeley.

Joshi, V. M. (1967). "Inadmissibility of the usual confidence sets for the mean of a multivariate normal population," Annals of Mathematical Statistics , 38, 1867-1876.

Peng, J. (1975). "Simultaneous estimation of the parameters of independent Poisson distributions," Technical Report No. 78, Department of Statistics, Stanford University.

Stein, D. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposium*, Vol. I, 197-206.

Stein, C. (1974). "Estimation of the parameters of a multivariate normal distribution I: Estimation of the means," Technical Report No. 63, Department of Statistics, Stanford University.

Stein, C. Efron, B. and Morris, C. (1972). "Improving the usual estimator of a normal covariance matrix," Technical Report No. 37, Department of Statistics, Stanford University.

Stein, C. (1962). "Confidence sets for the mean of a multivariate normal distribution (with discussion)," *Journal of the Royal Statistical Society*, B, 24, 265-296.

Stein, C. (1973). "Estimation of the mean of a multivariate normal distribution," *Proceedings of the Prague Symposium on Asymptotic Statistics*, edited by Jaroslav Hájek, 345-382.

Strawderman, W. E. (1971). "Proper Bayes minimax estimators for the mean of a multivariate normal population," *Annals of Mathematical Statistics*, 42, 385-388.

Goodness of Fit

Anderson, T. W., and Darling, D. A. (1952), "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes," *Annals of Mathematical Statistics*, 23, 193-212.

Anderson, T. W., and Darling, D. A. (1954), "A test for goodness-of-fit," *Journal of the American Statistical Association*, 49, 300-310.

Durbin, J., and Knott, M. (1972). "Components of the Cramer-von Mises Statistics, I," *Journal of the Royal Statistical Society*, B, 34, 290-307.

Durbin, J. Knott, M., and Taylor, C. C. (1975). Components of Cramer-von Mises statistics, II," *Journal of the Royal Statistical Society, B*, 37, 216-237.

Kac, M., Kiefer, J., and Wolfowitz, J. (1955). "On tests of normality and other tests of goodness-of-fit based on distance methods," *Annals of Mathematical Statistics*, 26, 189-211.

Shapiro, S. S., and Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)," *Biometrika*, 52, 591-611.

Shapiro, S. S., and Wilk, M. B. (1968). "Approximations for the null distribution of the W statistic," *Technometrics*, 10, 861-866.

Stephens, M. A., and Maag, U. R. (1968). "Further percentage points for W_N^2," *Biometrika*, 55, 428-430.

Stephens, M. A. (1970). "Use of Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables," *Journal of the Royal Statistical Society, B*, 32, 115-122.

Stephens, M. A. (1974). "EDF statistics for goodness-of-fit and some comparisons," *Journal of the American Statistical Association*, 69, 730-737.

Stephens, M. A. (1974). "Components of goodness-of-fit statistics. *Annals de l'Institut Henri Poincare, Series B*, 10, 37-54.

Stephens, M. A. (1976). "Asymptotic results for goodness-of-fit statistics with unknown parameters," *Annals of Statistics*, 4, 357-369.

Watson, G. S. (1961). "Goodness-of-fit tests of a circle, I," *Biometrika*, 48, 109-114.

Multivariate Data Analysis

Ball, G. H., and Hall, D. J. (1965). *ISODATA, A Novel Method of Data Analysis and Pattern Classification*. (AD 699616) California, Stanford Research Institute.

Chernoff, H. (1973). "The use of faces to represent points in k-dimensional space graphically," *Journal of the American Statistical Association*, 68, 361-368.

Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, 179-188.

Fortier, J. J., and Solomon, H. (1966). "Clustering procedures," In *Proceedings of the International Symposium on Multivariate Analysis*, P. R. Krishnaiah (Ed.), Academic Press, New York.

Friedman, H. P., and Rubin, J. (1967). "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, 62, 1159-1178.

Johnson, S. C. (1967). "Hierarchical clustering schemes," *Psychometrika, 32*, 1159-1178.

King, B. F. (1967). "Step-wise clustering procedures," *Journal of the American Statistical Association*, 62, 86-101.

Kruskal, J. B. (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, 29, 1-17 (a).

Kruskal, J. B. (1964). "Non-metric multidimensional scaling: a numerical method," *Psychometrika*, 29, 115-129 (b).

Kullback, S., and Fisher, Marian (1974). "Multivariate logit analysis," *Biometrische Zeitschrift*, to appear.

Kullback, S. and Ku, H. H. (1974). "Loglinear models in contingency table analysis," *The American Statistician*, 28, 115-122.

Kullback, S. and Reeves, P. H. (1974). "Analysis of interaction between categorical variables," *Biometrische Zeitschrift*, No. 8, to appear.

Kullback, S. (1974). "The information in contingency tables final technical report," U. S. Army Research Office—Durham Grant Number DAHCO 4-74-G-0164.

Kullback, S. (1973). "Estimating and testing interaction parameters in the log-linear model," *Biometrische Zeitschrift*, 15, 371-388.

Kullback, S. (1971). "Marginal homogeneity of multidimensional contingency tables," *Annals of Mathematical Statistics*, 42, 594-606.

Kullback, S., and Khairat, M. A. (1966). "A note on minimum discrimination information," *Annals of Mathematical Statistics*, 37, 279-280.

Kullback, S., Kupperman, M. and Ku, H. H. (1962). "Tests for contingency tables and Markov chains," *Technometrics*, 4, 573-608.

Kullback, S., Kupperman, M., and Ku, H. H. (1962). "An application of information theory to the analysis of contingency tables with a table of 2N ln N, N = 1(1)10,000," *Journal of Research of the National Bureau of Standards, Section B*, 66, 217-243.

Kullback, S. (1959). 'Information Theory and Statistics,' John Wiley and Sons, New York.

MacQueen, J. B. (1967). "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

Mezzich, Juan E. (1975). "An evaluation of quantitative taxonomic methods," Ph.D. Dissertation, Ohio State University.

Pearson, Karl (1901). "On lines and planes of closets fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559-572.

Sokal, R. R., and Rohlf, F. J. (1962). "The comparison of dendrograms by objective methods," *Taxonomy*, 11, 33-40.

Statistics of Directions

Anderson, T. W., and Stephens, M. A. (1972). "Tests for randomness of directions against equatorial and bimodal alternatives," *Biometrika*, 59, 613-622.

Fisher, R. A. (1953). "Dispersion on a sphere," *Proceedings of the Royal Statistical Society, A*, 217, 295-305.

Greenwood, J. A., and Durand, D. (1955). "The distribution of length and components of the sum of n random unit vectors," *Annals of Mathematical Statistics*, 26, 233-246.

Stephens, M. A. (1962). "Exact and approximate tests for directions," *Biometrika*, 49, 547-552.

Stephens, M. A. (1966). "Statistics connected with the uniform distribution; percentage points and application of tests for randomness of directions," *Biometrika*, 53, 235-240.

Stephens, M. A. (1967). "Tests for the dispersion and for the modal vector of a distribution on a sphere," *Biometrika*, 54, 211-223.

Stephens, M. A. (1969). "Tests for randomness of directions against two circular alternatives," *Journal of the American Statistical Association*, 64, 250-289.

Stephens, M. A. (1969). "Multisample tests for the Fisher distribution," *Biometrika*, 56, 169-182.

Watson, G. S. (1956). "Analysis of dispersion on a sphere," *Monthly Notices of the Royal Astronomical Society: Geophysics Supplement*, 7 153-159.

Watson, G. S. (1966). "The statistics of orientation data," *Journal of Geology*, 7, 786-797.

Watson, G. S., and Williams, E. M. (1956). "On the construction of significance tests on the circle and the sphere," *Biometrika*, 43, 344-352.

Seriation

Gelfand, A. E. (1971). "Rapid seriation methods with archeological applications," *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press, 186-1201.

Kendall, D. G. (1963). "A statistical approach to Flinders Petrie's sequence-dating," Bulletin of the International Statistical Institute, 34th Session, Ottawa, 657-680.

Kendall, D. G. (1969a). "Incidence matrices, interval graphs, and seriation in archaeology," *Pacific Journal of Mathematics*, 28, 565-570.

Kendall, D. G. (1969b). "Some problems and methods in statistical archaeology," *World Archaeology*, 1, 68-76.

Kendall, D. G. (1971a). "A mathematical approach to seriation," Philosophical Transactions of the Royal Society of London, Series A, 269, 125-135.

Kendall, D. G. (1971). "Seriation from abundance matrices," *Mathematics in the Archaeological and Historical Sciences*, Edinburgh, University Press, 215-252.

Robinson, W. S. (1951). "A method for chronologically ordering archaeological deposits," *American Antiquity*, 16, 293-301.

Sternin, H. (1965). "Statistical Methods of time sequencing," Stanford University Technical Report No. 112, Stanford University.



Michael Athans, Director of the MIT Electronic Systems Laboratory, has been a member of the faculty of the MIT Department of Electrical Engineering since 1964. From 1961 to 1964, Dr. Athans was employed by the MIT Lincoln Laboratory. Since 1964, he has served as a consultant to Lincoln Laboratory and to many industrial organizations. Dr. Athans was born in Greece and received his electrical engineering degrees from the University of California, Berkeley. He received the Donald P. Eckman Award in 1964 and, in 1969, the American Society for Engineering Education's first Frederick Emmons Terman Award as the outstanding young electrical engineering educator. He is a member of AAAS, Phi Beta Kappa, and Sigma Xi and is a Fellow of IEEE.

PERSPECTIVES IN MODERN CONTROL THEORY

Michael Athans

MIT Electronic Systems Laboratory
Massachusetts Institute of Technology
Cambridge, Mass.

Abstract: This paper reviews the development of modern control theory, with emphasis on future theoretical directions as motivated by expanding areas of application and innovation. Of particular interest are (a) large-scale systems and decentralized control, (b) control using microprocessors, and (c) dynamic system reliability and control under failure.

Modern system theory and its applications deal with decisionmaking under conditions of uncertainty. Of particular importance, and a major source of challenges and complexities, is the case in which the outcomes of decisions are related in a dynamic context; that is, the current outcome or output of a dynamic system depends on past decisions or control inputs. For example, consider the problem of maintaining a moving submarine at a constant depth below the ocean surface. In this case the main output variable of interest, the submarine depth, depends (among other things) on the past history of the positions of submarine control surfaces, the stern plane and the bow plane.

The development of any theory and associated computational algorithms for analysis and design almost always requires the abstraction of reality by approximate yet realistic mathematical relations. For control of dynamic systems these relations take the form of complex, linear or nonlinear, ordinary or partial differential equations, which relate the main system variables of interest, often called *state variables*, to variables that can be directly manipulated manually or automatically. The latter are often called *control variables*.

In addition to the inherent complexity associated with multivariable dynamic systems whose behavior is described by complex differential equations, the control engineer must deal with issues of *uncertainty*. Several sources of uncertainty that are of crucial importance in both analysis and design are:

Errors inherent in modeling a physical system by means of mathematical equations

Errors in the parameters that appear in differential equations of motion (e.g., the submarine hydrodynamic derivatives)

Exogenous stochastic disturbances that influence the time evolution of the system state variables in a random manner (e.g., the effects of surface waves on submarine depth)

Sensor errors and related noise in measurements.

Such uncertainties are modeled as *random variables* and/or *random processes*. Thus, the complete description of any real physical system requires the use of *stochastic differential equations*. Figure 1 is a visualization of the key elements of a stochastic dynamic system.

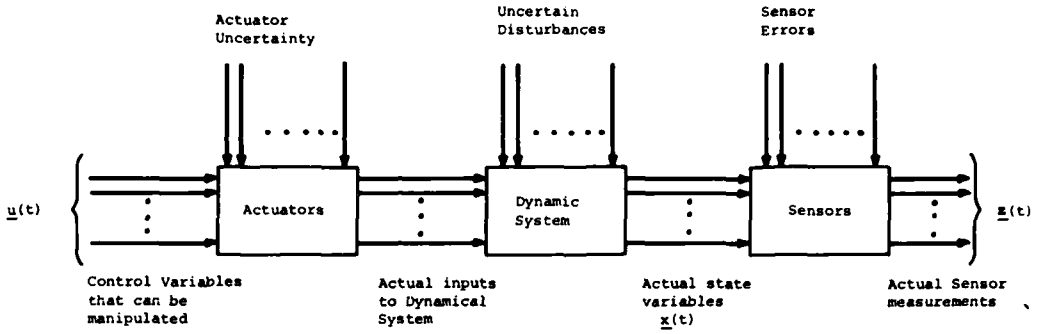


Figure 1—A realistic stochastic dynamic system. From a pragmatic point of view the only variables available for real-time measurement are control inputs u(t) and sensor measurements z(t).

WHAT IS THE CONTROL PROBLEM?

The control engineer is usually given a particular physical system (submarine, aircraft, power system, traffic network, communication system, etc.) that has been designed by others. More often than not, the performance of the system is unsatisfactory; this may be due to interaction of the exogenous disturbance inputs with natural system dynamics, causing unacceptable behavior of system state variables. For example, the system may be inherently unstable in the absence of control, due to the complex interaction of kinetic and potential energy; this is the case with all unaugmented helicopters, missiles, and certain high-performance aircraft. Even if a system is stable, its responses to changes in command inputs may be too oscillatory or too sluggish.

If the behavior of the unaugmented, or "open-loop," system is not satisfactory, then the only way it can be made satisfactory is by judicious manipulation of control variables as a function of the actual sensor measurements. This is often called "feedback control." The main thrust of the control system design problem is to deduce the transformation from the noisy sensor measurements to the control signals. This is illustrated in Figure 2; the device that accomplishes this transformation is called a controller, or compensator. Depending on the nature of the physical problem and the stringency of requirements for overall system performance, physical realization of the feedback controller can be exceedingly simple (e.g., a constant-gain analog amplifier) or complex (e.g., a special-purpose modern digital computer). The appropriate design of the feedback compensator or controller, so that not only is system performance satisfactory but also technological constraints on its implementation are observed, is the essence of the control design problem. These technological constraints can be both hardware and software considerations, cost, weight, reliability, and so on.

HISTORICAL PERSPECTIVE

In this section we present a necessarily very brief history of the techniques available for designing feedback control systems. We hope, however, to convey the intimate interrelationship

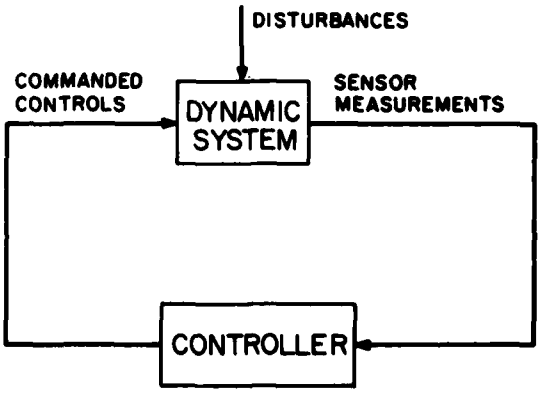


Figure 2—Structure of a centralized stochastic control system

among the development of the theory, motivating applications, available computational tools, and hardware technology for implementation.

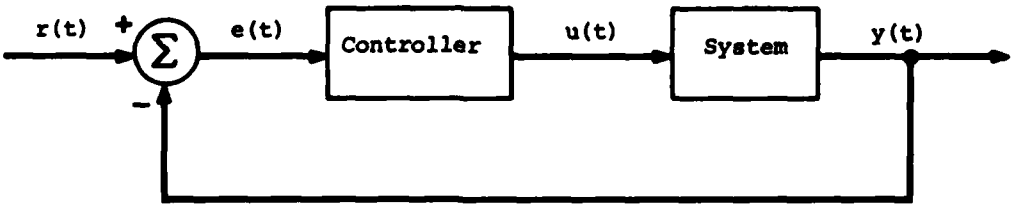
The first phase of the development of control theory took place in the period 1940-1960. We refer to this original brand of theory as *servomechanism theory* or *classical control theory.* During this period the theory was developed for systems described by linear differential equations with constant coefficients and characterized by a single control input. By means of the Laplace transform such systems could be analyzed in the frequency domain, so that the system dynamics could be represented by a transfer function. One of the main motivations for development of the design methodology was the need for accurate fire control systems for both naval and surface weapons systems [1-4]. Later during this period, feedback control of chemical and industrial processes provided additional motivation for theoretical refinements.

The design tools that emerged from classical control theory were, of necessity, greatly influenced by the computational tools and simulation facilities available. Most design tools were graphical in nature (like Nyquist diagrams, Bode plots, Nichol's charts, root locus plots). Closed-form solutions were sought. Since the available theory could not handle nonlinear systems and stochastic effects (with the notable exception of the work of Norbert Wiener [5]) extensive simulations were carried out on electronic analog computers, and much knob-twisting and common-sense engineering was used in arriving at a satisfactory design. Almost exclusively, implementation of the feedback system was by electromechanical and analog-electronic devices.

The basic development of classical control theory can be understood in reference to Figure 3. The basic idea was to have the actual output y(t) "follow" the reference input r(t) as closely as possible. The error signal e(t) was a measure of the undesirable deviation, which was then transformed by the controller into the actual control signal applied to the physical system. At the basic level the issue of how to design the controller so that the error signal would always remain small was the key design problem.

The second phase of the development of a more sophisticated and powerful theory of control is often referred to as *modern control theory.* Its origins are acknowledged to be around 1956, and it is still an extremely active research area. In its early stages, the theory was strongly motivated by the missile and aerospace age and in particular trajectory optimization. Aerospace systems can be extremely nonlinear and, in general, their motion and performance can be influenced by several available control inputs. Since classical control theory represented a scientific design methodology only for linear single-input systems, a much more general design methodology had to be developed for the stringent performance requirements of aerospace systems.

The development of modern control theory and the associated design methodologies were also greatly influenced by the appearance of the modern digital (maxi) computer in the early 1960s. The



r(t): reference input e(t): error signal (e(t) = r(t) - y(t))

y(t): actual output u(t): control input

Figure 3—The traditional servomechanism problem

digital computer greatly influenced the nature of "solutions" to control problems. To be more specific, in classical control theory one almost always sought closed-form solutions; in modern control theory a recursive algorithm is a perfectly acceptable solution to the control problem. This transition from analytical solutions to algorithmic solutions opened several important new research horizons and fresh ways of thinking.

The basic new ingredient of modern control theory was *optimization*. This new attention to "optimal design" was necessitated by the fact that it is difficult to simultaneously examine several control and state variables, as they evolve in time, in order to make a clearcut scientific decision on which design is preferable. Thus, for multivariable control problems it is important to translate the attributes of "good" system performance into a scalar mathematical *index of performance*. This must be optimized subject to the constraints imposed by the system differential equations, as well as additional constraints on the control and state variables, which arise from the physical nature of the problem.

Two powerful theoretical approaches were developed during the early phases of modern control theory. The first approach was an extension of classical calculus of variations methods to the optimal control problem; it was developed by the Russian mathematician L. S. Pontryagin and his students and was called the maximum principle [6-11]). The second approach, due to the U.S. mathematician R. Bellman, was based on the so-called "principle of optimality," an almost self-evident property of optimal solutions, which led to the so-called "dynamic programing" algorithm [12-14].

These two major theoretical breakthroughs in the late 1950s resulted in a worldwide flurry of research during the early 1960s. Several digital computer algorithms were developed to be used for numerical solutions of the complex nonlinear equations that define the optimal control solution, and the theory was applied very successfully to a variety of complex trajectory-optimization problems for both endoatmospheric and exoatmospheric aerospace systems.

Another byproduct of the initial research breakthroughs in dynamic optimization problems was the development of a systematic theory, with associated digital computer algorithms for problems of optimal stochastic estimation and optimal stochastic control.

In stochastic estimation one attempts to reconstruct estimates of key state variables and parameters of a physical system from noisy sensor data. An important class of applications that motivated, and later benefited by, the development of optimal stochastic estimation algorithms was the problem of tracking targets by radar or sonar. The radar or sonar generates noisy range and/or angle measurements; the stochastic estimation algorithms process the noisy sensor data to obtain (a) improved position estimates, (b) velocity estimates, and (c) target classification estimates. There exists a variety of stochastic estimation algorithms, which represent extensions of the celebrated Kalman Filter [15, 16], (the optimal stochastic estimation algorithm for linear dynamic systems subject to Gaussian uncertainties) to systems described by nonlinear equations with respect to their dynamics and measurements [17-19].

Stochastic estimation algorithms have been used extensively for improving position accuracy in inertial navigation systems. Some relatively recent studies show how to couple measurements of the inertial measurements units (IMU) with those obtained from gravitational and/or magnetic field anomalies so as to further improve the position accuracy of a ship or submarine.

Although stochastic estimation theory and the associated algorithms are important by themselves in a variety of applications (such as the tracking and navigation problems), they become even more important when coupled to the control problem. The theory and algorithms associated with optimal stochastic control deal with the overall problem of optimizing an overall system performance index subject to the constraints imposed by the dynamic stochastic differential equations that describe the system behavior as well as the available sensor configuration and their accuracy characteristics.

Most of the theoretical advances in optimal stochastic control have been made during the past decade [20-22]. Optimal stochastic control problems are relatively well understood, because the dynamic programing algorithm can be extended easily to the stochastic case. There remains, how-

ever, certain formidable real-time computational requirements. This class of problems not only combines the issues of deterministic optimization and stochastic estimation, but also includes a considerable interaction between the two. This is the so-called *dual control* problem [23-28]. Roughly, the problem is that in any dynamic optimization problem the *present* values of the control variables should cause the *future* values of the state variables to behave in an optimal manner, and this requires a relatively good knowledge of future system response. Unfortunately, especially in the case of nonlinear systems with uncertain parameters, such knowledge of the future is not available. It may turn out that by applying a control that excites certain modes, we could identify in real time certain key parameters, which would improve our knowledge of future responses. On the other hand, control inputs that are good for identification may not be the best for control.

The preceding argument demonstrates the conceptual complexity of the optimal stochastic control problem. Fortunately, the mathematical formulation of the problem automatically handles all of these complex tradeoffs, and provides the optimal control solution containing the correct balance between the tasks of identification and optimization of the performance index, as a function of time. The practical difficulty is that, at the present state of the art, the real-time computational requirements can be formidable for very complex nonlinear stochastic optimal control problems. To give the reader an idea of the complexity of the real-time computational requirements, it suffices to state that one needs to solve in real time coupled sets of nonlinear partial differential equations; such solutions are beyond the state of the art of current and projected maxicomputers.

The situation is not as grim, however, as one may imagine. Even if computation of truly optimal stochastic control cannot be accomplished, the mathematical theory provides insight into the nature of the optimal solutions. Such insight, together with commonsense engineering know-how about the specific physical problem, can be used for developing near-optimal solutions to several physical problems, still based on a general design methodology. The so-called Linear-Quadratic-Gaussian (LQG) method has been extensively analyzed during the past decade [29-31] and has been successfully applied to several complex problems. The resulting designs show a significant degree of improvement over conventional designs. Of particular naval interest are submarime control [32, 33], jet engine control [34-37], and supertanker control [38].

RECAPITULATION

We have attempted in the discussion so far to simultaneously provide an historical perspective as well as a survey of the state of the art of classical and modern control theory. At present we have good enough conceptual understanding, theories, and design algorithms that we can tackle complex control problems. Of course, there is a gap between available theory and applications. The trend in the past 5 years has been to apply modern control theory to several applications. Needless to say we need many more complex applications to fully reveal the advantages and shortcomings of modern control theory. The shortcomings can then motivate future research at the theoretical, algorithmic, and design methodological level.

In the remainder of this paper we shall outline some exciting future research topics and explain why they are important. Needless to say, the list of topics is not exhaustive. However, it does represent a consensus of international opinion on the most pressing areas for future research, based on diverse applications and the theoretical state of the art.

The need for future advances in control and estimation theory can only be appreciated by viewing this field of research as truly interdisciplinary, applicable not only to complex defense systems but also to other complex engineering and socioeconomic systems, such as interconnected power systems, urban transportation networks, and command-control-communications systems (C^3).

DECENTRALIZED CONTROL AND LARGE-SCALE SYSTEMS

The theories of both classical and modern control theory have been developed under a crucial

key assumption: *centralized decisionmaking*. This can be best understood in reference to Figure 2, where the objective is to design the feedback controller. Notice that the controller (or decisionmaker) has access to *all* the measurements generated by the noisy sensors and generates *all* the controls. Implicit in the theory and associated algorithms is that the controller also has *central* knowledge of (a) the entire system dynamics, (b) the probabilistic description of all uncertain quantities, and (c) the overall index of performance. Although such assumptions are perfectly valid in a variety of applications, it is clear that several complex systems cannot be handled within the existing framework. We present two oversimplified examples that illustrate the point.

Example 1—Consider the problem of defending a fleet of several vessels under attack. The overall objective may be to minimize the expected number of losses of men and equipment. Clearly the course of battle is a stochastic phenomenon, involving real-time decisions about the allocation of sensor resources (radar and sonar) and defense resources (torpedoes, missiles, guns, etc.). A purely decentralized strategy, in which each vessel defends only itself, cannot be optimal, since it does not use effectively the available fleet resources. On the other hand, it is unrealistic to visualize a purely centralized strategy, in which the command center directs at all instants of time every action of the fleet. A centralized strategy can be formulated conceptually, but it is unrealistic from the point of view of communication requirements and the vulnerability of the whole fleet to damage at the central command point. The proper way of handling this problem is to establish some sort of hierarchical command structure, in which the overall defense objective is divided into subobjectives, according to the remaining defense resources.

Example 2—Consider a geographically distributed command-control-communications (C^3) system, which consists of several nodes and links of different capacities, and which is required to handle messages of different priorities. Each node represents a decision point and it must make real-time decisions on how to route different classes of messages over the available links. Under heavy demand, and especially if certain nodes or links become destroyed, this is an exceedingly complex stochastic dynamic control problem. Once more a centralized control-decision strategy does not make sense. The entire resources of the network could be used to pass back-and-forth protocol and status information rather than to transmit useful messages. Once more the real-time optimal decisions, say with respect to routing strategies, can only be made with limited information exchange. For example, each node may be allowed only to communicate with its neighboring nodes. Hence, the optimal control strategy must be decentralized.

The above two examples represent problems of stochastic dynamic systems with distributed decisionmakers (or controllers) and limited communication interfaces. Several other areas, such as power systems, ABM defense systems, transportation networks, and economic systems, have similar general characteristics. In the control literature these are referred to as large-scale systems, and the methodology that must be used is called *decentralized control*.

One could go on and on describing other large-scale systems that require improved dynamic control strategies. However, let us pause and reflect upon their common attributes:

1. They are topologically configured as a network.
2. They are characterized by ill-understood dynamic interrelations.
3. They are geographically distributed.
4. The controllers (or decision points) are many and also geographically distributed.

This class of large-scale system problems certainly cannot be handled by classical servomechanism techniques. Current designs are almost completely ad hoc in nature, backed by extensive simulations, and almost universally studied in static, or at best quasi-static, modes. This is why their performance may deteriorate when severe demands or failures occur.

We do not have a large-scale system theory. We desperately need to develop good theories. The theories that we develop must, however, capture the relevant physical and technological issues. These include not only the traditional performance improvement measures, but also the key issues of (a) communication system requirements and costs and (b) a new word—"distributed computation."

In addressing the problems of large-scale systems and decentralized control we must also recognize that we are facing a critical technological turning point. We are in the beginning of a microprocessor revolution. These cheap and reliable devices offer us low-cost distributed computation. It is obvious that advances in the theory and design methodologies must take into account the current and projected characteristics of microprocessors, distributed computation, and decentralized control.

The development of a theory for decentralized control, with special attention to the issues of distributed computation by microprocessors, must represent a relatively drastic departure in our way of thinking.

Figure 4 shows the type of structure that we must learn to deal with. Once more we have a complex dynamic system that is being controlled by several distinct controllers. Each controller may consist of a single microprocessor or many microprocessors, so that they provide means for distributed computation.

As shown in Figure 4, we have now several controllers or decisionmakers. Each controller receives only a subset of the total sensor measurements and in turn generates only a subset of the decisions or commanded controls.

The key assumption is that each controller does not have instantaneous access to the other measurements and decisions. To visualize the underlying issues, imagine that the complex dynamic system of Figure 4 in an urban traffic grid of one-way streets. Each local controller is the signal light at the intersection. The timing and duration of the green, red, and yellow for each traffic signal is controlled by the queue lengths in the two local one-way links, as measured by magnetic loop detectors. In this traffic situation some sort of signal coordination may be necessary. In the general representation of decentralized control shown in Figure 4, the dotted lines represent the communication-computer interfaces. All boxes and lines with question marks represent design variables. To systematically design the underlying decentralized system, with all the communication and microprocessor interfaces, is the goal of a future large-scale system theory.

The conceptual, theoretical, and algorithmic barriers that we must overcome are enormous.

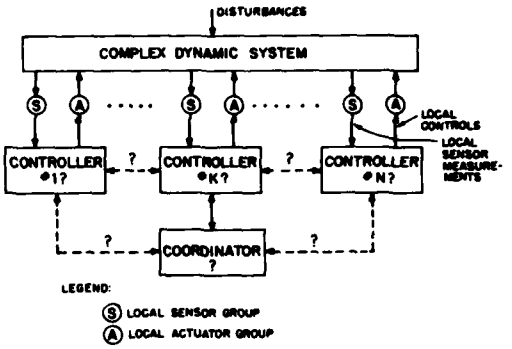


Figure 4—Structure of a decentralized system

There are many reasonable starting points that lead to pitfalls and nonsense [39, 40]. Such decentralized control problems are characterized by so-called "nonclassical information patterns" or "nonnested information structure." This means that each local controller does not have instantaneous access to other measurements and decisions.

Such situations can lead to complicated results. The classic paper of Witsenhausen [41] that demonstrated, via a counterexample, that a very simple linear-quadratic-Gaussian problem has a nonlinear optimal solution was an early indication of the difficulties inherent in decentralized control. Since that time some advances have been made in such fields as dynamic team theory [42-47] and dynamic stochastic games [48-52]. Nonetheless, we have only scratched the surface. We have not seen as yet spectacular theoretical breakthroughs in decentralized control. We are at a normative stage, where old ideas such as feedback are reexamined and new conceptual approaches are investigated.

My feeling is that, concurrently with the theory, we must obtain a much better understanding of the key features associated with different physical large-scale systems. Then, and only then, will we be able to obtain a deep understanding of the kinds of issues associated with large-scale systems, as distinct from the physical, technological and even sociopolitical peculiarities of each system.

We must answer the question of how important a bit of information is for good control. We may have to translate or modify certain results in information theory (such as rate-distortion theory)

to accomplish our goals. Perhaps the deep study of data communication networks will provide a natural setting for basic understanding; the commodity to be controlled is information, and the transmission of information for control routing strategies, or protocol as it is often called, shares the same resources, has the same dynamics, and is subject to the same disturbances.

In summary, the development of new theoretical directions and concepts in decentralized control promises to be one of the most exciting areas of research in decades to come. In spite of the tremendous conceptual and technical problems, the potential payoffs are enormous.

MICROPROCESSOR CONTROL, ALGORITHM COMPLEXITY, AND CONTROL SYSTEM DESIGN*

The potential of microprocessors for conventional control system design is a virgin area for theoretical and applied research. The development of classical and modern control theory was never greatly influenced by computer languages and architecture for two reasons. During the early phases of development, controller implementation was analog in nature. During the later phases, the availability of special-purpose minicomputers for digital control did not present any serious obstacle for implementation.

The availability of reliable low-cost microprocessors presents new opportunities for the design of sophisticated control systems. However, the peculiarities of microprocessors, their architecture, and so on do present problems that cannot be handled by the available theory. If control theory follows its tradition of rapidly exploiting technological innovations (such as the digital computer) for novel and improved designs, then it must face the challenges presented by microprocessors.

Of paramount importance is to incorporate in the overall index of performance not only quantities that pertain to the overall behavior of the control system but also quantities that reflect the complexity of the control algorithms. Besides the usual constraints imposed on the control and state variables by the physical system, we must also include constraints that reflect the use of microprocessors for signal processing and control, such as memory, finite word length, interrupts, and the like.

There is still another area that needs theoretical investigation; most of the existing methodology applicable to design of digital compensators is of the synchronous type. This means that the sampling of sensors and the generation of control commands are carried out at uniform time intervals. On the other hand, nontrivial applications of microprocessors will almost surely require asynchronous operation. Thus we see a divergence of existing theory and desired implementation. This clearly points out that available theory must be reevaluated, modified, and extended; perhaps we may even have to adopt a completely new conceptual framework to keep up with the microprocessor technological innovations. Perhaps the theory does not need a tremendous quantum jump, but certainly several concepts from computer science (such as computational complexity, parallel vs serial computation, automata theory, and finite-state sequential machines) must be incorporated in the formulation of the control problem. To be sure, the mixing up of "continuous" and "discrete" mathematics will lead to severe theoretical difficulties that must be overcome. For example, the author is not aware of any natural and general way of incorporating discrete-valued random variables in digital compensator design. Also, computer scientists interested in computational complexity have not examined in any detail the most common algorithms used in control systems (such as the Lyapunov equation and the Riccati equation). Even if such measures of computational complexity were available, it is not clear how they could be naturally applied either to constraints or to penalty functions in the overall performance index to the optimized. Since the mathematics have to "mesh" together, it is not clear whether variational techniques could be used to solve this class of new optimization problems.

At any rate the theory underlying optimal use of microprocessors and their interconnections for digital compensation has yet to be developed. The

*The material in this section was heavily influenced by a "white paper" recently written by one of my colleagues, Prof. R.L. Johnson [53].

resulting compensators will probably be of the finite-state, asynchronous operation variety, for optimal use of the computational resources. This type of structure may naturally incorporate the common implementation problems, such as model aggregation, interface design, saturation, fault handling, finite-state inputs and outputs, storage allocation, interrupt-handling, and alphabet and programing languages.

FAILURE DETECTION, CONTROL UNDER FAILURE, AND SYSTEM RELIABILITY

Another exciting area for future research deals with the overall problem of reliable control system design and operation. The motivation for studying these types of problems is self-evident, since reliable operation is critical in many applications.

We do not now have a systematic methodology or theory for handling such problems. Reliability theory, as a self-contained discipline, does not appear to be well suited for dealing with the complex dynamic and stochastic situations that one is faced with in control.

Although we do not have a general theory, several theoretical investigations and results emerging in the literature appear to represent promising entries to this very important problem. Several of these concepts were presented at MIT in August 1975 at a workshop, funded by the NASA Ames Research Center on "Systems Reliability Issues for Future Aircraft." The proceedings of this workshop will be published as a NASA Special Publication in the summer of 1977. It was evident from the presentations in that workshop that the present state-of-the-art in constructing reliable designs is to use triple or quadruple redundancy in critical actuators, sensors, and other key components.

With respect to future high-performance systems (aircraft, ships, etc.) the trend is to use larger numbers of control devices and sensors, under complete automatic control. Constructing each new sensor and actuator to be quadruply redundant will result in prohibitive expense. The idea is, then, to try to arrive at systematic means for designing the control system so that redundancy requirements are reduced. In case of sensor or actuator failures (when recognized), one should be able to reorganize the control system so that operative sensors and controllers can maintain safe system operation.

Failure detection and isolation is thus of paramount importance, and some extremely important work has been done in this area during the past 4 years. The field is well surveyed in a recent paper by Willsky [54]. Essentially, the idea of failure detection and isolation relies very heavily on the blending of dynamic stochastic estimation concepts (e.g., Kalman filters) with hypothesis testing ideas. Under normal operating conditions the residuals (innovations) of Kalman filters are monitored. A failure exhibits itself as a change in the statistical properties of the Kalman filter residuals. Once a failure has been detected, one can formulate a set of alternate failure modes and, through the use of generalized likelihood ratios, isolate the failed component.

Within the next 5 years we will see two or three case studies that will give us great insight into the entire issue of failure detection and isolation. From these we will obtain a much better understanding of the inevitable tradeoffs associated with

1. Rapidity of failure recognition
2. Rapidity of failure isolation and classification
3. False alarm probabilities
4. Computational complexity.

Failure detection and isolation is only the tip of the iceberg in the broad area of designing reliable systems. The whole issue of alternate ways of reconfiguring and reorganizing the control system *in real time* after a failure is a wide-open research area. Much research at both theoretical and applied levels must be carried out during the next decade. Of particular importance is the problem of what to do between the time at which a failure is declared and the time at which it has been isolated. During this critical transient one can certainly expect degraded operation of the control system, but the system's stability (under noncatastrophic failures) must be guaranteed.

It is imperative that such a unified theory dealing with failure detection and isolation be developed. The current trend is to concentrate mainly on sensor failures, but the theory and methodology must be extended to other types of

failures, such as abrupt changes in system dynamics, actuator failures, and computational failures. To be sure, redundancy of certain critical components will still be important. However, for military combat systems such as aircraft and high-performance surface-effect ships, it is desirable to distribute the redundant sensors on the vehicle to minimize the probability that an entire group of critical redundant sensors (such as gyros and accelerometers) will be destroyed by enemy fire. However, the geographical distribution of such redundant sensors presents additional problems, since their readings will be influenced by their location. Hence kinematic and structural dynamics must be taken into account in order to use even simple majority-rule voting procedures in triply redundant sensors. Thus, the short-term dynamics of the ship and aircraft, as well as important bending and vibrational modes, must be known relatively accurately so as to minimize the effects of false failure alarms.

In the long run we need a general theory of dynamic system reliability for the design of fail-safe, fail-operational, and fail-degradable control systems. We must develop a methodology that starts with an overall measure of desired reliability and control system performance and provides us with systematic computer-aided design techniques that determine the types of sensors and actuators, their accuracy, their inherent reliability, their redundancy level, their geographical distribution, and their backups (especially in the case of sensors) by software (based on stochastic estimation techniques) that can reduce the level of redundancy. Futhermore, such a theory must incorporate the real-time reconfiguration of the control system, following the onset of one or more noncatastrophic failures, so as to maintain acceptable system performance. To the best of our knowledge very little has been done in formulating in a precise mathematical way this class of problems, and several conceptual barriers must be overcome before a useful set of theoretical tools can be developed.

CONCLUDING REMARKS

We have attempted to define three major areas of future research in control and estimation theory. Such future theoretical directions build upon a solid theoretical foundation available today and are motivated by both significant application needs and technological advances. The theoretical issues and technical details that must be overcome are extremely difficult and diverse. In the development of relevant theoretical and algorithmic tools there is a need for significant interdisciplinary efforts by groups of control engineers, mathematicians, and computer scientists, as well as a great need for advanced applications for rapidly testing the advantages and disadvantages of the new theories.

ACKNOWLEDGMENT

The preparation of this paper was supported in part by ONR contract N00174-76-C-0346. The author is grateful to Dr. Stuart Brodsky of ONR for his critical review of the manuscript and constructive comments.

REFERENCES

1. H. M. James, N. B. Nichols, and R. S. Philips, *Theory of Servomechanisms*, McGraw-Hill Book Co., New York, 1947.
2. G. S. Brown and D. P. Cambell, *Principles of Servomechanisms*, J. Wiley and Sons, New York, 1948.
3. J. J. D'Azzo and C. H. Houpis, *Feedback Control System Analysis and Design*, McGraw-Hill Book Co., New York, 1960.
4. G. C. Newton, L. A. Gould, and J. F. Kaiser, *Analytical Design of Linear Feedback Controls*, J. Wiley and Sons, New York, 1957.
5. N. Wiener, *The Interpolation and Smoothing of Stationary Time Series*, MIT Press, Cambridge, Mass., 1949.
6. L. S. Pontryagin, "Optimal Control Processes" (in Russian), *Usp. Mat. Nauk*. **14**, 3-20 (1959); translated in *Amer. Math. Soc. Trans*. **18**, 321-339 (1961).

7. L. S. Pontryagin, et al., *The Mathematical Theory of Optimal Processes*, J. Wiley and Sons, Interscience, New York, 1962.
8. L. I. Rozonoer, "L. S. Pontryagin's Maximum Principle in the Theory of Optimal Systems I, II, III," *Automation Remote Contr.* **20**, 1288-1302, 1405-1421, 1517-1532 (1960).
9. M. Athans and P. L. Falb, *Optimal Control*, McGraw-Hill Book Co., New York, 1966.
10. E. B. Lee and L. Marcus, *Foundations of Optimal Control Theory*, J. Wiley and Sons, New York, 1967.
11. A. E. Bryson and Y.-C. Ho, *Applied Optimal Control*, Blaisdell, Waltham, Mass., 1969.
12. R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, N.J., 1957.
13. R. Bellman and S. E. Dreyfus, *Applied Dynamic Programming*, Princeton University Press, Princeton, N.J., 1962.
14. S. E. Dreyfus, *Dynamic Programming and the Calculus of Variations*, Academic Press, New York, 1965.
15. R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Trans. ASME, J. Basic Eng.* (Series D) **82**, 34-45 (1960).
16. R. E. Kalman and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," *Trans. ASME, J. Basic Eng.* (Series D) **83**, 95-108 (1961).
17. R. S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*, J. Wiley and Sons, New York, 1962.
18. A. H. Jazwinskii, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.
19. A. Gelb, *Applied Optimal Estimation*, MIT Press, Cambridge, Mass., 1974.
20. M. Aoki, *Optimization of Stochastic Systems*, Academic Press, New York, 1967.
21. K. J. Astrom, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
22. H. J. Kushner, *Introduction to Stochastic Control*, Holt, Rinehart, and Winston, New York, 1971.
23. A. A. Fel'baum, *Optimal Control Systems*, Academic Press, New York, 1967.
24. B. Wittenmark, "Stochastic Adaptive Control Methods: A Survey," *Int. J. Contr. 21*, 705-730 (1975).
25. M. Athans and P. Varaiya, "A Survey of Adaptive Stochastic Control Methods," in ERDA Report CONF-78067, *Systems Engineering for Power: Status and Prospects*, L. H. Fink and K. Carlsen, eds., pp. 356-366, Oct. 1975.
26. E. Tse, Y. Bar-Shalom, and L. Meier, "Wide Sense Adaptive Dual Control for Nonlinear Systems," *IEEE Trans. Automat. Contr.* **AC-18**, 98-108 (1973).
27. E. Tse and Y. Bar-Shalom, "An Actively Adaptive Control for Linear Systems with Random Parameters via the Dual Control Method," *IEEE Trans. Automat. Contr.* **AC-18**, 109-116 (1973).
28. Y. Bar-Shalom and E. Tse, "Concepts and Methods in Stochastic Control," in *Control and Dynamic Systems: Advances in Theory and Applications*, C. T. Leondes, ed., Academic Press, New York, 1975.
29. M. Athans, ed., "Special Issue on Linear Quadratic Gaussian Problem," *IEEE Trans. Automat. Contr.* **AC-16**, (Dec. 1971).
30. B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*, Prentice Hall, Englewood Cliffs, N.J., 1971.
31. H. Kwackernaak and R. Sivan, *Linear Optimal Control Systems*, J. Wiley and Sons, New York, 1972.
32. J. Griffin et al., "Advanced Concepts for Submarine Control," Analytic Sciences Corp., Report TR-662-1 (ONR-CR-289-001-1F), Reading, Mass., 1976.
33. D. L. Kleinman, W. Killingworth, and W. Smith, "Automatic Depth Keeping Control for the Trident Submarine," Systems Control, Inc., Report 101, Palo Alto, Calif., Oct. 1973 (Confidential Report, Unclassified Title).
34. D. L. DeHoff and W. E. Hall, "Multivariable Control Design Principles with Application to the F-100 Turbofan Engine," *Proc. 1976 Joint Automat. Contr. Conf.*, West Lafayette, Ind., July 1976, Amer. Soc. Mech. Engr., New York, 1976.
35. G. J. Michael and F. A. Farrar, "Development of Optimal Control Modes for Advanced Technology Propulsion Systems," United Aircraft Research Labs, Report N911620-2, East Hartford, Conn., Mar., 1974.
36. C. R. Stone, "Turbine Engine Control Synthesis," Honeywell Systems and Research Division, Final Report AF Contract F33615-72-C-2190, Minneapolis, Minn., 1976.
37. F. A. Farrar and G. J. Michael, "Analyses Related to Implementation of Multivariable Control Techniques for ten F100-F401 Class of Engines," United Aircraft Research Laboratory Report UARL-M177, East Hartford, Conn., 1973.
38. K. Astrom *et al.*, "Estimation and Control for Supertankers Using the Self-Tuning Regulator Method," Submitted to *Automatic*.
39. N. R. Sandell, P. Varaiya, and M. Athans, "A Survey of Decentralized Control Methods for Large Scale Systems," in ERDA Report CONF-750876, *Systems Engineering for Power: Status*

and Prospects, L. H. Fink and K. Carlsen, eds., pp. 334-352, Oct. 1975.
40. M. Athans, "Survey of Decentralized Control Methods," *Ann. Econ. Soc. Meas.* **4**, 345-356 (1975).
41. H. S. Witsenhausen, "A Counterexample in Stochastic Optimal Control," *SIAM J. Contr.* **6** (1968)
42. Y.-C. Ho and S. K. Mitter, eds., "*Directions in Large-Scale Systems*," Plenum Press, New York, 1976.
43. *Proc. IFAC Symp. Large Scale Syst. Udine, Italy, June 1976*. (G. Guardabrasi and A. Loratelli, eds.), Instr. Soc. Amer., Pittsburgh, Pa., 1976.
44. Y.-C. Ho and K. C. Chu, "Team Decision Theory and Information Structures in Optimal Control Problems—Parts I and II," *IEEE Trans. Automat. Contr.* **AC-17**, 15-28 (1972).
45. Y.-C. Ho and K. C. Chu, "Information Structure in Dynamic Multi-Person Control Problems," *Automatica* **10**, 341-351 (1974).
46. N. R. Sandell and M. Athans, "Solution of Some Non-Classical Log Stochastic Decision Problems," *IEEE Trans. Automat. Contr.* **AC-19**, 108-116 (1974).
47. C.-Y. Chong and M. Athans, "On the Periodic Coordination of Linear Stochastic Systems," *Automatica* **12** (1976).
48. J. B. Cruz, Jr., "Survey of Nash and Stackelberg Equilibrium Strategies in Dynamic Games," *Ann. Econ. Soc. Meas.* **4**, 339-344 (1975).
49. D. Castanon and M. Athans, "On Stochastic Dynamic Stackelberg Strategies," *Automatica* **12**, 177-183 (1976).
50. D. Castanon, "Equilibria in Stochastic Dynamic Games of Stackelberg Type," MIT Electronic Systems Laboratory Report ESL-R-662, Cambridge, Mass., May 1976.
51. Y.-C. Ho and F.-K. Sun, "Value of Information in Two Team Zero Sum Problems," *J. Optimization Theor. Appl.* **14**, 557-571 (1974).
52. T. Basar, "A New Class of Nash Strategies for M-Person Differential Games with Mixed Information Structure," *Proc. 1975 IFAC, Cambridge, Mass., 1975*, Instr. Soc. Amer., Pittsburgh, Pa., 1975.
53. T. L. Johnson, "Finite-State Compensators for Physical Systems," MIT Electronic Systems Laboratory Technical Memo ESL-TM-658, Apr. 1976.
54. A. S. Willsky, "A Survey of Design Methods for Failure Detection in Dynamic Systems," MIT Electronic Systems Laboratory Report ESL-P-633, Nov. 1975 (to appear in *Automatica*).

William Cummins is Associate Technical Director for Ship Performance and Head of the Ship Performance Department at the David W. Taylor Naval Ship Research and Development Center. Dr. Cummins directs theoretical, experimental, and computer-simulation investigations of resistance, propulsion, seakeeping, and maneuvering for craft ranging from hydrofoil, planing, ACV, and SES craft to all types of displacement ships, submarines, and cable-towed body systems. He serves as consultant to a number of other Navy organizations, including the fleet, on questions of ship hydromechanics. Dr. Cummins received a B.S. in Naval Architecture and Marine Engineering from the Webb Institute of Naval Architecture and a Ph.D. in Mathematics from American University. He has received the David W. Taylor Award for outstanding achievement, the Davidson Gold Medal of the Society of Naval Architects and Marine Engineers, and the Navy Distinguished Civilian Service Award.



HYDROMECHANICS RESEARCH AND THE NAVY: A PROJECTION

W.E. Cummins

David W. Taylor Naval Ship R&D Center
Carderock, Md.

This paper is about the relation of hydromechanics research to the Navy, its importance, how it contributes to development and design, how this contribution can be increased, and the directions research might take in the future. We will not be specifically concerned with the research of the past and present, though we will briefly review the past for the lessons it should have taught us and the present for trends that are likely to continue. The foundation for our projection will be a review of the important unsolved hydrodynamic problems arising from design of both traditional ship types and radically different types of promise.

The Navy has had a strong positive attitude toward research throughout most of this century, and particularly since World War II. However, it has not always been skillful in exploiting this research. The technology available has frequently been far beyond that exhibited in the designs of our fighting ships. We will examine some of the reasons for this and suggest ways to improve the process of translating applied science into design practice.

In this period of economic and political constraint, the support of research as an act of faith is no longer tenable. The question, "Do we need it?" spoken or unspoken, is in the minds of the financial decisionmakers. We must ensure that the Navy gets the greatest possible benefit from the dwindling research dollar.

The problem has two complementary aspects—(a) insuring that the right research is carried out, and (b) exploiting these results when they become available. As we have hinted, the second aspect presents the greater difficulty. Many results of good research which relate to real problems are incorporated in fleet hardware only when circumstances force the issue.

An example from early in my career shows that the problem is not new. I first saw a model of a fleet-type submarine many years ago at the Taylor Model Basin. I asked why the hull form used was so unsuited to submerged operation, with a ship-like shear line, an unstreamlined superstructure, and a submerged speed far below its speed on the surface. It was obvious that radical improvements could be made. The argument that nothing better was needed held until the Germans gave us great difficulty during World War II with a true submersible, capable of much higher submerged speed. The rapid development of the "guppy," by merely cleaning up our fleet type, and the later development of the first modern U.S. submarine, the experimental *ALBACORE*, showed that the technology was there, waiting to be used!

Research in hydromechanics is as old as humanity's move into the waterways and the oceans. The pioneers had never heard of scientific method, but it is obvious that they practiced it—sometimes very effectively. There have been

many successful ship and boat types, both historic and prehistoric. The Viking ships were well suited for their means of propulsion, and the Thames barges were shaped to carry cargoes easily through the restricted passages of the British canals. The clipper ships of the last century achieved average speeds over long voyages which exceed those of most ships of the age of steam.

These successes were not achieved by magic. They were the result of evolutionary or revolutionary ideas, conceived after careful observation of then current practice and tested by sometimes daring innovation in design. Few of the failures have been recorded, leaving an apparent history of continuous success, but we can be sure that failures were there—the galley that was hard to row, the barge that had difficulty in narrow quarters, the clipper ship that had excessive passage times. However, there was steady progress throughout the centuries, and it is certain that it was the result of something very akin to the scientific method. The lessons learned may not have appeared in learned journals, but they became visible improvements in ships that went to sea. We could emulate them to our benefit!

It should be noted that until fairly recently, naval hydromechanics has been extremely conservative. There have been radical variations in size, in construction, in source of power, and in mission, but rarely as the result of hydromechanical breakthroughs. The "inventions" tended to focus on other features, and hydrodynamic innovations followed as needed.

This century has seen a change, with the advent of hydrofoil craft, the air-cushion concept, the low-waterplane catamaran, and various hybrids, but it remains true today that the feasibility of any new type depends as much on the availability of an efficient, lightweight power source as it does on hydromechanic performance. It also remains true that the hydrodynamicist is frequently given a very difficult assignment of making some strange configuration successful.

It might be concluded that hydromechanic research is no great thing, since naval architects have repeatedly demonstrated the ability to reach a near optimum design for a given set of constraints. However, this neglects the growing cost of failure, and a hydrodynamic failure can be spectular. Water is a very unforgiving medium. A fault in hydrodynamic design can result in a vessel that cannot reach its design speed, in vibration so bad that structure and equipement fall apart, in a propeller that erodes from cavitation in a few hours, or in an inability of the vessel to perform its mission in the weather conditions of its operating area. (If the reader suspects that such dangers are exaggerated, it may be noted that every one cited has been suffered in a design of the recent past.)

The first "modern" research in ship hydromechanics began in Great Britain over a century ago with the work of William Froude. He built the first operational ship model towing tank and carried out an analysis of model and ship resistance. He recognized that there were two principal components to the resistance to moving a ship through water—a frictional resistance, analogous to the drag on water moving through a pipe, and a resistance due to waves generated on the water surface. He hypothesized that if he could separate these two components in the drag of a ship model and project each to full scale by its appropriate law, he could predict full-scale resistance. The procedures have since been sharpened and given a more reasonable scientific foundation, but operational change has not been great. We now know that Froude's procedure is nothing more than a good approximation, and we shall see that current developments are taking us beyond its range of validity. Still, his brilliant and successful application of an essentially empirical approach has been an inspiration to hydrodynamicists ever since.

True hydromechanic research began in the U.S. Navy when Admiral Taylor built the Experimental Model Basin in 1898. Admiral Taylor's efforts quickly brought the U.S. Navy to the front. Taylor recognized a need of the designer well beyond that satisfied by Froude's method—a knowledge of the laws relating ship resistance to ship form. Since there was no theoretical foundation for establishing these laws, he also used an empirical approach based on systematic variation in model shapes. His exposition was well suited to the use of the designer, and to this day it remains a primary tool in preliminary phases of design. He was also responsible for many other innovations in ship design, experimental techniques, and applied theory.

Progress during this century has tended to be evolutionary, with the emphasis on improving the

ability of the designer to reach his design objectives with confidence. The emphasis of the work has been predominantly on "conventional" ships, since these are the types that the Navy has most often built. This pattern has radically changed in the recent past, with much greater attention being given to "exotic" vessel types. This is a consequence of concern by both investigators and high-level decisionmakers that there might be a better way to meet the Navy's requirements. This trend toward radical innovation is very healthy, but the emphasis on reliable techniques and data must be maintained, together with careful and objective evaluation of the true merits of competing concepts.

This recent history shows that we must allow hydromechanic research to follow two complementary paths. The continued development of technology in support of the ships actually being designed as part of our fleet must not be neglected. These designs are not usually radical departures from past practice, but rather evolutionary developments in both form and function. It is essential that they be designed with confidence, and, as we shall see, this frequently requires intense research of quite narrow scope. On the other hand, it is equally important to develop new concepts and offer improved options to satisfy the Navy's mission needs. But these options, if they are to become reality, require the same kind of careful development of a suitable technology base as the conventional options of the present.

HYDRODYNAMIC PROBLEMS IN DESIGN

The stated purpose of this paper is to project trends in naval hydrodynamic research and offer suggestions for how it can be more effectively used. Starting with a review of the problems arising in design efforts being carried out today, we break the discussion into two parts—the unsolved design problems of conventional ships, where the object is to be able to design with confidence, and the new frontiers in hydromechanics, which arise from revolutionary concepts of current interest.

Conventional Design

"Conventional," as used here, means single-hull vessels, supported primarily by buoyancy. While it is true that the form of many of the better ships being designed today would appear unremarkable to designers of the last century, it is also true that the term "conventional" embraces some rather strange shapes. For these hulls, function is all-important, and the hydrodynamic designer must make the best compromise he can. Enormous bulbs may be fitted to the forefoot to house sonar equipment, the transom width may continue the midship beam, form coefficients may move outside the old ranges, and appendages may be elephantine. The important consequence for the hydrodynamicist is that the old empirical rules do not always work, and the opportunity for serious design error is greatly increased. The old technology, even when based on sound scientific principles, is not quite relevant. It was established over decades of careful research, and an extension suitable for current needs would be an endeavor of nearly the same magnitude. There is neither time nor funding for such a program. Thus, there is great interest in developing a more basic understanding based on the laws of physics. This would provide a foundation for solving the problem of creating a successful design of a hull type which is unlike any existing parent. In other words, designers must resort to basic principles much more than in the past.

There follows a review of a number of the problems which face the hydrodynamicist today. While some of them are not new, in their current aspects they go far beyond old experience.

The unusual sizes, proportions, and forms of some of the new designs make it extremely difficult to predict the full-scale resistance. As we have mentioned, the classic Froude method of separating frictional and wavemaking drag of a ship model and projecting each to full scale by its own law is an approximation at best. The frictional drag is estimated from flat plate drag at the proper Reynolds Number, even though the hull may be far from a flat plate. The residual drag (total minus frictional) is treated as wavemaking, even though it includes any error in the pure frictional drag, form drag due to separation of the flow around hull or appendages, and other known and unknown effects. The standard practice is to make a prediction using the Froude procedure, and then to add a "fudge factor," called a correlation allowance, to get the proper full-scale value.

The correlation allowance is based on past experience with similar hulls. But suppose there are no similar hulls in our data bank. Undesirably large errors can result, which may mean that the ship will not reach design speed. These problems have always existed, but they are worse today because we stray further from the beaten path. In this wandering, we sometimes even encounter new phenomena. For example, predictions of the power needed for the supertankers in use today have been found to require very large corrections in order to make them fit into the traditional scheme. Spray resistance associated with the very unusual bow forms is believed to be the source of the difficulty, but there are other possibilities, associated with peculiar apsects of flow about the model or the ship. Sometimes there are problems at model scale—the presence of laminar rather than turbulent flow in the boundary layer, the greater tendency toward flow separation at model scale, the effect of surface tension on model waveforms, and many other effects assumed to be small. The correlation allowance hides many sins.

There is an increasing need for knowledge about details of flow in local regions of the ship. For example, flow in the neighborhood of the stem is not well understood, in spite of the many ships that have been built. The design of the stem is frequently made by the lines draftsman, with only esthetics and past habits as a guide. An error here, though, can result in ventilation or cavitation that can degrade other aspects of performance (more on these phenomena later). Further aft, flows about appendages, particularly separated flows, can be sources of vibration or degrade the performance of the propeller. The flow in the neighborhood of the propeller is of the greatest importance and will be discussed in some detail. Also, at the stern, as at the bow, there exists little real guidance for the designer on how to lay out the transom demanded by ship function or arrangement. What are the penalties of width, depth, and area?

The powerplant of preference today for many designs is the gas turbine. The gas turbine in its present configuration has no ability to reverse, as the steam turbine does. The usual means for permitting the ship to go astern is to adopt a controllable and reversing pitch (CRP) propeller. This requires an oversize shaft to contain the pitch change actuators, an oversize hub to contain the very complicated pitch changing mechanism, oversize struts to support the oversize shaft and hub, propeller blades limited in geometry so they can pass each other when reversing pitch. It also poses a propeller strength problem requiring for solution precise knowledge of the hydrodynamic loads. The hydrodynamic penalty is an approximately 10% increase in ship drag from the oversize appendages and an extremely difficult propeller design.

No matter how good the hull design, the ship is not satisfactory unless it has an efficient propeller, compatible with the ship's powerplant and well adapted to work in the highly turbulent environment under the stern. In one sense, propeller design is an advanced technology, taking advantage of the science of airfoil and wing theory. The basic procedures were laid out many years ago, and there have been continuous improvements. Propellers can now be designed to minimize vibration and to recover much of the energy lost by the hull in frictional resistance. However, there is one major requirement; the designer must know in detail the environment in which the propeller is working. There are two difficulties. The only source of information about the wake is a model test. The usual procedure is to measure the wake structure in the plane of the propeller (without the propeller present, of course) and to correct it for the presence of the propeller by certain integrated measures determined from an experiment with an existing or "stock" propeller. The propeller, however, is operating in the region of the model most sensitive to "scale" effect, because the difference between model and full-scale boundary layers is greatest at the stern. Thus, the ship propeller is designed for the model wake. The error is important since both the efficiency and vibration-preventive qualities of the propeller depend very much on the details of the flow. The second difficulty is that the propeller itself modifies the inflow, so that the measured wake structure is altered by the presence of the propeller, perhaps differently at model and full scale.

One of the concerns of the propeller designer is to ensure that it has adequate strength under all operating conditions. For this, he needs to know the loading, chordwise and spanwise. The great-

est stresses very likely occur under some condition other than the primary design condition, (for example in a crashback maneuver in which the engine goes from full ahead to full astern as fast as possible). Two things are needed, neither of which is adequately known—the inflow that the propeller experiences during the maneuver, and the loading on the propeller blade in such off-design conditions.

A final problem of increasing concern to the propeller designer is cavitation. One of the unique aspects of hydrodynamics as opposed to aerodynamics is the vapor cavity that forms when the pressure in the fluid drops below vapor pressure. Boiling under reduced pressure is a type of cavitation. It occurs near a body (such as a propeller blade) when the dynamics of the flow cause the pressure to fall below a certain threshold. There are a number of types of cavitation that may occur on conventional propellers. The core of a tip or hub vortex is a low-pressure region that can explode into a very stable cavity extending far aft. Steady or unsteady cavities can form on either the pressure face or the suction face of the blade. The higher the ship speed, the greater the load on the propeller; at some speed the pressure will go below this threshold, and the propeller will start to cavitate, with a resultant degradation in performance. It cannot be avoided, but by careful design it can be delayed to higher speed. Therefore it is most important to be able to predict inception speed.

The phenomenon would appear to be one that can be readily explored and predicted by model techniques. The procedure is to test the propeller in a variable-pressure water tunnel where the absolute pressure can be scaled. But most remarkably, the techniques that have been developed are suitable for qualitative judgment only. Predicted inception speeds can be in error by a factor of two or more (usually a nonconservative overestimate). It has become evident that cavitation inception is a very complex process indeed, and that we are far from understanding its physics. The traditional empirical approach as well as the elementary theoretical approaches have failed. The designer is left with sometimes inconsistent experience as his only guide.

This brief review of the problems facing the designer of the "conventional" ship would not be complete without a discussion of ship dynamics, or, more explicitly, design for performance and survival in a real ocean. Until recently, this was treated according to rules of thumb learned from experience. The variability of storm conditions made the environment difficult to describe and to treat in a rational way. Experimental and theoretical research was carried out on models or ships moving in regular sinusoidal waves, but the conditions were considered so unrealistic that designers gave the results little credence. If the rules of thumb were sound, the design was good. If they were deficient or irrelevant, the design could be a failure.

Ensuring that a design has satisfactory seakeeping qualities is a problem that has received proper scientific treatment only in the last two decades. This work borrowed the techniques of stochastic processes and adapted them to a workable description of a storm sea. The earlier "academic" results on regular waves were an essential building block of the new approach.

The design community now recognizes that seakeeping considerations must be introduced into the design process in a much more refined fashion if we are to consistently obtain a successful design, a ship that is able to reliably carry out its mission in the environment in which it is expected to operate.

In spite of the fact that much progress has been made on the basics of ships' response to storm seas, there remain a number of extremely difficult problems for investigators and designers to consider.

Our knowledge of the environment is deficient. We have a usable model, but we lack the body of statistical data needed to give it substance. Man has traveled the oceans for thousands of years, but there is little quantitative information on the nature and variability of natural waves. We know that there is usually a local sea due to the local wind. It is a function of windspeed and fetch (the distance over which it has been blowing). However, this is superimposed on one or more wave trains or swells coming from distant storms. The resulting spectrum of sea conditions at any point in time and space can have as much individuality as a fingerprint. Until we have much more information about the "population statistics" of ocean wave spectra, there will be important gaps in the

rational treatment of seakeeping in ship design. Specifically, the designer should know the frequency of sea conditions that will degrade the performance of the ship and its various systems to an unacceptable level.

The variability of sea conditions and our lack of knowledge of their statistics are elements in another problem facing both the designer and the buyer of a ship. The designer can address effectively only those requirements that can be specified with precision and that relate to some aspect of ship performance which can be measured against the specification. Seakeeping is a quality that is surprisingly difficult to pin down in such a fashion. Qualities such as smooth-water speed, propeller performance at design speed, turning radius, etc., are easily specified and easily measured, but performance in rough water involves many responses, both rigid body and elastic, that are functions of the particular seaway in which the ship finds itsef. Any of these responses can be at a level that degrades the ability of the ship to carry out its mission. All should be considered in establishing a measure for seakeeping performance. This has turned out to be a very complex and as yet unsolved problem in logic.

The question of actual prediction of ship responses, given the description of the sea in which it is operating, is the part of the problem that has received greatest attention, both experimentally and theoretically. Some of the simplifying assumptions are rather drastic and are known to be in some error, but the success has been comparable to that of the Froude hypothesis of a century ago. To an extent this is a consequence of the fact that the uncertainties in the sea environment are greater than the uncertainties in the prediction techniques. Nevertheless, certain important deficiencies have become evident. Most of the responses have been successfully treated by linear techniques. That is by assuming that if the wave excitation can be subdivided into a set of components (say, sine waves), the response to the total is the sum of responses to the components. At higher levels of excitation all responses can be expected to become nonlinear, and this greatly increases the difficulty of treatment, both by theory and by experiment. Sometimes the modification is slight, as in pitch and heave, but even here the slight modification could have an important effect on such questions as the amount of freeboard the designer should incorporate to ensure dry decks. Sometimes the effect is large, as in the case of roll, which does not yet have a usable scientific foundation. Another important nonlinear problem is the prediction of the time history of the pressure distribution when a ship slams. (A slam occurs when the ship bow emerges from the surface of the water and then crashes back, setting the entire hull into a low frequency vibration). Local damage to the shell plating as well as dangerous stresses in the hull girder can result.

In summary, the demands of the designer of the conventional ship cannot be satisfied by the essentially empirical techniques of the last century. For proper design of hulls for resistance and seakeeping and for the achievement of efficient, vibration-free propulsion, the designer needs much more detailed information on the interaction of the hull and propeller with the environment. This can be achieved only by a much better understanding of the mechanisms at work. Thus, these hard demands of the designer for guidance in making engineering decisions are forcing the hydrodynamicist to examine his basics. Empiricism is no longer enough, and in fact can become a great danger. Because of the very large number of variables, the risk of misinterpreting a few pieces of data is great. Not every investigator or designer can be as clever as Froude or Taylor. Empiricism remains important, but it must be supported by insight and understanding whenever possible.

Radical Options

The term "radical options" is considered here to include all design concepts other than conventional displacement vessels. It thus represents a wide variety of configurations intended for an equally wide variety of applications. For our present discussion, there is one feature that they have in common; they have no foundation of technology to support design comparable to that available for conventional craft. If we were to follow the path that was selected for conventional vessels, we would need to repeat the efforts of at least a century of research for each configuration. This is impossible, if such a concept is expected to be a realistic option for the ship buyer. Thus, the

hydrodynamicist is forced into very nearly the same situation as with conventional craft. That is, he must resort to basic principles, not in this case to establish detailed knowledge for precision in design, but to develop insights into the factors that govern their performance. The range of problems is enormous, as the variety of configurations is virtually unlimited.

We will start with a brief catalog of concepts that are of current or recent interest, recognizing that an enthusiastic inventor can expand the list at any time.

Floating platforms—These are nonshiplike structures intended to remain, more or less, in particular locations; they usually include working platforms, buoyancy chambers, and connecting members. They are widely used in the oil industry but have also been used for oceanographic purposes and have been considered as floating naval bases. Static stability and survivability in storm seas are their main hydromechanical requirements.

Catamarans—The catamaran is an age-old concept that has many attractions for modern applications. The USS HAYES, designed as an oceanographic research ship, is an example. The twin hulls give great transverse stability at the expense of high roll acceleration. Low damping in pitch can result in hydrodynamic impact on the bridging structure between the hulls in seas that are synchronous with the natural period of the ship in pitch. The large wetted surface of the two hulls means frictional resistance will be high in comparison with that of a conventional hull. Nevertheless, this is an excellent low-speed platform for certain special applications.

SWATH—(Small Waterplane Area, Twin Hull) This represents an attempt to remedy the problems of the catamaran. It is a twin-hull configuration in which the beam of the hulls is greatly reduced where they pass through the water surface. They have good seakeeping characteristics under most conditions and can be designed to have low wavemaking resistance. Wetted surface is great, so frictional resistance is high. This is an attractive configuration for applications that require medium to moderately high speed, very good seakeeping qualities, and large deck area.

Planing craft—This is another old concept. It depends for its support on the dynamic lift on its bottom rather than on hydrostatic displacement. It is an inexpensive configuration with low drag, suitable for intermediate to high speed in smooth water. It is deficient in seakeeping qualities in most configurations, so speed is degraded rapidly as the sea state rises. Technology is better established for this concept than for most other radical options, but there remain important gaps.

Hydrofoil craft—This, like the planing craft idea, is a dynamic lift concept. The hull is supported above the water on strutlike columns attached to foils or wings that run below the surface. There are a variety of configurations, suitable for applications from intermediate to very high speed. Configuration options include fully wetted foils with active controls, fixed surface-piercing foils, and supercavitating or ventilated foils.

Air-cushion vehicles (ACV)—These craft ride on a cushion of air, contained by an air-filled toroidal elastic bag. They can be amphibious, and the elasticity of the toroidal bag permits them to ride over moderate obstacles. They are capable of fairly high speed. Loss of air from the cushion must be replaced by a blower, and power consumption can be high even though the water drag is fairly low. However, this is a very useful configuration where its amphibious qualities can be used (in river rapids, over mud flats, and up beaches, for example).

Surface-Effect Ships (SES)—This is an air-cushion vehicle that has rigid sidewalls and elastic end closures to contain the air. The sidewalls extend down into the water to form a more effective seal. The SES gains efficiency in cushion air use, at a loss of amphibious capability. Some configurations are suitable for very high speed operation (80 n.mi./h or more). Accelerations in waves at high speed can be a problem. Plans to build a 3,000 ton experimental craft are underway.

Hybrid craft—These include many concepts that combine buoyancy, dynamic lift from foils or planing surfaces, and air cushions. An example would be a SWATH with a foil between the hulls, or a planing craft partly supported by a hydrofoil. The field is wide open for the clever inventor, and some possibilities are very attractive for certain unique applications. By clever combination of the various elements it may be possible to create a

point design that overcomes the disadvantages of any "pure" configuration.

Any one of these vehicle types has a list of associated problems far more extensive than that detailed for the displacement hull. In many cases, our knowledge is so limited that we cannot even define the problems precisely. Therefore, our discussion of the demands on the hydrodynamicist will take a different approach; we shall discuss generic problems that are common to some or all of the types but take different forms for different configurations.

We start with the Froude problem—the analysis of vehicle resistance and prediction of full-scale values. The difficulties we have discussed for conventional ships, important though they are, fade into insignificance. We have many new components, or old components that now assume a greater percentage of the total. More physical phenomena contribute, so increased difficulty with scale effects is probable. In addition to the usual wavemaking and frictional resistance, there may be interference drag among struts, buoyancy elements, lifting surfaces, and other appendages; frictional or wavemaking drag on elastic elements such as inflated bags; spray drag; drag associated with air supply in air-cushion systems; drag associated with ventilation and cavitation; and many others. Some of these are subtle and hard to identify. Others are immediately evident but difficult to analyze. A simple example of the sort of problem that can occur is characteristic of one of the simplest configurations, the planing craft. Model and full-scale prototype planing craft do not run at the same trim angle (angle between the base line and the horizontal) due to differences in the frictional resistance coefficient and other hydrodynamic factors. Wavemaking resistance, the greatest single component, is extremely sensitive to trim angle, because it affects the geometry in a fundamental fashion. Therefore, prediction of the resistance at full scale involves corrections based on an uncertain estimate of the differences between model and full-scale attitude.

The next problem we might call the Taylor problem—prediction of the variation in resistance due to changes in configuration parameters. (For example, the effect of varying distance between hulls on multiple-hull configurations or of varying the arrangement and spacing of struts on hydrofoil craft). Since the configuration are frequently complex, the number of parameters needed to describe any particular configuration can be enormous. An empirical approach like Taylor's "Standard Series" is out of the question. Other methods must be used to establish a foundation for arriving at an efficient design.

Propulsion is a universal problem both in finding a suitable propulsor and in arranging an efficient and practicable geometry of the propulsor in relation to the other components of the configuration. Some configurations require propellers on power-wasting struts, others require air screws, some make the inefficient water jet seem very attractive. For the very high speed applications we have the supercavitating propeller, the ventilated propeller, and the partially submerged propeller. The latter is most attractive because in some geometries it is possible to reduce or eliminate the drag of exposed propeller shafts and struts. From the point of view of the designer and the hydrodynamicist, there is no adequate technological base for the design of any of these concepts. In any application it is necessary to "cut and try" at model scale and to pray vigorously at full scale.

We have mentioned cavitation and ventilation. At the speeds at which some of these craft are expected to operate (60 n.mi./h or higher), cavitation is a fact of life and cannot be avoided, so the designer must include its existence in his plans. The vapor cavities are no longer incipient, but are fully developed and may extend well aft of the cavitating surface. Thus, the low-pressure side of a propeller blade may be completely hidden in a stable cavity. Such conditions are called "supercavitating." Ventilated cavities resemble vapor cavities but are filled with air, sometimes at atmospheric pressure, instead of vapor. Ventilation can be useful, as when it is used to stabilize a cavity that could otherwise be intermittent. It can also be harmful or even destructive. When a vapor cavity vents to the atmosphere and suddenly becomes ventilated, there is a large step change in the forces on the cavitating body. If not quickly controlled, the craft can become unstable. Another example of the danger of sudden ventilation is provided by a hydrofoil craft in a turn. The near-vertical struts become lifting surfaces, and

the low-pressure sides of the struts may cavitate or ventilate. Sudden ventilation can cause an instantaneous reversal in the strut lift, which may throw the craft out of control. We have mentioned that cavitation inception has not been successfully modeled, even when pressures have been properly scaled. Full vapor cavities behave much better, and a mathematical theory of cavity flow has been useful for the design of supercavitating lifting surfaces. But ventilation does not always behave like cavitation, particularly in its dynamic phases. Current thinking is that the study of ventilation may require full-scale speeds at atmospheric pressure rather than just cavitation scaling.

Many of these exotic options are attractive in smooth water. Their geometries are well suited to rapid travel over a flat water surface. When the surface is roughened by storm or swell, there is an important reordering of these concepts. The SWATH, for example, with its long natural periods in all modes of response, becomes attractive in spite of its rather high frictional drag. Its performance is degraded only in following or quartering seas at certain wavelengths, and these should be operationally avoidable. The hydrofoil craft is also attractive, since buoyancy excitation from waves is completely eliminated and the variation in lift due to the orbital fluid motions in waves can be eliminated or reduced by a well-designed control system that controls the angle of foils. Planing craft, on the other hand, may go to the bottom of the list because of poor performance and high acceleration in a seaway.

A problem arises from the need for establishing the relative merits of the various concepts. We have mentioned the analogous problem for surface ships. Here, the situation is much worse, for the range of variation is enormous. Some types may behave well in a wind sea from ahead but roll badly in a long swell from abeam; others may behave badly in a short swell from ahead, or have trouble holding course in quartering seas. These differences can be the overriding consideration in selection of type, and the differences are meaningful only in relation to the operational environment and the mission. The need for realistic environmental information is just as important here as it is for conventional ships.

Many of the craft have dynamical features that are not well understood. The unsteady planing surface, for example, has not received adequate theoretical treatment. A more complex example is the air cushion of the ACV and SES concepts. The dynamics of the cushion when the craft experiences vertical motions while traveling in waves are a function of air compressibility, the elasticity of the elastic bags, and the dynamics of the cushion air supply, as well as of Froude Number and Reynolds Number. Thus, the scaling problem is very difficult. If a model test is carried out at atmospheric pressure, the air in the cushion is not sufficiently compressible at model scale. Model tests may be suitable for qualitative studies or for validation of the controlling equations, which then could be used for full-scale prediction, but they are unreliable for providing design information directly.

We have given a sampling of the problems facing the developer and the hydrodynamicist who supports him, who are together responsible for turning one of these concepts into reality. It is obvious that there is an important difference between this situation here and that of conventional ships. Here we are working at the frontiers, and we need to map out the gross features of the technology. Ultimately, as we proceed through exploratory and advanced development, we are thrown back into the mode of providing reliable design support information of the sort discussed earlier. Chances and consequences of error are orders of magnitude greater than for more conventional options. The danger of failure will always be finite, but the hydrodynamicist must do everything possible to reduce this danger to an acceptable level. Otherwise, these concepts will remain merely attractive ideas, which we do not have the will to turn into reality.

TRENDS OF RESEARCH

We have reviewed in some detail the hydrodynamic problems generated by trends in conventional design and by competitive new options. We now examine how hydrodynamic scientists are responding to these needs, and the more promising directions this research is taking.

Theoretical hydrodynamics has not been considered a particularly useful discipline by naval

hydrodynamicists until fairly recently. The facts that most problems of interest involve a turbulent boundary layer and that the theoretical approaches either ignore viscosity completely or are limited to very low Reynolds Number flows which are completely laminar, was taken by the empiricists to suggest that hydrodynamic theory was an interesting but academic exercise. It is evident from his writings that Admiral Taylor did not completely accept this attitude, and the success of airfoil and wing theory in the early decades of this century was strong evidence that there was value here. The last few decades have seen a tremendous increase in the use of theory to attack a wide variety of practical problems.

The usual methods follow a classic pattern, the ingenious adding of solutions to the partial differential equations that govern the flow to build a solution that satisfies the various boundary conditions. This requires that the problem have an important quality; both field equations and boundary conditions must be linear. Unfortunately, many of the problems of the naval hydrodynamicist involve nonlinear conditions. The boundary-condition equation at the free surface, for example, is quadratic in the velocity of the fluid. The approach, of course, has been to linearize, sometimes ruthlessly. It is remarkable how successful these techniques have been—not for all problems, and not for all cases of a class of problems, but often enough to provide the hydrodynamicist a useful tool.

Ship motion theory is an outstanding example. By a rather crude technique of superimposing essentially two-dimensional solutions stacked in vertical layers along the length of the ship, the boundary condition at the ship is approximately satisfied. The solution, in principle, assumes that the hull is vertical at the free surface and that the waves to which the ship is responding are infinitesimal. However when this very simplified theory is used to predict the behavior of a real ship model in finite waves, it usually works. One can obtain from it engineering quality solutions that are supported by experiment. One of the problems of the hydrodynamicist is that the design engineer may become so confident in the results of the technique that he will forget its weak theoretical foundation.

Another recent example that illustrates the power of these techniques is provided by research on the SWATH concept. This success is particularly remarkable in view of the long history of very mediocre results from corresponding research on conventional craft. The problem is the theoretical prediction of wave resistance. Perhaps because the wave resistance is fairly low for these configurations, a very usable theory has been developed. More important, it can and has been used to optimize hull form for a given set of design constraints. This is most valuable because the inherent high frictional drag of this configuration makes it important to minimize all other contributions.

These procedures have been extremely powerful for generating insights into how the solutions depend on the various defining parameters. There is no doubt that they will continue to be used, because they can map out the "global" character of a configuration more efficiently than any other currently available technique. The limitations are important; the procedures are limited to linear or weakly nonlinear problems in which the boundary layer and the other viscous effects play an insignificant or identifiable and separable role. There is always a need for validation, because it is not always obvious where the methods will break down.

We may expect continual improvement in these ideal fluid techniques from two sources. The methods themselves are continuing to evolve as more and more sophisticated techniques are incorporated. For example, the recent use of matched asymptotic expansions makes it possible to tie local solutions, which are valid only in the neighborhood of the body, to far-field solutions, which are valid only far from the body. The second source of progress is the modern computer. The solutions frequently appear in the form of very formidable multiple integrals. Not long ago, when the analyst reached this point, he had to stop further efforts. Continuing progress in computer capability has had a revolutionary effect.

We terminate this review of ideal fluid techniques with a discussion of two apparently simple problems that have not been satisfactorily treated:

The question of flow about the ship stem has been mentioned as a problem of concern. It can be crudely idealized as the free-surface flow past a near-vertical wedge extending indefinitely aft. In

this idealization, the flows for various stream velocities can be considered as all self-similar. That is, if one were to define a Froude Number based on some length associated with the wave disturbance, all of these flows would be nondimensionally equivalent. Secondly, since the wedge angle is finite, one can expect that the wave slope at the stem is also finite and independent of velocity. This discussion applies only to the local flow, but it suggests that on an actual ship, no matter how low the speed past the stem, the free-surface condition is inherently nonlinear.

Now, if we open up our wedge to a large included angle and make it enter the water vertically rather than translating horizontally, we have an idealized geometry of a ship slamming. Here again we have a set of self-similar flows; it is only necessary to adjust the time and length scales to make them all look alike. Also, again, there are strong nonlinearities. There is a spray sheet climbing up the sides of the wedge, so that slopes, velocities, and surface elevations reach levels that invalidate the linearized equations. One can imagine that the planing surface, whether steady or oscillating, has similar difficulties.

Because of inherent limitations, these "classical" techniques can take us only so far. Even when viscosity is neglected, their capacity is limited. Fortunately, the high-speed, large-memory computer has opened the way for a set of completely new techniques. They are collected under the general label of "numerical hydromechanics," and they include procedures based on finite-element and finite-difference analyses and Fast Fourier Transform algorithms. They are not very much affected by nonlinearities, either in the field equations or in the boundary conditions. Time-dependent flows introduce an additional variable, but appear to be within current computer capability. It is not intended that classical procedures be rejected out of hand; where they can be blended in to reduce the computational load, they will be used. The important result is that the numerical hydrodynamicist has moved theoretical hydromechanics forward.

The limitations (other than computer capability) arise mainly from our knowledge of the physics involved. The boundary layer, if introduced at all, must be introduced in the form of empirical equations rather than the Navier-Stokes equations for viscous flows. Cavitation inception cannot be treated until we have a satisfactory understanding of the mechanisms involved. Even so, these techniques make it possible to attack many problems that the theoretical hydrodynamicist has avoided in the past because of strong nonlinearity. The problems listed at the end of the discussion of classical techniques are obvious candidates.

An old problem that is under attack by these methods is the theoretical calculation of the waves generated by a conventional displacement vessel moving in smooth water. This problem has been attacked by a series of brilliant investigators, using classical approaches, for most of this century. The results must be considered rather mediocre, as we have noted. The cause is attributed variously to the linearization of the free-surface boundary condition, the linearization of the body surface condition, the neglect of the boundary layer, and some combination of these. It will be most illuminating to see if removing the nonlinearities leads to a better solution.

We have made continual reference to the boundary layer as an essential, complicating factor in most problems of hydrodynamics. Much of our knowledge of the boundary layer is purely empirical, because of the highly complex mechanisms at work. Even so we know a great deal about them, and we are learning more. We are concerned with two types, the laminar boundary layer, in which the particles move in steady, smooth paths, and the turbulent boundary, which involves locally violent mixing. Naval hydrodynamicists have usually rejected the laminar boundary layer, in both thought and experiment. Since at full scale laminar flow is restricted to a very small region near the stem, they would like to ignore it at model scale because it confuses resistance analysis. They sometimes use mechanical devices such as sand, studs, or wires bonded to the models' surfaces, to trip the boundary layer into a turbulent mode. They have been much more concerned with gross effects, such as those of roughness, than with any of the physical mechanisms actually at work. In other words, they have tended to be engineering empiricists, rather than scientists, seeking answers suitable for application. Recent developments suggest that they have been somewhat rash and that sev-

eral important unsolved problems involve mechanisms in the boundary layer.

It is worthwhile to review some aspects of the boundary layer on a model or ship. When a body is moving through quiet water (no environmental turbulence), the flow starts off as laminar. As we progress aft there comes a point at which the smooth streamlines start to oscillate (Tollmien-Schlichting waves). These waves increase in amplitude until we reach a second point (the transition point), where the waves break and the boundary layer becomes completely turbulent. We know that the location of transition is a function of free-stream velocity and the pressure distribution over the body. A favorable pressure gradient (pressure decreasing in the direction of the fluid velocity) delays transition to a point further aft than it would be on a body with no pressure gradient. Empirical criteria have been developed to make it possible to estimate the location of transition, and it appears that elasticity of the body wall (as in the skin of a porpoise) may also delay it. Roughness may trigger it (hence the use of a wire or sand trip), but if the trip is placed too far forward in the stable region, the flow may again become laminar.

Why does this process concern naval hydrodynamicists? It is mainly because some systematic empirical procedures in which they placed great faith may have led to incorrect conclusions about the relation between ship form and ship resistance. It now appears that laminar flow near the model bow was of significantly different extent on different forms tested, and differences in measured resistance were sometimes due to differences in frictional resistance rather than differences of form. Another reason for interest will appear in our discussion of cavitation.

Turbulent boundary layers have been the subject of extensive research because they represent the main mechanism of frictional resistance. The research has taken the form of rational analysis of empirical data. Techniques have been developed for calculating boundary layers on both flat plates and bodies of revolution. When these techniques have been extended to cover an arbitrary body, the door will be open for the solution of one of the most important practical problems, design of a propeller to operate in a full-scale wake.

This discussion of boundary layer research will not be complete without reference to the effect of additives on frictional drag. If a long-chain polymer is dissolved in the fluid that makes up the boundary layer, the boundary layer's profile is drastically changed and resistance is greatly reduced. The mechanism is not understood, but the effect has been well demonstrated. Research has been concentrated on the relative merits of various polymers, both for reducing drag and for avoiding the degradation that takes place as the molecules travel aft in the turbulent boundary layer.

The need for more information about cavitation and ventilation was a recurrent theme in our discussions of the problems of both "conventional" and "radical" design concepts. Many of these problems are peculiar to a particular configuration; in this section, where we are concerned with the more scientific or general aspects, we will not discuss them further. Pressure scaling is necessary, and a variety of facilities for cavitation research has been built over the years, including recently some remarkable new ones. The variable-pressure water tunnel has been the traditional facility. These are generally limited to flows without a free surface and are used mainly for studying components such as propellers and the associated appendages. The desire to simulate both cavitation number (pressure scaling) and Froude Number (wave scaling) has led to the development of variable-pressure free-surface water channels and, most recently, the variable-pressure towing tank at Wageningen, Holland, where 10-m ship models can be towed under reduced atmospheric pressure. In this remarkable facility, models are taken in and out and attached to the carriage by remote control, all without breaking the vacuum. The David Taylor Naval Ship Research and Development Center is building a very high speed carriage capable of 100 knots for cavitation work at atmospheric pressure. Froude Number scaling is rejected, but at these speeds wave effects are believed to be secondary. More important are the natural atmosphere and high Reynolds Number.

From the point of view of the writer, of even greater importance is research on the physics of cavitation, directed toward an improved understanding that might explain some of the many paradoxes. As we have noted, pressure scaling is

not enough. Froude Number scaling offers no significant improvement. Air content has been believed to be a factor and was measured for many years. It is now recognized that dissolved air is of secondary importance but that air nuclei trapped on solid particles can be very important. (As the water in the Wageningen vacuum tank ages, it appears to be losing its nuclei, and cavitation is becoming more erratic with time.) Nuclei in a crevice on the surface of the body can contribute just as much as free nuclei. None of these effects seems to explain the early inception of cavitation at full scale. Some current research suggests that mechanisms of the boundary layer are important, particularly in the laminar region just forward of the transition point. There appear to be nonsteady reductions in pressure in the region of the Tollmien-Schlichting waves above the wall of the body. These reductions can reach down to vapor pressure with a resulting transient cavity—in other words, transition. If this tentative result is confirmed by further research, it could help explain the scaling problem, since transition on the model and at full scale differs geometrically and dynamically. Much more research is needed, both theoretical and experimental, to fully explore these relationships. The desired result would be a mathematical model, based on true physical mechanisms, which would be suitable for predicting full-scale performance.

Little if any work has been done on the related phenomenon of ventilation, but it is greatly needed. We have just enough information to know that the differences between ventilation and cavitation are as important as the similarities. This is virgin territory and deserves a vastly increased effort.

CONCLUSION

We have reviewed the same subject matter—problems and trends affecting hydrodynamic research—from three quite different positions. First, we have examined the needs of the ship designer for accurate, detailed, and relevant hydrodynamic information to support decisions to be made in the design of a well-defined ship for a well-defined purpose. Second, we have discussed the demands made on hydrodynamicists by engineers who are attempting to develop radically different options, where neither mission nor details of configuration may be specified or even understood. Finally, we have examined the problem areas from the point of view of the applied hydrodynamicist himself, who must develop techniques to satisfy his two customers. While there are relationships in the three views, the differences are great. However, one theme is common—a need for an understanding of the mechanisms involved more complete than that given by the purely empirical procedures so popular in the past. For the designer, empiricism without understanding may lead to an important design error. For the developer of new concepts, there is neither time nor funding for developing a technology on a purely empirical base. He needs early guidance to the paths most likely to be profitable. For the hydrodynamicist who must satisfy these needs, the old technology is not adequate to the new demands being placed upon him. As quickly as possible, he must provide reliable information for his customers. More powerful techniques, solidly based on the physical mechanisms at work, are an absolute requirement.

In our introduction we noted the difficulties the Navy has experienced in getting its new technologies into hardware, and we promised some comments on how this transfer could be improved. We offer no grand solution; the problem is inherently difficult and is beyond the scope of this review, but we can identify some contributing elements that *must* be part of any solution that has a chance of working.

There are several participants in the system who contribute to the problem: the research manager, who modulates the flow of funds to support research; the hydrodynamicist, who carries out the research; the designer or developer, who applies the results of the research, if it is to be applied; and the ship buyer, who orders the ship that provides an opportunity to exploit the research.

Two other characters in the drama sometimes play very important roles: the "advocate" and the operator. By "advocate" we mean an individual who has taken the position of enthusiastically defending a certain concept approach, no matter what its merits with respect to alternatives. The

operator is the person who has actual operational experience with the technology in question.

The research manager's role is critical; he should be the source of wisdom in the system. He has power, because he controls the money and because he leads the drive for additional funds when needed. It is his responsibility to see that proper support is provided to satisfy the needs of the designer and the developer. Indeed, his most important problem is a strategic one—what is the appropriate division of funding between the short-term needs of the designer, the longer term needs of the developer, and the fundamental needs of the hydrodynamicist himself.

The hydrodynamicist sometimes thinks of himself as a scientist and sometimes as an engineer. He is very pleased when his work is recognized by his peers, and he is equally pleased when it is used by the designer. He is puzzled about why the designer does not immediately exploit his wonderful findings, and he is annoyed when the designer misinterprets them. He is also annoyed when the research manager cuts off the support of some attractive line of research.

The designer has little time to study scientific papers. He has a schedule to meet, and if the information is not available when he must make a decision, he makes it anyway. He would like all hydrodynamic problems to be solvable by a few minutes' work at a computer terminal, using programs already developed and stored. In fact, he has shown readiness to respond to new technology, if it is available in a form that can interface with his operation.

The ship buyer is interested in a highly effective, low-cost, and, most important, low-risk design. He has at most a nonprofessional interest in research and new technology. If some untried new development is proposed, his response is likely to be, "Not on my ship." He is not reactionary; it is just that he cannot afford to gamble.

The advocate is very frequently an inventor who is outside the system. He can equally well be in the system—he is particularly likely to be a research manager or a hydrodynamicist—and when his advocacy is discovered by the other players he should be relieved of all decisionmaking authority. His objectivity is suspect. The advocate is important, though, because he sometimes performs the important task of perturbing the system, forcing attention to options that would otherwise be rejected. The available technology is rarely sufficient to justify his claims, but his enthusiasm may be sufficient to stimulate its development. His authority, however, must be restricted.

The operator is usually a minor character, ignored by the other participants. This is unfortunate, because he can be a source of information about actual problems that the others have neglected. He may not be a scientist or engineer, and he may not be able to state his knowledge in the terms they would prefer, but he can inject a touch of the real world.

How can the system be made more effective? First, by recognizing the peculiar natures of the participants and the constraints under which they operate. The manager, since he has the role of the wise man, must be familiar with the objectives of the hydrodynamicist. He should recognize potential for the Navy and measure its importance. He must listen to the needs of the designer and respond to them. He must understand the requirements of the ship buyer. Above all, he must remain objective.

The hydrodynamicist must go most of the way in making his research accessible to the designer; he must learn to adapt his findings to this purpose. He should learn how his work fits in the larger context of the Navy and avoid carrying his work beyond the point of useful return. He must recognize that he is part of the system and that his independent existence cannot be justified.

The designer can help by anticipating needs before they become critical. He must maintain an awareness of the potential of research results and advise the hydrodynamicist on how they can be made most useful to him. He should allow the hydrodynamicist an advisory role in design; this will both improve design and educate the hydrodynamicist. He must advise the ship buyer as to options and help him in developing specifications that will lead to real advances.

The ship buyer must consider objectively options that exploit advances in technology. While risk taking must be limited, careful advanced development can reduce it to an acceptable level. Above all, experimental prototypes should be supported when the basic technology has been sufficiently established.

In short, any management system that is expected to improve the Navy's ability to exploit research must foster among the participants a common understanding and respect of their various roles, greatly improved communication, and true cooperation.

BIOLOGICAL AND MEDICAL SCIENCES



Robert D. Myers has been Professor of Psychological Sciences and Head of the Laboratory of Neuropsychology at Purdue University since 1965. Dr. Myers is also Director of the Psychobiology Program at Purdue and holds a joint appointment as Professor of Biological Sciences. He is a regional editor of *Pharmacology Biochemistry and Behavior,* advisory editor for *Physiology and Behavior,* and editor of the series *Methods in Psychobiology.* He has been a member of the National Institute of Mental Health, Advisory Panel on Alcoholism and of the National Science Foundation Advisory Panels on Psychobiology and Neurobiology, has written more than 100 articles and other scientific papers, and has lectured widely in the United States and abroad. He taught at Colgate University from 1956 to 1963. From 1963 to 1965 he was a Visiting Scientist in Physiology and Pharmacology at the National Institute for Medical Research in London, and in 1969 he returned there as *Visiting Professor of Neuropharmacology. Dr. Myers earned a* B.S. from Ursinus College, and M.S. and Ph.D. degrees in Psychology from Purdue University. He took postdoctoral training in neurophysiology in 1960-1961 at the Johns Hopkins University School of Medicine. He received the Ursinus College Outstanding Alumnus Award in 1967 and, in 1971, the Sigma Xi award for meritorious research in the Neurosciences at Purdue University.

CHEMICAL FACTORS IN THE BRAIN INVOLVED IN LIFE-SUSTAINING REGULATORY MECHANISMS

R.D. Myers

Purdue University
West Lafayette, Ind.

Within the human brain resides the most intricate chemical "laboratory" known to mankind. Quite remarkably, different substances are newly synthesized within this organ, degraded metabolically, carried by ultrastructural means by one of a dozen transport processes, and even act as transcellular messengers. What is really fascinating about all of this are two additional facts. First, those compounds and elements found in the brain are the same as those that circulate throughout the bloodstream and likewise exist in other tissue and organs of the body. Second, they are distributed in the brain and compartmentalized very precisely according to subunits—individual anatomical structures. This second fact dispels an historical notion that cerebral tissue is amorphous and an undifferentiated mass of neurons—a "bowl full of nervous jelly."

In this article, I want to describe how so-called nerve transmitters and other neurohumoral factors present in the brain of every higher animal and human alike can operate to control and maintain our vital, life-sustaining processes. To do this, we shall first explore the nature of a transmitter, where it is located, how it comes into being, its final disposition, and, lastly, how it works. Next will be considered the compelling concept of a "neurohumoral code" as it pertains to the special functions of hunger and feeding, thirst, sexual behavior, stress, sleep, emotion, and aggression. Finally, the regulation of body temperature will be used as a special case to illustrate a model of the neurochemical "coding" process. Thermoregulation was selected because, at this writing, more information is available about the brain's cellular "thermostatic" control system than any other.

Special applications of the present state of knowledge pertaining to distinctive neurohumoral "codes" will be pointed out. With reference to the future directions that this field will take, the prospects that lie ahead of us are very great indeed with respect to a clear understanding of basic control mechanisms in the brain.

NEUROTRANSMITTER CONCEPT

Historically, the transmission of an impulse from one nerve to another or from a nerve to the surface of a muscle was always thought to be mediated by a surge of electrical current at their junction—i.e., the synapse. With the monumental work of Loewi, Dale, and other European scientists, it soon became apparent that a nerve impulse is instead carried across the tiny cleft of a synapse by means of a lightning-fast chemical process. At the terminal of the sending neuron, a compound later identified unequivocally as acetylcholine (ACh) is released. This packet of released ACh touches upon ultrasensitive recep-

tors located just across the cleft on the next neuron; these receptors have chemical features that make them specifically reactive to ACh [1].

Although all of the early findings were gathered from experiments on the peripheral nervous system—mainly at the neuromuscular junction—it was realized, even as long as 40 years ago, that the principles of transmission learned from the periphery could just as easily apply to elements of the central nervous system (CNS). ACh could also be released in the same way from nerve endings in the brain which would in turn activate or trigger an impulse on the following neuron. If the process were repeated, a chain of activity of neurons along a pathway would be established. This makes sense in the CNS; since ACh is degraded extremely rapidly, the necessity is fulfilled for exceptional speed in the repetitive firing of neurons. Such swift reactions are required by the acts of seeing, hearing, thinking, moving, and a myriad of other cerebrally mediated processes. In fact, almost all of these necessitate an immediate sort of processing mechanism of the neurons in the brain. Today, ACh is considered to be a transmitter in the CNS just as it is in the periphery.

With the succession of discoveries in the early 1950s that many other substances also occurred endogenously in the brain of the mammal, the idea soon became prevalent that ACh is not the only compound that is released from the presynaptic nerve terminal onto a post synaptic receptor complex. Over the last 15 years, elegant microscopic methods have enabled the scientist to visualize the chemical individuality of neurons that fluoresce differentially. In fact, with each passing year, anatomical "maps" are being constructed continually which provide us with pictures of the pathways that neurons take as they traverse the brain. Tracks of chemically distinct bundles of nerves are being traced.

As a first step, a major future direction of the research in this field will be to identify the precise chemical makeup of each subunit of the brain. A second step will be to isolate the miniscule fiber-like connections between each of these individual structures. Ultimately then, one can reconstruct the enormously complex "chemical wiring" diagram of the human brain. With enterprise, this may be achieved by the end of this century.

The Synapse

In addition to ACh, only three of the many other chemical factors will be dealt with in detail here: serotonin (5-HT), norepinephrine, and dopamine. Their selection is based simply on the large amount of literature accumulated on them. These three compounds, termed monoamines, are distributed unevenly in the brains of rats, cats, humans, and other species. What is more, some nerves in the CNS contain mainly, if not exclusively, only one of these chemical substances.

Where much of the functional action takes place in the nervous system is at the synaptic junction between two nerve cells. Figure 1 presents a schematic diagram which details the

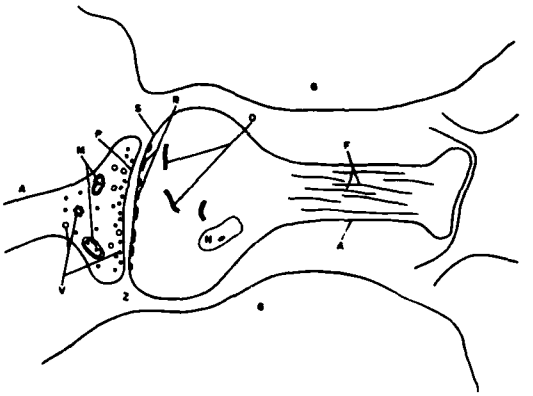


Figure 1.—A schematic representation of a neuron positioned between the axonal ending of a second neuron and the dendritic process of a third. The fine detail of the synaptic coupling and other ultrastructural elements are illustrated which could be affected by the chemical applied to this tissue. Abbreviations are as follows: A-axon membrane, F—neurofibril, G—glial cell, M;—mitochondrion, N—nucleus, O—organelle, P—presynaptic membrane, R—receptor site, S—postsynaptic membrane, V—vesicle, Z—synaptic zone or cleft [2].

terminal end of one nerve cell (at the left) and the next cell (center) upon which it impinges. Especially notable are the tiny vesicles which line the presynaptic membrane. It is these vesicles that contain the transmitter substance. Juxtaposed across the synaptic cleft is the long line of receptor sites which stand ready to receive the transmitter material. After the substance (i.e., ACh, serotonin, norepinephrine, or dopamine) is synthesized locally within the respective cell, it is stored within the vesicles at the nerve terminus.

Depending on its vesicular constituent, an individual neuron is thus called a cholinergic (ACh), serotonergic, noradrenergic (containing norepinephrine), or a dopaminergic neuron. A collection of these individual amine-containing neurons with their long processes (axons) forms a bundle of chemically specific fibers. Thus, terms such as a "cholinergic fiber pathway" or a "noradrenergic bundle" are used to describe a given piece of anatomical architecture, delineated according to its own unique chemical feature.

At the nerve ending, storage pools serve to keep extra transmitter substance that is manufactured. It is believed today, however, that the "functional pool" comprised of the vesicles at the edge of the synapse (Figure 1) releases the transmitter substance as soon as the appropriate physical stimulus to do so is received at the nerve ending.

Synaptic Funtion

A nerve cell is one of several excitable tissues in the body. An impulse is propagated along an individual nerve fiber by a local change in ionic current on the cell's membrane. Briefly, the change in polarity (so-called depolarization) of the nerve membrane occurs as the positively charged sodium ions, located externally, enter the neuron. This results in the outside of the neuron membrane being transiently negative relative to the inside of the nerve membrane. This wave of negativity, as it is propagated along the nerve fiber constitutes the physical element which is transmitted.

As this negative potential reaches the terminal end of the nerve cell, the charge acts in a millisecond flash to evacuate the transmitter material from its presynaptic depot—the vesicles. Once the transmitter substance exits into the synaptic cleft, it readily attaches itself to the receptors on the postsynaptic membrane across the way on the next nerve cell (Figure 1). The transmitter-specific receptors are a protein complex which, through literally thousands of innovative experiments, has been characterized pharmacologically and classified according to several arbitrary designations. If sufficient receptor protein is temporarily activated by the impact of incoming transmitter from the previous cell, the cell membrane itself opens up the local channel for sodium ions. The subsequent entry of this ion species once again depolarizes this region of membrane and the negativity cycle begins anew for this next cell.

At present, researchers in nerve physiology and nerve chemistry are sorting out the chemical nature of the ultrastructural elements of the receptor complex. Future research of an exacting nature will be directed toward the electron or other microscopic identification of a given receptor complex. The chemical isolation, separation, and characterization of receptor material are also on the virgin threshold. When this knowledge is available, a most important forward advance will occur in science. Why? If the makeup of receptor material is known, drugs can be developed that act specifically to either block, partially inhibit, or perhaps potentiate receptor activity of a particular neuronal pathway. It is easy to see, therefore, how demyelinating disorders and Parkinson's and other diseases could be ameliorated by such compounds.

The Neurohumoral Code

The concept of neurochemical "coding" is derived from the geneticists' explanation of the patterning of genes. In the neurosciences, the term coding refers to a particular systematization of physiological signals and events within the brain. In the general sense, a neurochemical systematization at the level of the synapse would provide the mechanism which dictates whether a specific response would be enacted or whether it would be blocked. In other words, the function controlled by neurons in a very circumscribed part of the brain would be altered by both the enhanced presynaptic release of one humoral factor and by the inhibited release presynaptically of its opposing counterpart. How would this actually operate? An incoming signal to the brain relayed by some sort of physiological imbalance first would be sensed locally in that region. Then the signal would trigger the release of one transmitter factor (e.g., ACh) and retard the release of its functionally opposite transmitter (e.g., norepinephrine).

The principle of a neurochemical code can explain theoretically how a collection of one set of neurons in a specific part of the brain is capable of mediating excitation or inhibition of a physiological response or a behavioral action. For example, two functionally opposing chemical substances determine the on-off nature of separate excitatory-inhibitory sets of neurons or pathways.

Here it may be helpful to give a few examples of the neurochemical dualism underlying the coding process in the brain of the cat, monkey, or other animal. If the endogenous substances found in the brain are artificially applied by injection to the region where they are stored, one substance may stimulate a specific response, whereas another can counter the response. These examples are taken from the *Handbook of Drug and Chemical Stimulation of the Brain* [2] as follows (1) Serotonin injected in the hypothalamus increases local blood flow, but norepinephrine reduces blood flow when applied similarly (Ch. 3). (2) The hormone, progesterone, deposited in the basal hypothalamus suppresses the synthesis of progestin, but estrogen applied at the same locus facilitates its synthesis (Ch. 5). (3) Serotonin elevates body temperature when infused into the forward or anterior part of the hypothalamus, whereas a norepinephrine infusion lowers temperature (Ch. 6). (4) Norepinephrine injected in the anterior hypothalamus evokes feeding, whereas a peptide hormone, angiotensin, reduces eating behavior (Ch. 7). (5) Dopamine injected into a structure involved in motor activity, the caudate nucleus, antagonizes the intense tremor evoked by ACh applied at the same locus (Ch. 10). (6) ACh injected into the outer edge of the hypothalamus causes a rat to kill its prey, whereas norepinephrine given at the same site suppresses killing (Ch. 11). Some of these examples will be elaborated upon in succeeding sections.

Ways to Examine the Neurochemical "Code"

Ingenious procedures have been developed in laboratories scattered throughout the world for studying the local chemical activity of neurons within specific structures of the brain. One straightforward method involves the post mortem dissection of the brain into its component parts. Thereafter, each part is examined by analytical chemical procedures, including spectrofluorometric or chromatographic ones. The disadvantage of this is twofold: the animal must be killed for the analysis so that it cannot serve as its own control and the anatomical separation of the parts is often beleaguered by imprecise dissection because of the minuteness of the structures.

Two other methods are used painlessly in the live animal. First, an endogenous transmitter or humoral substance is microinjected directly into a particular structure of the brain [2]. Although a traditional approach uses the cerebrospinal fluid as the route of injection, this anatomical alternative of microinjection into the brain substance in a specific region is singularly advantageous. If careful, the scientist can mimic the action of the endogenous compound and sometimes can characterize the features of the postsynaptic receptor sites. Above all, the effect that one observes can be localized anatomically.

A second way of examining the coding process is only now coming into practice as a useful physiological tool. The procedure involves the localized perfusion of an area comprised of chemically distinct neurons [2]. As the fluid washes the site, a transmitter or other substance that is released locally is collected in the perfusate. Changes in endogenous activity that occur in the region as the result of a given stimulus can be detected by the analysis of the samples of perfusate. If the release of one substance is enhanced at the same time that its opponent is inhibited, the existence of a specified code may be postulated. It is marvelously encouraging when the changes in release of the compounds correlate identically with their pharmacological actions upon microinjection. Then the evidence for a functionally specific chemical coding becomes firm.

The apparatus systems whereby a chemical is delivered to a local region or a perfusion of that region is undertaken are relatively sophisticated. They entail the implantation of a very fine needle by means of stereotaxic surgery. After the animal recovers from surgery, a solution is delivered in an exceedingly small volume through miniature catheters. Naturally, many controls are required for this type of delicate experimentation; final histological studies reveal the locus of the needle implant. In the following sections of this article,

the delineation of the neurochemical processes that underlie cerebral functions is based in large measure on these two methods.

BRAIN'S EXECUTIVE ACTION: CURRENT EVIDENCE FOR NEUROCHEMICAL CONTROL.

As alluded previously, many of our vital functions are thought to be under the neurochemical control executed by the brain. Here, in the following sections, we shall see how a transmitter code can operate decisively to provide the most subtle of finely balanced reactions in the nervous system. These responses enable us to survive constant environmental challenges, such as temperature, as well as internal challenges such as water deficit, sleep, and hunger.

Hunger and Feeding

The concept of a neurohumoral "code" applied to the act of feeding combines both physiological and behavioral events. Input to the brain in the form of blood-borne nutrients including carbohydrates and lipids reflects an excess, deficit, or balance in their respective levels [3]. One speculation is that the individual balance in nutrient titres constitutes the physiological signal that impinges directly upon neurons in the hypothalamus responsible for eating and how much and what kind of food is consumed. Today, many scientists believe that the pathway that the nutrients key in upon is a noradrenergic (norepinephrine-containing) system of nerves. As the noradrenergic synapses are activated in a specific portion of the hypothalamus, intense feeding is caused.

In the same context, once the condition of satiety is achieved, then a functionally opposing substance should be released to inhibit these neurons. By this we mean that a satiety signal should also be coded but in opposition to the noradrenergic neurons. Up until now, however, no endogenous substance present in nerve endings has been discovered which serves to inhibit feeding consistently. Possibly, ACh at certain sites in the hypothalamus represents the most likely candidate for a satiety transmitter.

Noradrenergic Feeding System—As documented now, relatively strong evidence is gathering to support the idea that norepinephrine mediates eating behavior [3]. First, dopamine and norepinephrine nerve endings are located within subunits (nuclei) of the hypothalamus which through classical experiments are implicated in hunger and satiety mechanisms. Second, when norepinephrine is applied to these hypothalamic regions, at least in the rat or monkey, the animals eat food voraciously even though they are already fed. Third, specific pharmacological antagonists of the norepinephrine receptors injected at the same loci attenuate if not entirely abolish the animal's appetite. Fourth, a lesion produced by a chemical neurotoxin (6-hydroxydopamine) injected into the hypothalamus or by a knife cut placed along the brain-stem pathway (which depletes both dopamine and norepinephrine) causes a disastrous effect on the regulation of food intake. A lesioned rat, for example, usually loses its appetite completely, and its body weight declines precipitously. Unless the rat is offered a really palatable food, a chocolate biscuit or the noodles from chicken soup, it starves itself to death.

Fifth, when a rat is deprived of food for a given amount of time, the content of norepinephrine in its hypothalamus decreases. A similar sort of fast also affects the new synthesis of both dopamine and norepinephrine in the neurons of the hypothalamus. Both observations indicate an active process taking place within the hypothalamus which is directly correlated with nutrient activity.

Norepinephrine Release—In the 1960s we found that a norepinephrinelike substance was released from the hypothalamus of the hungry monkey. This finding was verified later in the rat with even more clear-cut results. As shown in Figure 2, radio-labeled (^{14}C) norepinephrine, which is applied to the hypothalamus as a tracer of norepinephrine activity before the experiment begins, washes out of a localized perfusion site in the hypothalamus over time. However, the moment that the hungry animal begins to eat its special rat pellets (0 perfusion) norepinephrine activity (DPM) increases dramatically at this site (A). This effux is shown in the left panel of Figure 2. Incredible anatomical specificity is revealed in the right panel of Figure 2. When the perfusion

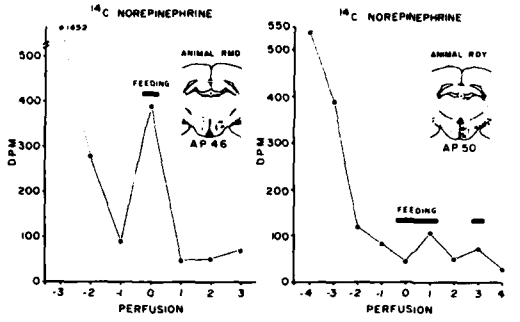


Figure 2—Changes in ^{14}C norepinephrine (NE) activity in DPM in push-pull perfusion fluid collected at a rate of 20 µl/min from a site in the ventromedial (left) and dorsomedial (right) hypothalamic areas of the rat. The sites had been labeled by a microinjection of ^{14}C-NE 1 hr before the first perfusion. The points on these ^{14}C washout curves were obtained at 30 min perfusion intervals until 0 perfusion when food was offered to the rat [4].

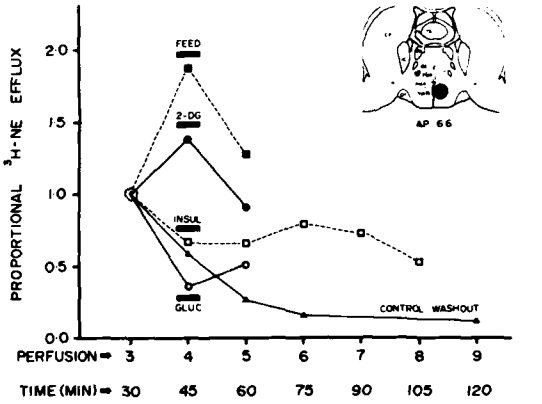


Figure 3—Changes in norepinephrine release ^{3}H-NE efflux from the rat's hypothalamus during four different experiments. During perfusion 4, at 45 min time, either 2-DG, insulin (INSUL), or glucose (GLUC) was added to the fluid which perfused the hypothalamic site (dot in inset). Or the animal was given and ate food (FEED). The site was always labeled with ^{3}H-NE injected in a 1 µl volume, 30 min before each experiment began. [5].

site is located only 2 mm away (▲) in the hypothalamus, hardly any change in norepinephrine activity occurs during the course of the feeding intervals. The upshot of this is that noradrenergic neurons at a circumscribed locus increase their activity as soon as food is available and eating commences [4].

A major question is the actual trigger that stimulates norepinephrine release. Just recently, we found that the norepinephrine activity in a hypothalamic feeding site is exceptionally sensitive not only to the local excess or deficiency in glucose but also to the presence of insulin [5]. Figure 3 illustrates an identical type of experiment as that depicted in Figure 2, but there is one difference. In the midst of the washout curve of norepinephrine activity (^{3}H-NE efflux), glucose, insulin, or 2-DG (a compound that depletes glucose locally) is added to the perfusion fluid (Perfusion #4 at 45 min time). The norepinephrine activity measured at the site reveals quite clearly that glucose suppresses noradrenergic activity. At the same site of perfusion, its competitor (2-DG) has the same net effect as hunger-induced feeding, as it depletes locally the glucose stores: increase in norepinephrine release. On the other hand, insulin also evokes norepinephrine release but it has a delayed action on the noradrenergic system. This corresponds with its known metabolic action, that of inducing hunger when taken systemically.

Undoubtedly, future research in this field will be devoted to the issue of how lipids, amino acids, carbohydrates, and other substances related to one's nutritional state affect the brain's neural control mechanisms. We still do not understand today how the so-called "set-point" for body weight is established. For example, why do some individuals maintain a slim figure with ease while others lapse into static obesity despite considerable dietary efforts to the contrary. Currently, obesity is a major health hazard worldwide, and the all-too-common lack of willful control over food indulgence is equally puzzling. The answer to the clinical treatment of the obese patient would seem to lie in continued research on these basic mechanisms for the neurochemical control of feeding.

Thirst

An exceptional balance exists between the peripheral and central processes responsible for the regulation of body water and maintenance of salt balance. Receptors in the mouth, stomach, and other tissues monitor the condition which engenders thirst or the craving for water. Although dehydration can be brought about in sev-

eral ways, the forward or anterior part of an animal's (and presumably human's hypothalamus) possesses specialized detectors that monitor osmotic and volumetric changes in the blood.

One type of chemical signal that impinges upon neurons of the anterior hypothalamus is sodium. In excess of its normal concentration in the bloodstream, sodium causes an osmotic disturbance, compensatory thirst, the resultant search for and drinking of water. The other signal is in the form of a hormone manufactured by means of a kidney principle, called angiotensin II. A primary purpose of angiotensin II is to constrict the blood vessels which thereby sustains normal blood pressure. This vital hormone has several other important actions, one of which is to cause water-seeking behavior. When it is applied locally to receptors in the hypothalamus, the animal drinks copiously.

The neurohumoral "code" proposed to be involved in the restitution of a water deficit is subserved by a cholinergic system. Evidence for a cholinergic thirst system in the hypothalamus has been accumulated over the last 30 years by endocrinologists, psychologists, and physiologists [2]. Illustrative are experiments on the local application of ACh to the anterior hypothalamus. Here, ACh causes secretion of a pituitary hormone that prevents the kidney from losing water in the form of urine, and thus body water is conserved effectively. ACh and other drugs that act on cholinergic receptors when applied to the hypothalamus and at points throughout a very large "circuit" of neurons in the brain also cause a rat to spontaneously drink water. This intake of fluid occurs even though the animal is in a perfect state of water balance and is not ostensibly thirsty. Literally hundreds of experiments have been undertaken to demonstrate the existence of a cholinergic thirst circuit. There are some convincing pharmacological studies with substances that block the cholinergic receptors along the thirst circuit of neurons.

A striking illustration of the neurohumoral code for both of the ingestive behaviors, eating and drinking, is presented in Figure 4. When norepinephrine or its chemically close analog (epinephrine and dopamine) are applied to the hypothalamus of the rat, food intake is evoked (Figure 4, left). Conversely, when ACh or its

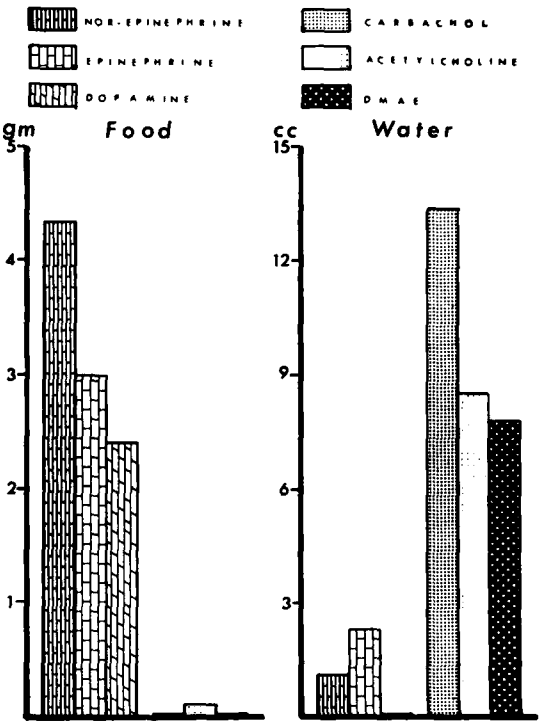


Figure 4—Effects of adrenergic and cholinergic stimulation of the hypothalamus on food and water intake of sated animals during a 1 hr poststimulation period [6].

analogs (carbachol and DMAE) are applied at the identical site in the rat's hypothalamus, water is taken in remarkably large volumes. In both circumstances, the animal is fully satiated with both food and water before the experiment is started [6].

Whether other chemical factors in the hypothalamus modulate the drinking of water remains to be determined. The scientific researcher in the future will most likely investigate the commonality of transmitter factors that mediate drinking and certain other functions. An example is the condition of overheating and heat stress, which, because of abundant perspiration, deplete body water. Naturally, the drinking of water ensues not only because of an almost immediate effect of body cooling but also because the loss of water due to perspiration or in some animals from the airways during rapid respiration is rectified. How the "codes" for these functions coalesce is an intriguing question.

Sexual Behavior

Among the most extraordinary features of the brainstem of an animal is the special sensitivity to certain hormones. The finding that several areas of the hypothalamus have an affinity for the female hormone estrogen, with respect to its accumulation and binding to neuronal elements, is truly significant. It has led to the speculation that sex steroids circulating in the bloodstream exert a direct influence on the activity of neurons in the central nervous system. As reviewed earlier [2], estrogen crystals deposited by fine needle in the hypothalamus alter the growth rate and size of the reproductive organs. The consequent secretion of sex steroids and the anatomical characteristics of the pituitary gland are correspondingly affected. In the cat and other animals, the local deposition of tiny pellets of synthetic estrogen even causes persistent copulatory behavior in the female despite an earlier ovariectomy.

Quite extraordinarily, an implant of estrogen crystals in the hypothalamus of the female monkey, as shown by Michael [7], drastically improves the sexual performance of her male partner. His aggressiveness and typical threatlike behavior always seen in the wild primate is equally affected by the estrogen implant in the female. Particularly interesting is the fact that the female hormone also acts on the hypothalamus of the male animal to inhibit copulatory activity of the male. Again, much research must be done towards the thorough characterization of the various types of responses, behavioral and endocrinological, that are able to be elicited by the direct action of a sex hormone on the brain.

New findings suggest that a neurotransmitter is an intermediary in the hormone's effect on nerve tissue. Some observations show that dopamine and norepinephrine pathways are activated (or deactivated as the titre of hormone circulating in plasma rises or falls. For example, the hypothalamic application of either of these two neurohumors influences the secretion from the pituitary gland of the trophic principle, luteinizing hormone. Morever, norepinephrine, but not dopamine, also shifts the period of ovulation in the rat. That a neurochemical "code," as yet unspecified, functions in the endocrine control process is a reasonable possibility.

Although many questions arise almost daily in this field, the transmitter pathways that are probably involved in hormone secretion are now in the process of being traced by Swedish and other histologists. Yet to be done is the precise anatomical localization of an effect of a neurotransmitter on a glandular process. Another crucial question revolves about how a neurotransmitter is released from a pool of neurons by a gonadal hormone.

Response to Hormonal Stress

Physiological stressors such as cold, pain, hemorrhage, pressure, anoxia, and abnormally rapid movement are perceived immediately by systems in the brain. As such, these stressors exert a powerful impact on the nervous system, whose response is translated into a full-blown reaction by the adrenal gland. In recent years, neuroendocrinologists have suspected that the hypothalamus and other structures in the central nervous system exert a direct influence on the adrenal gland. Once stimulated, the outer layer of the adrenal gland secretes hormones to combat the stressful situation and overcome the resultant deleterious effects.

The neurochemical "code" in the hypothalamus that facilitates the adrenal response is not solved as yet. Nevertheless, ACh release from cholinergic neurons within the hypothalamus probably elevates the level of circulating adrenal steroids by way of stimulating the output of corticotrophin releasing factor (CRF). This factor in turn stimulates the production of adrenocorticotrophic hormone (ACTH). The latter is a trophic hormone secreted from the pituitary gland which rests just beneath the hypothalamus. Thus, the final common pathway for the adrenal response may indeed be cholinergic, because cholinergic antagonists inhibit the *in vivo* production of adrenal steroids by the adrenal gland. This pharmacological blockade predominates after anticholinergic drugs are administered in spite of the variety of noxious stressors to which an animal is exposed [8].

The Soviet endocrinologist Naumenko [9] and his colleagues have studied extensively the hypothalamic role played by serotonin in the secretion of adrenal steroids. Essentially, they find

that serotonin activates the releasing factor (CRF) from the hypothalamus. This in turn stimulates the production and liberation of the pituitary hormone, ACTH. It is Naumenko's contention that the activity of serotonin in several structures in the brain, comprising an emotional "circuit" of neurons, can stimulate the pituitary axis directly. If the endogenous release of serotonin could be demonstrated, which would reflect an enhanced local activity of serotonin, this concept would be strengthened.

Figure 5 (top) illustrates the potent local effect of serotonin on the basal brain of the guinea pig. Even though the brain has been transected (denoted by the solid black line) so as to block the outflow of nerve impulses to the body, an injection of 5-HT into the implanted tube nevertheless evokes the output of the adrenal corticosteroid (17-OHCS) (bottom). The direct action of serotonin as a neurochemical transducer for the brain (CRF)—pituitary (ACTH)—adrenal (17-OHCS) pathway is astoundingly documented here.

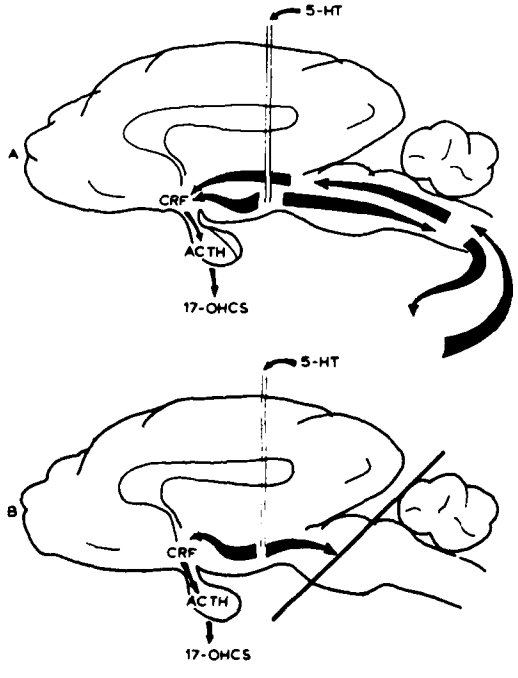


Figure 5—The influence of serotonin, injected locally into the brain, on the hypothalamic-pituitary-adrenal system. The effect is stimulatory not only when efferent neurons are intact (A) but also after the blocking of descending nervous pathways by transection (B). In the latter case the effective mechanism must involve ascending pathways [9].

Emotion and Aggression

In harmony with the hypothalamus, which is strategically located at the very base of the brain, several structures beneath the surface of the cortex form an integrated anatomical "circuit" by virtue of rich connections of nerve fibers. These include the amygdala, septum, midbrain, hippocampus, and the thalamus. This circuit, delineated by Papez in the 1930's, forms the anatomical basis for the expression of our emotions.

Both Soviet and American research workers have found that ACh, if infused into certain portions of any of these aforementioned structures, induces a variety of striking changes in the emotional behavior of an animal. Depending on the site of direct injection, for example, ACh will provoke fear and escape behavior as well as a syndrome that is likened to human rage. Eventually an attack on an animate or inanimate object occurs.

Within minutes after the application of ACh to the hypothalamus of a rat, the animal suddenly attacks viciously either a mouse or frog placed in its cage. The rat then may kill its prey swiftly, even though under normal circumstances it is known to be a nonkiller from earlier tests [10]. The specificity of a cholinergic mechanism subserving this sort of killing and other aggressive action has been well documented. Pharmacological antagonists that block cholinergic receptors in the circuit prevent an ACh-elicited emotional outburst.

Predatory attack does not necessarily reveal other easily recognized components of emotional behavior. Killing can occur without any demonstrable signs of emotional turmoil. Yet the outward and measurable expressions of emotion also are served by a cholinergic circuitry. For example, when ACh is injected locally into the hypothalamus of the cat, a startling array of emotional symptoms is generated. The cat's heart and respiratory rates are markedly elevated; its fur bristles; the pupils are dilated; salivation, baring of teeth, and loud growling occur; and the animal adopts a crouching stance as if preparing to pounce or attack. At the same site, the local application of a substance such as norepinephrine often has the opposite effect. The animal becomes placid, docile, sedate, or calm in contrast to its

otherwise normally active state. Figure 6 illustrates the two opposing emotional responses caused by ACh and norepinephrine injected independently on different days but at the same hypothalamic locus.

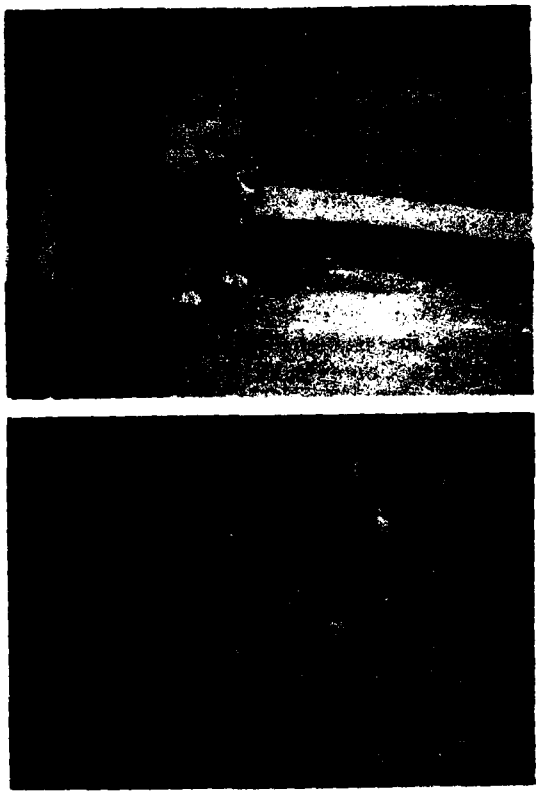


Figure 6—Top: Cholinergic stimulation (10 μg carbachol) of lateral hypothalamus. Note pupillary dilation, piloerection, spitting, and a fear-like withdrawal from a piece of tubing. Pronounced hissing and spitting accompanied the eventual attack of the tubing. Bottom: Adrenergic stimulation at the same locus (10 μg epinephrine) results in proneness, pupillary constriction, and absence of emotional behavior. A sleep-like state persisted for nearly 1 hr following the 1 μl injection of the drug. [11]

More research must be done to delve into the difficult question of how aberrant patterns of emotional behavior can be therapeutically subdued. The precise neurochemical effects that result from treatment with antidepressant type drugs and tranquilizing agents are still unknown. The anatomical locus of action of these drugs is equally perplexing. What is required, of course, are studies of the changes in endogenous activity of the transmitter and other neurohumoral factors both during periods of emotional crisis and following the efficacious therapy with tranquilizer and other psychoactive drugs.

Sleep

Although scientific controversy surrounds the actual purpose of sleep, most physiologists agree that sleep is a restorative process beneficial to all organ systems of the body. The recuperative process itself has slowness as its hallmark, for fully one-third of our entire life (on the average) is occupied by the state of sleep. Additional disagreement centers on how we enter into the condition of somnolence. Major questions are still to be answered by future research workers: Is there a blood-borne factor that gives rise to the state of sleep? Or do independent hypnogenic substances within the brain accumulate to signal "sleepiness" to the appropriate neurons? How is arousal triggered after an 8-hour period of sleep has ensued?

In the mid-portion of the brain, a major anatomical substrate devoted to the sleep mechanism has been uncovered. According to the French scientist Jouvet and other workers, an imbalance in the release of neurohumoral factors within this region is responsible for the onset of sleep. One key factor is serotonin. Following injections of a drug that depletes serotonin stores in the brain of the laboratory animal, a pronounced insomnia develops. When applied at certain midbrain sites, serotonin also causes drowsiness, a sleeplike state, and changes in the electrical activity (EEG) of the animal's cerebral cortex, as recorded by superficial electrodes.

Two cholinergic pathways seem also to be involved both in the state of waking as well as in the induction of sleep. Since ACh is decidedly implicated in the maintenance of the electrical activity of the cortex, its role in maintaining behavioral arousal can easily be understood. When ACh is injected into selected sites in the hypothalamus and midbrain of the cat, either arousal or a sleep-like condition is produced. The status of norepinephrine and dopamine in the functions of sleep and waking is still not clear. Some inves-

tigators find that their local injection into brainstem loci causes arousal; other workers report that drowsiness and what appears to be a deep sleep are elicited by these two substances.

One difficulty is worth mentioning that faces scientists in the future who undertake sleep experiments in mammals. The chief problem that has always plagued the researcher is that any sort of experimental manipulation (e.g., switching on recording equipment) tends to disturb and awaken a sleeping animal. Thus, to detect the ongoing changes in release of transmitter substances during various stages of wakefulness, perfusion tubes will have to be positioned and samples taken in a way that entails a minimal physiological disturbance. Nevertheless, as telemetering, remote stimulation, and sensing devices become perfected and more widely adopted, the prospects for understanding the brain's internal neurochemical code of the sleep and arousal processes are enhanced.

BODY TEMPERATURE CONTROL— A SPECIAL CASE

> New urgencies are bringing man back to the sea from whence all life has sprung. Clearly, if man hopes to crack the mystery of his murky beginning, he must go back to Mother Sea for the final answers.... One day we may learn that the initial living cell was sparked by the heat from an undersea volcano rearranging the sea's rich ionic solution. Perhaps the great pressure of the deep was the catalyst in this vital chemical reaction. Few deny our ancestral link with the sea; our saline blood, the salty sweat on a man's brow, the gill slits in the human embryo, all recapitulate evolution and betray man's ocean genesis.
>
> J. Piccard and R. Dietz, 1961, *Seven Miles Down*

Temperature regulation is considered here as a special case. The main reason for this is that the processes governing body temperature give us a somewhat comprehensive picture or model of how a neurochemical "code" actually performs its task. In this case, two substances oppose each other functionally within the same hypothalamic area in terms of their distinctive mediation of heat gain and heat loss. A third substance carries the messages, derived from the balance mechanism of the first two substances, downstream from the hypothalamus. Finally, an ionic mechanism has been proposed for the set-point that holds our body temperature at approximately 37°C (e.g., 98.6°F) throughout life.

Conceptually, the body's set temperature of 37°C is defended assiduously against a great variety of incoming thermal challenges. These challenges are often severe and prolonged, as exemplified by a trek in the torrid jungle or across an open glacier. Sometimes they are sudden, as typified by the plunge of a Navy frogman into the frigid waters of the North Atlantic. The heat problem encountered within the confines of a ship's engineroom illustrates the practical side of the thermal stressor.

First, we will deal with the mechanism hypothesized to establish the set-point temperature. Second, current views on the regulation around this set temperature, as achieved neurochemically, will be presented.

Set-Point Temperature

Physiological constants are at the heart of the homeostatic process, which maintains internal physiological equilibrium by specific responses to changes that arise inside or outside of the body. Although it is easy to conceive that the temperature of 37°C depends solely on the rate of metabolism of tissues throughout the body, much experimental evidence indicates that a central process establishes a set-point. One example is the defined rise of one's body temperature during a fever. Here thermoregulation occurs to defend the new fever level no matter what sort of external cold or heat stimulus is applied.

The quotation taken from the memorable account of Piccard and Dietz contains a profound insight. The set-point mechanism seems to be an ionic one. This is not totally unexpected. Such a mechanism would have to be most fundamental, biologically speaking. Further, we know that the set-point temperature (1) is present at birth, (2) is

universal across all species of mammal, and (3) possesses the cardinal element of stability. Inherent in these characteristics is that which is fulfilled, for the most part, by the rich ionic nature of the extra-cellular milieu. In itself, this milieu is generally invariant.

Several years ago we discovered that sodium ions perfused in the hind portion (posterior) of the hypothalamus cause a runaway rise in the temperature of a cat or monkey. Calcium ions at the same site exert the opposite effect and produce a sharp fall in body temperature. Unlike all other chemical substances ever tested, sodium and calcium are the only ones that can drive an animal's temperature upwards or downwards to the brink of death. Furthermore, after a new set-point temperature is established with this hypothalamic disturbance of the ratio of sodium to calcium ions, the animal regulates its temperature perfectly well when it is exposed to heat or cold. Figure 7 portrays a typical sodium rise and calcium fall in temperature produced when the ions are perfused at sites (dots) in the hypothalamus. Note how the temperature change reverses once the perfusion of either of the ions stops. Two other important features of the set-point have been elucidated recently.

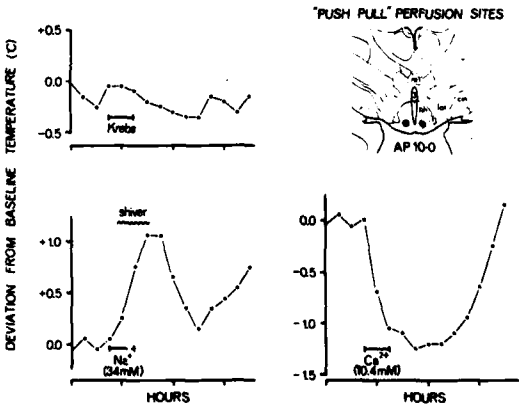


Figure 7—Changes in the colonic temperature of an unanesthetized cat in response to the local perfusion of the posterior hypothalamus for 30 min with a Krebs solution alone (upper left); Krebs solution plus 34 mM excess sodium (lower left); Krebs solution plus 10.4 mM excess calcium (lower right). The site of each bilateral perfusion is designated by the dots in the inset. Shivering occurred as indicated [12].

First, if bacteria (e.g., typhoid) are administered systemically, the animal develops an intense fever. Accompanying the onset of this fever is a sudden shift in the level of calcium ions within the posterior hypothalamus. Calcium ions leave this part of the hypothalamus probably through membrane unbinding or transport. And this is happening at the same site at which an artificial disturbance in the ion ratio, by perfusion of sodium, causes a rise in temperature identical to that following a bacterial insult.

Second, when the cage in which the animal lives is subjected to cooling or warming, the animal seems to actively defend its set-point by a quantitative shift in calcium activity, again within the posterior hypothalamus. Figure 8 illustrates the ionic response. Calcium is retained in the hypothalamus as the animal's environmental temperature is raised to 40°C. This retention

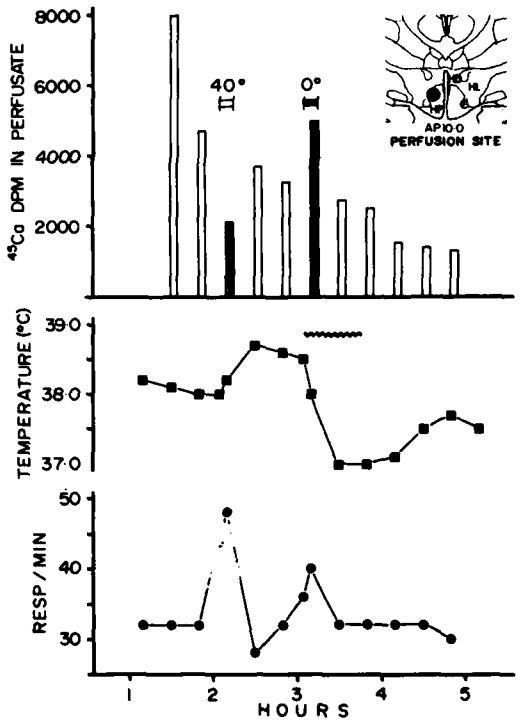


Figure 8—Efflux (top) of $^{45}Ca++$ in successive push-pull perfusates collected at a rate of 50 µl min from the perfusion site denoted by the dot in the histological inset. The site had been labeled with 1.0 µCi $^{45}Ca++$ 18 hr earlier. The chamber temperature of the cat was raised to 40°C or lowered to 0°C just preceding and during the third and sixth perfusions, repectively, as denoted by the bars. Colonic temperature (middle) and respiratory rate (bottom) were recorded continuously. Shivering is designated by the zigzag line (middle) [13].

corresponds to the heat loss evoked by calcium perfusion (Figure 7). Next, during the interval when the temperature of the animal's cage environment is lowered to 0°C. calcium ions are expelled from the hypothalamus (Figure 8). This expulsion of calcium ions would enable the sodium ions to predominate. Again this corresponds precisely with heat production which sodium evokes when it is perfused in the same locus.

Neuronal signals that impinge upon the posterior set-point region of ionic balance arise from the forward part (anterior) of the hypothalamus. This region contains thermally sensitive neurons which change their firing rate when they are heated or cooled. The peripheral pathways that relay the information about external temperature from the skin also terminate in the anterior region. As discussed in the next section, the thermoregulatory system is believed to be located here.

Thermoregulatory "Coding" Mechanism

Interlaced pathways of serotonin and norepinephrine-containing neurons ascend through the brain and terminate in the anterior hypothalamus. When serotonin is injected into this region, the temperature of the cat or monkey rises transiently. This observation led to the theory that serotonin is responsible for mediating heat production signals in this thermosensitive area [2]. Within the same site a perfusion carried out to collect locally released serotonin gives confirmatory data. Cooling of the animal's environment enhances the release of serotonin. The two results, pharmacological and physiological, taken together provide experimental evidence for serotonin's role in regulating against the cold [14].

When norepinephrine is injected into the anterior hypothalamus, the temperature of a cat or monkey falls. During the perfusion of this area at the same time that the environment is warmed, norepinephrine is released. The physiological and pharmacological concordance of these results supports the theory that norepinephrine is responsible for mediating the nerve impulses for the loss of body heat.

That the signals for heat production are carried by a network of cholinergic neurons has also been demonstrated by the same type of experiments. When ACh is microinjected at sites all along descending hypothalamic pathways, the temperature of the animal rises briefly. As the animal is cooled, ACh is released from the same sites at which the cholinergic compound causes heat production. Representative experiments that show the serotonin and norepinephrine effects on temperature and the transient action of ACh are illustrated in Figure 9. This graph portrays the

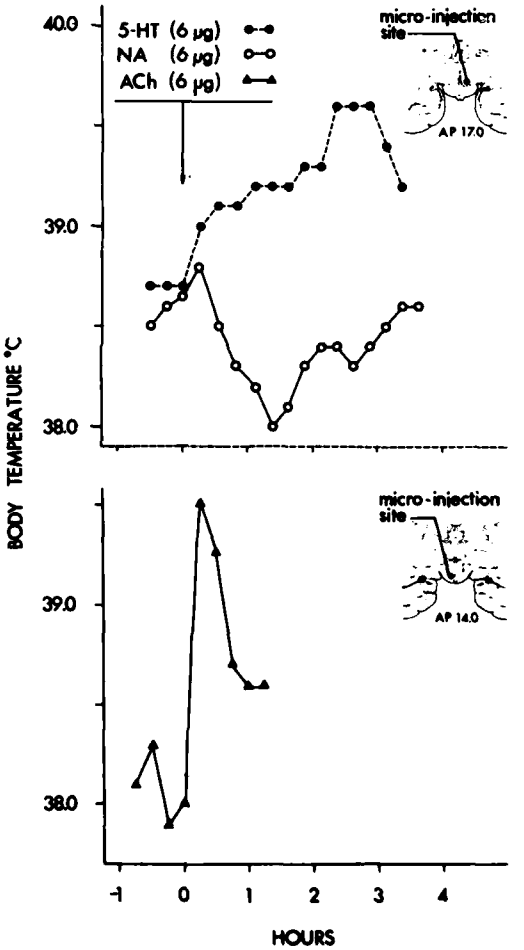


Figure 9—Top: Temperature responses of two monkeys following microinjections at 0 hr in the anterior hypothalamus at AP 17.0 (inset) of 6 µg 5-HT (•—•); of 6 µg norephinephrine (NA) (o—o) 5-HT and norepinephrine were given in the same animal and acetylcholine in the other. Bottom: Temperature response following microinjection at 0 hr in the posterior part of the ventromedial hypothalamus (•) at AP 14-0 (inset) of 6 µg acetylcholine-eserine mixture (ACh) (▲—▲) [15].

temperature responses produced by injections at specific loci involved in the temperature control mechanism.

A neurochemical "code" for thermoregulation seems to reside within the anterior hypothalamus. The way that this code could operate is by way of a finely tuned balance between the endogenous release of serotonin and norepinephrine [14]. Thus, as heat gain or heat loss is called for, the respective substance is liberated from its nerve endings, while the release of the other is attenuated. The anterior hypothalamic output, relayed by ACh to the posterior hypothalamus, operates in harmony with the ionic set-point mechanism. Overall, it now appears that incoming impulses that signal a displacement in body temperature are sensed by the serotonin-norepinephrine cells. They in turn telegraph the appropriate corrective response to the posterior hypothalamic neurons, which integrate all the output control signals. A shift in sodium-calcium balance here then serves to excite or inhibit mutually the firing of neurons that comprise the heat production and heat loss pathways.

Future researchers are left with the difficult issues pertaining to the kinetics of ion flux, the precise anatomical tracing of the fiber connections between the two temperature areas and their relation to other regions of the brain. How these processes function in the hibernator also is unknown. Can we eventually induce hibernation in a nonhibernating mammal such as man? The applications here would be as far reaching as the implications of such a possibility are exciting.

CONCLUDING REMARK

Although the functional description of the given "codes" in this article could lead to a conclusion that a code simply serves as a switchlike on-off process, this is undoubtedly an oversimplification. The activity of a transmitter represents more than an electrical-type switching mechanism. A bundle of chemically specific neurons can be damped down, enchanced, or modulated according to a recruitment gradient. Thus, a continuum of graded responses is achieved by a coded system. There is even a strong possibility that a third substance can modulate the actions of two opposing neuronal factors.

The complexity of each individual neurochemical code that is delegated to a particular function is acknowledged. But, in spite of this, there is great promise for the continual cracking of each code as research in this field continues. Indeed, the ultimate knowledge of how two, three, or more of the codes interact with one another will also be attained.

With this in mind, a quote is taken here from the concluding section of the *Handbook of Drug and Chemical Stimulation of the Brain*. This volume presents an account of thousands of experiments which contribute, each in its own way, toward the elucidation of different neurochemical coding processes. The "black box" referred to is a term used metaphorically by behaviorists to describe the brain. In their terminology, input to it can be defined (external stimuli) and the output from it can be quantitated (response measures).

> In all quarters, great strides are being taken continually by neurophysiologists, biochemists, electron microscopists, neuropharmacologists, and many others who are successfully sorting out the contents of this dark box. To be sure, the ultrastructure, the interconnections and the chemical dynamics persist in being bewilderingly complex. But immense gains in factual knowledge make it safe to conclude that the hue of the box can no longer be considered as black. Instead its overall color, in my opinion, has taken on a conceptual cast of gray.
>
> Only through the tremendous out-pouring of research from laboratories, large and small throughout the world, has the lightening of this box been achieved. As each one makes a distinguishing impact in one way or another, the tight lid of this now gray box is already wedged open, and by a formidable wedge at that.
>
> R. Myers, 1974

REFERENCES

1. H. McLennan, *Synaptic Transmission*, W. B. Saunders Company, Philadelphia, 1963, pp. 1–134.
2. R. D. Myers, *Handbook of Drug and Chemical Stimulation of the Brain*, Van Nostrand Reinhold Co., New York, 1974, pp. 1–760.
3. R. D. Myers, *Pharmacol. Biochem. Behav.* **3**, 75–83 (1975).
4. G. E. Martin and R. D. Myers, *Am. J. Physiol.* **229**, 1547–1555 (1975).
5. M. McCaleb and R. D. Myers, to be presented at the 6th Annual Meeting of the Society for Neuroscience, (November 1976, Toronto) and subsequently published in *Neuroscience Abstracts*.
6. S. P. Grossman, *Int. J. Nueropharmacol.* **3**, 45–58 (1964).
7. R. P. Michael, *Exp. Med. Int. Cong.* **184**, 302–309 (1968).
8. J. Kaplanski and P. G. Smelik, *Acta Endocrinol.* **73**, 691–699 (1973).
9. E. V. Naumenko, *Brain Res.* **11**, 1–10 (1968).
10. R. J. Bandler, *Nature* **224**, 1035–1036 (1969).
11. R. D. Myers, *Canad. J. Psychol.* **18**, 6–14 (1964).
12. R. D. Myers, in G. E. B. Wolstenholme and J. Birch, editors, *Ciba Foundation Symposium on Pyrogen and Fever*, Churchill, London, 1971, pp. 131–153.
13. R. D. Myers, C. W. Simpson, D. Higgins, R. A. Nattermann, J. C. Rice, P. Redgrave, and G. Metcalf, *Brain Res. Bull.* **1**, 301–327 (1976).
14. W. Feldberg and R. D. Myers, *J. Physiol.* **173**, 226–237 (1964).
15. R. D. Myers, in J. Barchas and E. Usdin, editors, *Serotonin and Behavior*, Academic Press, New York, 1973, pp. 292–302.



Enoch Callaway, M.D., is Professor in Residence and Chief of the Research Division, Department of Psychiatry, at the University of California, San Francisco, where he has been since 1958. His interest in relationships between brain function and human behavior has resulted in numerous papers and in the book *Brain Electrical Potentials and Individual Psychological Differences*. Dr. Callaway received A.B. and M.D. degrees from Columbia University and did postgraduate work at Grady Hospital (Emory University), at Worcester State Hospital and Worcester Biological Foundation, and at the University of Maryland. In the Navy, he served on active duty at the Army Chemical Center, U.S. Naval Hospital, Bethesda, and Naval Medical Research Institute.

ELECTRICAL "WINDOWS" ON THE MIND: APPLICATIONS FOR NEUROPHYSIOLOGICALLY DEFINED INDIVIDUAL DIFFERENCES

Enoch Callaway, M.D.

Langley Porter Neuropsychiatric Institute
San Francisco, Calif.

How can we pick the best job for a person and the best person for a job? Training programs further complicate matters by demanding three-way matches between person, training, and job. That traditional nest of problems can now be attacked in a new way by using measurements of human brain electrical potentials. In this paper we review some of these new techniques and consider the possibility for developing others.

This use of brain wave measures rests on two simple notions: (1) that the electrical activity recorded at the scalp can tell us things about the human brain that are hard to learn in other ways and (2) that an individual's brain ultimately plays the crucial role in the person-training-job interaction. Put another way, the sorts of things a person will do depend on the nature of his brain, and his brain will also determine the electrical signals we can record from his head. Thus, we come to suspect that a person's brain waves may tell us useful things about what we may expect from him in the way of behavior. Finally, at the basis of most studies of the mind lies the hope that if we understood more about how our minds worked, then we could make better use of them.

Of course we must be clear that classifying people neurophysiologically does not justify a false determinism. Ultimately, the best way to see what a person can do is to let him try. It is perverse when some theory or some statistical relationship is used as an excuse to set limits on a person's opportunities. Humans are at their very best when transcending apparent limitations. Blind reliance on statistics would exclude a stuttering Demosthenes from the debating society and an epileptic Caesar from military command. On the other hand, sometimes the cost of a training program or the consequence of a failure on a job may make the simple "try and see" approach unworkable. Then, after giving due and principal weight to the individual's own motivations and aspirations, some additional help in selecting for trainings and for jobs can be welcomed and helpful. Finally, there is the hope that some day we will recognize that individual differences represent one of mankind's greatest assets and we should capitalize on them instead of trying to erase them.

Human brain electrical activity was first described by Berger in the 1920s. Since that time there has been a fairly steady and consistent effort to find relationships between such electrical activity and individual psychological differences. For a long time there was very little to show for the effort, and that early failure is not hard to understand. The brain is generally doing a lot of things all at the same time. The electrical activity recorded from the surface of the head represents a jumble of underlying activity. In this gross mixture of electrical activity it is very hard to see

anything except gross changes in state. Now the raw EEG is sensitive to changes in state, and it has been very useful in studying such things as sleep, general level of arousal, epilepsy, states of intoxication, and death.

Real progress towards a more fine-grained window on the mind began when digital computers became generally available. The waking brain is busy at a variety of tasks, so if we want to gain insight into more specific cognitive activities we need some way of separating out more or less specific electrical activity from the background of other ongoing operations. It was the advent of the relatively low cost computer which made it practical to perform such separations on an everyday basis.

When confronted with the problem of picking something out of a random background the most obvious way of proceeding is by averaging. Thus, if one can get the brain to repeat a specific act several times and if the incidental electrical babblings of the brain are random with respect to the specific activity in question, then one can distinguish between the repeated specific electrical events and the other unrelated (random) events by averaging. The interfering events will cancel out and average to zero, thus leaving an accurate representation of the more or less consistently repeated event of interest. Other approaches have been developed, and advances in computer technology have made a practical reality out of what were once wild mathematical theories. We will return to some of the more exotic approaches later, but averaged brain electrical events (called averaged evoked potentials) provide a practical introduction to the area [1].

The easiest way to make the brain do the same thing repeatedly is to present a simple repeated stimulus such as a click or a flash. The results of averaging responses to such a simple stimulus are shown in Figure 1 which has been adapted from Picton et a. [2]. Both axes are nonlinear, for the early responses are of both low voltage and high frequency while the later responses are slow and large. Since the background activity is more or less the same throughout the averaging period, one must average many more trials to disclose the small, early responses than to show the large, late ones. In fact, with special care, some late responses can be studied in the EEG just as it comes

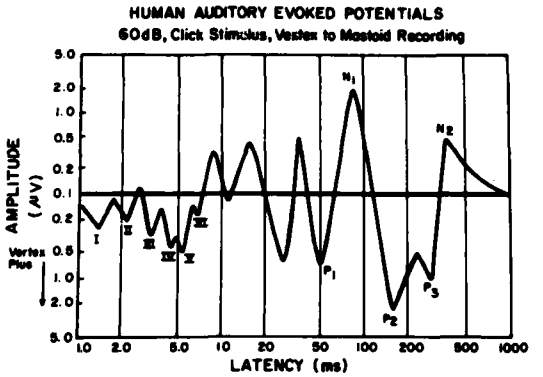


Figure 1–Schematic averaged evoked potential. Modified from Picton, et al. [2].

from the head, and such single-trial evoked potentials may be of value in monitoring the state of a person while a demanding task is being done.

The early responses reflect the passage of the signal along sensory pathways up to the primary sensory stations in the cortex. These early components allow averaged evoked potentials to be used in testing for specific sensory defects. In Figure 1, for example, we suspect that wave I is the cochlear microphonic, II is eighth nerve potential, III is cochlear nucleus activity, and IV is superior olivary nucleus.

By 100 ms (N1 in the figure) we find waves that have complex relationships to more subtle cognitive functions. The size of the N100 is related to a kind of primary sensory attention. For example, when clicks and flashes are mixed and the subject concentrates on the clicks, the click-evoked N100 will be larger than that to the flash. If the subject is uncertain about the stimulus, a new positive wave occurs between P2 and N2 at about 300 ms. Later waves, and in particular the P300, are sensitive to even more complex cognitive aspects of attention and become larger when more uncertainty is resolved by the stimulus. For example, if a string of regular clicks is presented with an occasional one omitted, the omitted click will evoke a response, and this emitted response will be larger in subjects that are counting these missing clicks.

The possibility that between-individual differences in brain electrical potentials might reflect interesting individual differences in psychological performance was given credibility by the within-individual findings such as those described previ-

ously. Some of the early work addressed itself to the gross individual differences found among psychiatric patients, and a variety of interesting correlations have been reported. For example, in normal subjects, N100 increases in size with stimulus intensity up to a point, but very strong stimuli may actually evoke smaller N100's than slightly less intense stimuli. By contrast, the N100 wave continues to increase in amplitude as the flash intensity increases over an unusually wide range in the manic patient, reflecting perhaps their stimulus-seeking propensities. P300 is smaller and more variable than normal in schizophrenics, reflecting perhaps the distracted and unstable cognitive states of these individuals.

At the same time this work on psychopathology was being done, evoked-potential workers were addressing individual differences in intelligence, that is to say, individual differences in certain types of performance tasks. The implicit goal was improved means of making individual-job-training matches. Now the issue of mental illness is not irrelevant in personnel selection, and research on brain potentials in mental illness is of interest in its own right. But we will for the present consider three evoked-potential measures that show correlations with IQ in normal subjects. Later we will show how these two streams converge.

The first correlation between IQ and averaged evoked potentials were obtained by measuring latencies (or delays) of waves evoked by light flashes. There is now a long history of controversy surrounding latency/IQ correlations, but, in general, it seems that bright people are likely to have short (fast) latencies. The correlations are not large, but the phenomenon is real enough to stimulate attempts at explanation. The first simple idea that fast (early) peaks in the averaged evoked potential meant a fast (smart) brain now seems unlikely. First, the correlation is low so that some very smart brains have slow peaks and vice versa. Next, the correlation is found only with flashes. Finally, the evoked potentials must be recorded from the side of the head with electrodes astride the motor area of the brain, not at the vertex. There is no apparent reason why quick responsiveness at the vertex should not be just as useful as quick responsiveness on the side of the head.

Several other theories have been advanced and now seem as equally unconvincing as the fast brain/fast mind idea. The author suspects that the observed IQ/latency correlations may reflect a relationship between the corticothalamic circuits and some as yet unidentified personality variable that is, in turn, weakly related to IQ.

The second evoked-potential correlate of IQ to be discussed is variability. Evoked potentials vary from trial to trial. This variability is increased in schizophrenia and is high in young children and the aged. It is lower in brightest subjects, and this shows up in comparisons between bright and dull age-matched military recruits. Like short latency, low variability as a correlate of IQ has a sort of face validity. Low variability evoked potentials suggest a good stable mind, but this simple analogy may be as misleading as the "fast evoked potential—fast mind" analogy.

At least in schizophrenics, who in general have lower IQ scores than would be expected from education, social background, and so forth, the increased evoked-potential variability is not evident throughout the entire evoked response but is found only in the later "cognitive" portion. In the early "sensory" portion, variability is actually lower among schizophrenics. One possibility is that the early evoked-potential variability reflects a kind of sensory preprocessing in which variability in brain state is compensated for by altering the early responses to the incoming data. In this way, a well-organized brain might impose variations on early responses in order to provide a more constant signal for later, more complex steps of processing. Such a well-organized brain would show more early variability and less late variability than would a more poorly organized brain. By contrast, a less efficient brain might respond more regularly to the immediate impact of the stimulus, but, having failed to adjust this initial response, later evoked activity might be more irregular.

Finally, there are IQ/evoked potential correlations that seem related to differential responsiveness of various cortical areas. Some years ago the brain was generally looked upon as a largely homogeneous organ where almost any part could do almost any job. The fantastic plasticity of the brain is still recognized, but specialization of the cortical areas has become of such great interest

lately that it is often jokingly referred to as the new phrenology. For example, we now believe that for most people the left hemisphere (which controls the right or dominant side of the body) is concerned with sequential, logical, verbal operations that are called "propositional" while the right or nondominant hemisphere is concerend with holistic, intuitive operations which have been called "apositional." In general, the left visual-evoked response is smaller than the right in high-IQ subjects, and this may reflect the fact that when such verbally gifted subjects are watching a flashing light they are likely to have the left hemisphere employed thinking verbal thoughts and the left hemisphere, thus occupied, is less responsive to the light. On the other hand, brain damage can also produce asymmetry, and brain damage is likely to produce a low IQ. Although there are many studies on this topic and some controversy, it is not unfair to summarize by saying that a moderate degree of asymmetry is characteristic of the bright, well-functioning individuals. An absence of asymmetry or an excessive asymmetry may be found more often among people who are not functioning so well.

Now to summarize what I have said about correlations between evoked potentials and intelligence: the correlations are real, they are low, and, in general, we can characterize bright, well-functioning individuals as being likely to have short latencies, low variability, and a moderate degree of asymmetry. There are considerable between-individual differences among the well-functioning group but even greater between-individual differences among the poorly functioning group. Imagine a space where each dimension of the space represents some evoked-potential measure and a point in that space represents a person. One will then find a loose group or cluster representing optimally functioning individuals, but that cluster will be entirely surrounded by other clusters representing different varieties of dysfunctioning. Some of these clusters might represent specific diagnoses of mental illness. Others might indicate normal variants who are more or less gifted for a particular task.

The process of looking for correlations between intelligence and evoked potential measures has now served its purpose by indicating that there are statistically significant predictors of human performance in the evoked-potential measures. The possibility now presents itself that the evoked-potential measures may be able to define clusters of individuals, and these clusters may provide more meaningful differentiations than conventional psychological categories and diagnoses [3].

The problem of IQ is a case in point. Poor performance on an IQ test may represent cultural disadvantage, normal variation on a continuum, or any of a variety of specific dysfunctions which might include dyslexia (the congenital disability in reading), schizophrenia, or even some temporary toxic state. The issue can perhaps be framed in another way. There is generally only one way to perform a task optimally. There are an endless number of ways to perform it poorly. Most conventional psychological measures simply determine whether the individual has performed well or poorly. The various ways of performing poorly are usually (though not always) lumped together in performance scores. There is a possibility that other techniques such as brain electrical potential measures may be able to distinguish different causes of poor performance. Now we perhaps should shift our methods, and, instead of looking for correlations between the evoked-potential measures and conventional psychological tests, we should instead cluster individuals on the basis of evoked-potential typology and then look for common characteristics among individuals who fall into a particular cluster.

Our work to date suggests we should abandon simple IQ/brain potential correlations for more elaborate cluster and analytic approaches. On the other hand, study of well-defined pathological groups can provide insights of a more theoretical nature, and these can supplement the atheoretical or brute force statistical work represented by cluster analysis. So now we return to the consideration of psychopathology to see how brain potential correlates of psychopathology suggest something about the brain mechanisms that might relate to performance and brain potentials in normals.

We have already noted how variability of late components in the evoked potential is a characteristic of schizophrenics and how subsequent research suggested this late variability may reflect a failure in early, preattentive processing of sensory data. There is considerable indication that a pro-

pensity for schizophrenia may not be maladaptive if coupled with a high level of ability, particularly in a truly threatening environment. Jarvic and Chadwick have coined the term "Odyssean personality" in discussing this idea [4]. It was Odysseus' habitual suspiciousness and almost pathological vigilance that finally brought him home to Penelope. A tendency to clean up sensory data before processing might lead to stable performance when careful attention to a limited set of data is desirable. But this might be dangerous in a totally unpredictable environment, and a talent for filtering incoming data might lead one to filter out a cue that was essential to survival. Questions for the future could involve identification of training strategies and job assignments that might capitalize on the way an individual's brain is likely to filter and organize incoming data.

In the early days of evoked potential response work, the special purpose averagers could not measure variability, and such measures had to be done off-line on larger computers. Multivariate statistics such as factor and cluster analysis were even more formidable, demanding the full force of a major computer center. Progress in minicomputers has changed all that. Variability is computed on-line by the most minimal of the general purpose machines, and complex multivariate analysis is usually carried out in-house.

This same increase in computer accessibility has stimulated a variety of other approaches to the analysis of brain electrical activity. The variety of approaches from which one could select an example include spectral analysis, coherence, discriminate function analysis, basis function analysis, and so forth, but we will take one developed in the author's own laboratory.

We have already remarked on the new interest in cortical specialization and how a moderate, averaged evoked-potential asymmetry is associated with good IQ scores. This evoked-response asymmetry is suspected of reflecting a difference in the way the two hemispheres are operating. We wanted to pursue the study of differences in the ways the cortical hemispheres operate, but we also wanted to study psychological operations that are more complex and continuous than response to light flash. It occurred to us that the level of communication between cortical areas might be reflected in correspondences between patterns of brain electrical activity picked up over these areas. We developed a method that depended on "decoding" the brain waves from two areas for a certain period of time and then determining the "information transmission" between the two areas during that period. The method turned out to be extremely efficient and has disclosed some interesting things [5, 6].

Before discussing some results with this method, however, a few words of warning for the more technical readers are in order. The terms decoding and information transmission are mathematically correct from the standpoint of information theory, but that is no guarantee that what we decode is related to what the brain considers to be a message. By the same token, what we measure as information transmission is only the contingency between the two artificial codes and may have no relationship to information transmitted back and forth between two brain areas. That warning is essential, for the author is biased. Although having no proof, the author suspects that our decoding scheme is related to the one actually used by neurons and that our information transmission measure does in some weak and inexact way reflect information transmitted between areas of the brain.

However, to avoid a too-literal interpretation of "information transmission" we will speak instead of electrical "coupling" between cortical areas.

Our first studies showed that different psychological operations resulted in different patterns of coupling, and the patterns were what one would expect from current concepts of cortical specialization. For example, when subjects read a book, then coupling from the occipital (visual) area to the left (propositional, language) hemisphere increased relative to the occiput-right coupling, and, when the subject looked at a picture without thinking of words, then there was a relative increase in coupling between occiput to right (apositional, spatial) hemisphere.

Other studies confirmed the impression that coupling did change as one would expect it to if indeed it reflected mutual involvement of brain areas in a data-processing operation. Now since our first steps showed that by varying the task we could show similar coupling effects among a group of subjects, we next decided to see if we could use coupling measures to detect

differences in the way individuals had their brains organized.

All the technical problems are not yet solved, but one experiment shows what we may hope for. It has been noted that evoked potentials to light flashes tend to have a greater coherence or similarity when they are recorded from two different areas over the right hemisphere. Evoked potentials to clicks, on the other hand, show greater coherence between the same two areas on the left hemisphere. This was thought to reflect the fact that sequential linguistic data were usually auditory, while spatial, holistic data are generally picked up visually. We set out first to see if a similar observation could be made using our coherence measure. We presented clicks and flashes at about 3/s and measured coherence between the frontal and parietal electrodes on the left and right sides of the head. As we suspected, flashes increased information transmission over the right hemisphere as relative to the left hemisphere, and clicks had the opposite effect.

It then occurred to Dr. Lekh Bali that dyslexic subjects might show a different pattern. He selected adults who had a long history of inability to read, for there is evidence in the literature to suggest that such people have defective cortical specialization. If the responses to clicks and flashes reflected differential cortical specialization, then one might expect that these dyslexic adults should show a different pattern of response than that of normals.

Figure 2 [7] shows the results of one experiment. One axis on the figure is labeled "left hemisphere," the other "right hemisphere." The scores on each axis are *t* scores. Let's take the left hemisphere as an example. We measure the coupling between brain waves recorded from the left parietal area and brain waves recorded from the left frontal area. The information transmission was measured once each second for a period of about 60 s while the light was flashing and for a period of about 60 s when a tone was sounding. Thus, we had a sample of transmissions during clicks and a sample of transmissions during tones. The numbers on the axis reflect the difference of coupling during clicks from coupling during flashes divided by a measure of variability. Such values we called *t* -scores. They are usually used in the *t* test which is a statistical procedure for

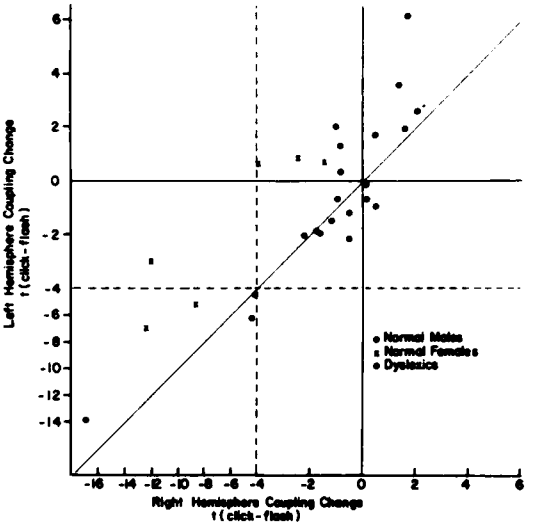


Figure 2–Cortical coupling responses to clicks and flashes. [7].

determining the significance of the difference between the two sets of data. For our purposes, the *t* score is more useful than the actual information transmission measure since it also reflects the variability.

A diagonal line is drawn on the figure, and all of the data would fall on the diagonal line if the click/flash difference were the same in both hemispheres. Since we expect clicks to produce a larger coupling in the left hemisphere and flashes in the right hemisphere in normals, we would expect all the normals to fall to the left of the diagonal line. This is the case in the figure. The dyslexic subjects, however, tend to fall very close to the diagonal line, and, in fact, all the dyslexic subjects are either closer to the diagonal line (showing less cortical specialization) than any of the normals or are to the right of the diagonal (suggesting reversed dominance).

One incidental finding is of some passing interest. You will notice that there is only one male subject who overlaps the females. In general, the males and females fall into quite separate clusters. This seems to reflect the fact that the females show greater relative shifts in the right hemisphere than do males. There is some evidence in the literature to suggest that females tend to rely more heavily on the left or language hemisphere than males, and, hence, may be showing less changes in their left hemisphere than males.

pensity for schizophrenia may not be maladaptive if coupled with a high level of ability, particularly in a truly threatening environment. Jarvic and Chadwick have coined the term "Odyssean personality" in discussing this idea [4]. It was Odysseus' habitual suspiciousness and almost pathological vigilance that finally brought him home to Penelope. A tendency to clean up sensory data before processing might lead to stable performance when careful attention to a limited set of data is desirable. But this might be dangerous in a totally unpredictable environment, and a talent for filtering incoming data might lead one to filter out a cue that was essential to survival. Questions for the future could involve identification of training strategies and job assignments that might capitalize on the way an individual's brain is likely to filter and organize incoming data.

In the early days of evoked potential response work, the special purpose averagers could not measure variability, and such measures had to be done off-line on larger computers. Multivariate statistics such as factor and cluster analysis were even more formidable, demanding the full force of a major computer center. Progress in minicomputers has changed all that. Variability is computed on-line by the most minimal of the general purpose machines, and complex multivariate analysis is usually carried out in-house.

This same increase in computer accessibility has stimulated a variety of other approaches to the analysis of brain electrical activity. The variety of approaches from which one could select an example include spectral analysis, coherence, discriminate function analysis, basis function analysis, and so forth, but we will take one developed in the author's own laboratory.

We have already remarked on the new interest in cortical specialization and how a moderate, averaged evoked-potential asymmetry is associated with good IQ scores. This evoked-response asymmetry is suspected of reflecting a difference in the way the two hemispheres are operating. We wanted to pursue the study of differences in the ways the cortical hemispheres operate, but we also wanted to study psychological operations that are more complex and continuous than response to light flash. It occurred to us that the level of communication between cortical areas might be reflected in correspondences between patterns of brain electrical activity picked up over these areas. We developed a method that depended on "decoding" the brain waves from two areas for a certain period of time and then determining the "information transmission" between the two areas during that period. The method turned out to be extremely efficient and has disclosed some interesting things [5, 6].

Before discussing some results with this method, however, a few words of warning for the more technical readers are in order. The terms decoding and information transmission are mathematically correct from the standpoint of information theory, but that is no guarantee that what we decode is related to what the brain considers to be a message. By the same token, what we measure as information transmission is only the contingency between the two artificial codes and may have no relationship to information transmitted back and forth between two brain areas. That warning is essential, for the author is biased. Although having no proof, the author suspects that our decoding scheme is related to the one actually used by neurons and that our information transmission measure does in some weak and inexact way reflect information transmitted between areas of the brain.

However, to avoid a too-literal interpretation of "information transmission" we will speak instead of electrical "coupling" between cortical areas.

Our first studies showed that different psychological operations resulted in different patterns of coupling, and the patterns were what one would expect from current concepts of cortical specialization. For example, when subjects read a book, then coupling from the occipital (visual) area to the left (propositional, language) hemisphere increased relative to the occiput-right coupling, and, when the subject looked at a picture without thinking of words, then there was a relative increase in coupling between occiput to right (apositional, spatial) hemisphere.

Other studies confirmed the impression that coupling did change as one would expect it to if indeed it reflected mutual involvement of brain areas in a data-processing operation. Now since our first steps showed that by varying the task we could show similar coupling effects among a group of subjects, we next decided to see if we could use coupling measures to detect

Further research will be required to test this point.

Our coupling procedure allowed us to make a perfect separation between the dyslexic and normal subjects. It must be noted, however, that these were severe adult dyslexics and represented an extremely homogeneous group. To apply it to the classification of a more heterogeneous group such as military recruits, one would perhaps do better to include it in a test battery that would then be submitted to factor and cluster analysis. It is also possible that cortical coupling measures could tell us other things about individual differences in cortical specialization. Approached from another point of view, coupling measures might tell us about the different demands various tasks and training procedures make on the human brain.

Now, let's bring together the threads of this paper, even though in the real world the threads are still somewhat loose. We have evidence that brain electrical potentials can tell us things about individual differences. The application of brain potential measures to small groups of pathological subjects resulted in information about our subjects and in information about our brain potential measures. The means are now at hand to combine all we have learned from pathological groups and from correlational studies into a test battery that may yield sensible clusters if data were gathered on a large enough sample of some group of interest, as for example, military recruits. Such performance and personality characteristics that are common to members of a cluster could then provide us with a powerful technical tool for prediction and, at the same time, starting with the psychological typology, we could work back through the cluster through the factor scores to pinpoint the neurophysiological phenomenon that sets a cluster apart and in this way add to our basic knowledge of the human mind.

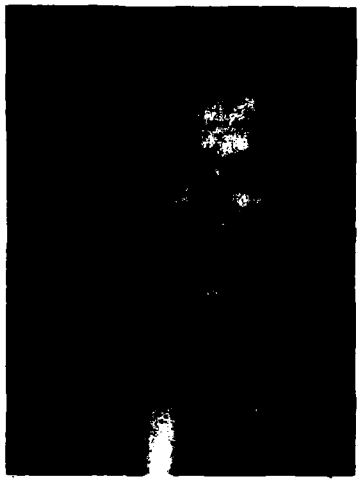
Much needs to be done. The relationship has barely been scratched. The notion that preattentive processing of sensory data can be studied via evoked-potential variables is scarcely a year old. The reasons for the correlation between visual evoked-potential latency and IQ remain a complete mystery. Yet practically useful classifications may (and probably will) be developed before a satisfactory theoretical basis is worked out.

All we have discussed illustrates again the blurred distinction between basic and applied research when we consider actual cases. Basic research is required to further the applied goal, and the applied work, in turn, supplies data pertinent to basic research questions. Meanwhile, after more than five decades of investigating brain electrical potentials and after some 15 years of computer analysis of these potentials, Hans Berger's dream that brain electrical potentials might provide a window on the mind seems about to come true.

REFERENCES

1. E. Callaway, *Brain Electrical Potentials and Individual Psychological Differences*, Grune & Stratton, New York, 1975.
2. T. Picton, S. Hillyard, H. Krausz, and R. Galambos, "Human Auditory Evoked Potentials, I. Evaluation of Components," *Electroenceph. Clin. Neurophysiol.* **36**, 179–190 (1974).
3. E. R. John, "How the Brain Works—A New Theory." *Psychology Today*, 48–52 (May 1976).
4. L. Jarvic, and S. Chadwick, "Schizophrenia and Survival," in M. Hammer, K. Salzinger, and S. Sutton, editors, *Psychopathology*, John Wiley & Sons, New York, 1973.
5. E. Callaway, and P. Harris, "Coupling Between Cortical Potentials From Different Areas," *Science* **183**, 873–875 (1974).
6. A. Yagi, L. Bali, and E. Callaway, "Optimum Parameters for the Measurement of Cortical Coupling," *Physiol. Psychol.* **4**, 33–38 (1976).
7. L. Bali et al., "Hemispheric Asymmetry in Normals and Dyslexics: Applications for a New Measure of Cortical Coupling" (in preparation).

PSYCHOLOGICAL SCIENCES



Alphonse Chapanis is Professor of Psychology at The Johns Hopkins University. Dr. Chapanis joined the Systems Research project there in 1946 and has been associated with the university ever since. He took a leave of absence in 1960–1961 to serve as liaison scientist in the Office of Naval Research Branch Office at the Embassy of the United States in London. Dr. Chapanis received a B.A. from the University of Connecticut and M.A. and Ph D. degrees from Yale. He is past President of the Society of Engineering Psychologists (1959–1960) and of the Human Factors Society (1963–1964) and was elected President of the International Ergonomics Association in 1976. In 1963 he received the Franklin V. Taylor Award of the Society of Engineering Psychologists and, in 1973, the Paul M. Fitts Award of the Human Factors Society.



Ederyn Williams is Director of Psychological Research for the Communications Studies Group, University College, London. Dr. Williams has been associated with this group since 1971. He received B.A. and M.A. degrees in Psychology (both with honors) from Cambridge University and a D. Phil. in Social Psychology from Oxford University.

HUMAN CONSIDERATIONS IN INTERACTIVE TELECOMMUNICATIONS

Alphonse Chapanis

Johns Hopkins University
Baltimore, Md.

and

Ederyn Williams*

University College
London, England

On July 31, 1971, and at various times during the following 2 days, millions of people settled comfortably in front of their television sets to watch two astronauts, David R. Scott and James B. Irwin, walk and drive around on the cold, bleak, airless surface of the Moon. From time to time two television cameras on the Moon, one near the landing capsule and one mounted on the lunar roving vehicle, moved and refocused in response to commands from Mission Control in Houston, Tex., nearly half a million kilometers away. Meanwhile, television viewers on Earth listened to verbal interchanges—jokes, bits of information, instructions, and questions and answers—between the astronauts and Mission Control in the Manned Space Flight Center [1].

Although these color telecasts were a spectacular and historic achievement, one of the most remarkable things about them was that they were accepted as an ordinary, everyday occurrence by the millions of people who witnessed them. Yet the technology that made these telecommunications possible was largely developed during the lifetimes of many people alive today. While the first commercial radio broadcasts were made from station KDKA as early as 1920 and the first commercial television broadcasts by the National Broadcasting System in 1939, most of the communications technology that we now accept so matter-of-factly has been developed only within the 30 or so years since World War II. Indeed, the rapidity with which technological developments have followed one another during the past few decades has been characterized as a "communication explosion" [2].

The systems that have been the end products of this technology have not all been success stories. Some, in fact, have been colossal failures, e.g., *Picturephone*®, and they have been failures because they did not meet human needs. At the same time, the explosion in communication systems has produced new problems, largely associated with the way people use, interact with, or respond to these systems [3]. This state of affairs has led Newsom [4] among others to conclude that advances in modern technology have surpassed our understanding of their human consequences and of the ways they need to be designed to match human needs, capacities, and limitations. In this paper we discuss some aspects of the communication explosion, concentrating on human considerations in the design and use of interactive telecommunication systems.

*This paper was prepared while Williams was a visiting research scientist in Chapanis' laboratory at The Johns Hopkins University.

SOME DEFINITIONS

Telecommunication means simply communication at a distance. It is usually used to refer to communication mediated electronically, such as by telegraphy, telephony, radio, or television. More broadly defined, however, the word also includes communication at a distance via nonelectronic means, for example, by whistle signals and semaphore. In this paper we use the more inclusive definition of the word. Although our interest is primarily in telecommunication, we shall also have a great deal to say about face-to-face communication because it is the standard against which the effectiveness of mechanically or electronically mediated communication is usually compared.

Interactive Communication

In communication research it is important to make a distinction between interactive and unidirectional communication. For years psychologists and other scientists have been concerned with the effectiveness of unidirectional modes of communication, such as highway signs, books, lectures, and television broadcasts. In unidirectional communication, the person to whom a message is addressed is a passive recipient of information. Nothing that the recipient does or says affects the communicator, the communication process, or the content of a message.

In interactive communication, by contrast, the participants are both senders and receivers of information. Communicators, the communication process, and the contents of messages can be and usually are affected by all the participants. Conferences, arguments, seminars, and telephone conversations are examples of interactive communication. Our paper is concerned entirely with such interactive telecommunication. We shall also use the term teleconferencing as a synonym for interactive telecommunicating even if only two people are involved.

The kinds of telecommunication we are concerned with are likewise characterized by their immediacy. Interactions can be or are made with the speed of electricity or very nearly so. For that reason, we deliberately exclude such slow and tedious forms of communication as the mails, even though in a certain sense they may be thought of as interactive.

Human Factors

Finally, we limit ourselves to the design and use of telecommunication systems from the standpoint of the users of those systems. The purely engineering or technical aspects of these systems are of interest to us only as they impinge on the system's effectiveness for human communication or as matters of general interest. For example, in discussing audio systems, it makes no difference to us whether the linkage between two telephones is a microwave beam, coaxial cable, or laser beam, provided that the intelligibility of the speech and interactive features of the communications are not affected. Most users are concerned only that their voices can be clearly transmitted from here to there. They are not concerned with how engineers make that happen. Neither are we.

AN OVERVIEW OF TELECOMMUNICATION MODES

In this section we describe briefly the major forms of person-to-person telecommunication systems, together with some of their principal advantages and limitations. Although we describe and identify some systems by name, our interest is not so much in particular communication systems as in the general characteristics of classes of systems. However, a difficulty with all psychological investigations involving equipment is that findings cannot be entirely divorced from the kind of hardware or apparatus that is used to produce the stimuli or vary the independent variables. The only way to extract generality from that kind of situation is to study a large number of different equipments that have certain psychological features in common and to concentrate on the human performance that is associated with or is a consequence of those general features. So, although audio systems come in hundreds of different variations, their common psychological characteristic is that they have an audio channel that allows

people to talk to but not see one another. If being able to talk to someone without seeing that person is always associated with certain difficulties or certain kinds of performance, irrespective of the particular kind of electronic or mechanical linkage involved, we have the beginnings of a valid generalization. With that in mind, we find, from a human factors point of view, four major forms of telecommunication systems.

Audio Systems

Audio telecommunication systems are among the most familiar since they include the ubiquitous telephone. This class of systems, however, also includes a number of variants of the ordinary telephone, for example, intercoms, sound-powered telephones, and citizen-band radio systems. Although most audio systems are used only for one-to-one communication, almost all of them can be used for multiperson communication as well. For example, most telephone systems allow subscribers to make conference calls through three or more telephones at separate locations connected together for group conversations. Usually the lines are completely open so that any participant may speak at any time. Some limitations of conference calls are their expense, the time required to establish the connections (currently about an hour), and the limitation on the number of participants. The cumulative effects of background noise as additional telephones are added to the network usually mean that no more than about six persons can be accommodated.

Loudspeaking telephones, such as the British Post Office LST4, or the Bell Speakerphone, also allow interactions between groups at separate locations. Loudspeaking telephones were developed primarily to allow executives to communicate in a hands-free manner so that they could handle documents and other materials. Because these telephones have omnidirectional microphones, a single instrument will pick up the voices of a number of people around it. The combination of conference call connections and loudspeaking telephones as terminals allows flexible patterns of audio conferences among several locations with several participants at each location.

Television Systems

A number of electronic systems simulate face-to-face communication with various degrees of fidelity. Some of the simplest are videophones, telephones with attached television cameras and cathode ray tubes as viewing screens. Such systems are operating in Sweden, the United Kingdom, France, and the Netherlands. Probably the most ambitious sales and promotional effort for such systems is being carried on by A. T. & T. with its Picturephone®. The Picturephone® screen is small, 12.7 by 14.0 cm, and has a black and white picture with a resolution of 250 lines per frame. Although the camera can be zoomed, it is intended to show only the head and shoulders of one person. Graphic material can be transmitted but the resolution is not good enough for reading a full page of ordinary type. Videophones are normally used for one-to-one conversations but can be connected together for multiperson conversations.

At the other end of the spectrum are video conference facilities that make use of larger screens and may use color. Westinghouse, for example, has recently set up conference facilities between its Air Arm Division near Baltimore, Md., and another of its plants in Lima, Ohio. The facilities have voice-captured cameras and project images in color that measure 1.32 by 1.73 m in size. Transmission is via the ATS-6 satellite.

Telewriting Systems

Although not very well known, there are several commercially available versions of telewriting systems, for example, Telautograph's Telepen and Victorgraphic's Electrowriter. In general, these systems have both a send and a receive unit at each station. Although a number of stations may be "hard wired" together, messages are normally carried through ordinary telephone circuits. In the latter case, the number of stations that may be interconnected is limited by the amount of noise that builds up as additional stations are added to the network.

These telewriting systems permit a sender's handwritten message or hand-produced diagram or drawings to be transmitted simultaneously to

all the other receive units connected to it. Other persons may either make their own additions to messages already received or may initiate their own replies on fresh pieces of paper.

Somewhat more sophisticated versions of telewriting systems are the RAND tablet, which operates through a computer, or the electronic blackboard still in the experimental stages at the Bell Laboratories.

Teletypewriting Systems

Conference teletypewriter systems have multiple teletypewriters, or input-output writers, connected together so that whatever one person types is produced simultaneously on all the other teletypewriters to which it is connected. In half-duplex systems, a teletypewriter at one location is used both to send and to receive messages. In this case, only one message can be transmitted at a time. More elaborate systems may have two or more machines at each station, one writer to send, the other(s) to receive messages. In the latter case, messages may be sent and received at the same time.

The most advanced systems of this type are those that make use of a computer [5, 6]. Each participant types in a message which then is stored in a computer. Messages can be retrieved, either en masse or selectively, at any time. This type of system has both advantages and disadvantages. Some of the main advantages are that conferences can be asynchronous, that is, participants may type in their messages at any time they are free and may catch up with prior conversations whenever they please. Since the messages are stored in a computer, the sorting services of the computer can be used to retrieve messages according to certain dates, participants, or content. Some of the principal disadvantages would appear to be that conferences can stretch out over such long periods of time that the responsive nature of really interactive communication is lost. Participants may also be flooded with irrelevant messages and so lose sight of the intent and primary purpose of a conference. However, this kind of conferencing is still so new that it has not yet been properly evaluated in carefully conducted laboratory or field trials.

ON THE ROLE OF TELECOMMUNICATIONS IN SOCIETY

Of all the technological advances we have witnessed since World War II, some of the most sweeping and far reaching are those associated with our vastly increased ability to communicate. Telecommunication is the glue that holds modern society together. It provides us with the power to direct and to organize at a distance, that is, to coordinate human activities in one place with those in other places. This immense power to organize makes possible trade, business, industry, and travel as we know them today. Our complete reliance on modern communication for the conduct of even our most ordinary activities is revealed most dramatically when we are deprived of them, even for a short time, as has happened during power blackouts, after fires that have destroyed central communications stations, or in the aftermath of natural catastrophes such as earthquakes.

Communication as an Expander

It has been said that communication has shrunk our world. The truth of the matter is that the shrinkage has occurred only in time. Communication has greatly expanded our world in terms of the personal contacts and experiences it provides us. Through the marvels of modern communication, there is much more for us all to see, hear, and absorb. We are also called upon to try to understand happenings and the affairs of people in remote regions of the world, to assimilate greater amounts and more varied information than heretofore, and to make critical decisions about our own affairs and events in more distant places. In bringing us all closer together, communication has also revealed how dependent we are on one another.

Some Characteristics of the Communication Explosion

Three things characterize the communication explosion: its geographic coverage, its volume, and its technical complexity and diversity.

Geographic Coverage—The world today is so blanketed with communication facilities that events in many parts of it can be viewed in most metropolitan areas as they are happening and discussed interactively by reporters on the scene with those in studios hundreds or thousands of kilometers away. News from much of the rest of the world can be received within minutes or at most hours, and delays in the receipt of news in excess of a day from any part of the world are rare and are generally the result of natural disasters, accidents, or unusual handicaps. Yet only a century ago, the response time to events in distant regions was at least several weeks and sometimes several months. Of the many sets of statistical data that could be used to convey some idea of the quantitative dimensions of the communication explosion, we have selected those in Figure 1 to show the increase in geographic coverage provided by one kind of communication facility within the span of two decades.

Communication Volume—The second characteristic of the communication explosion is the volume of telecommunications that now takes place. The raw numbers are so large that they almost literally defy comprehension. A. T. & T. estimates that more than 300 billion person-to-person telephone conversations are held per year throughout the world. There is now more than one telephone for every 10 persons, one radio receiver for every 4 persons, and one television set for every 11 persons throughout the world. To be sure, these facilities are not distributed evenly throughout the various regions of the earth. The

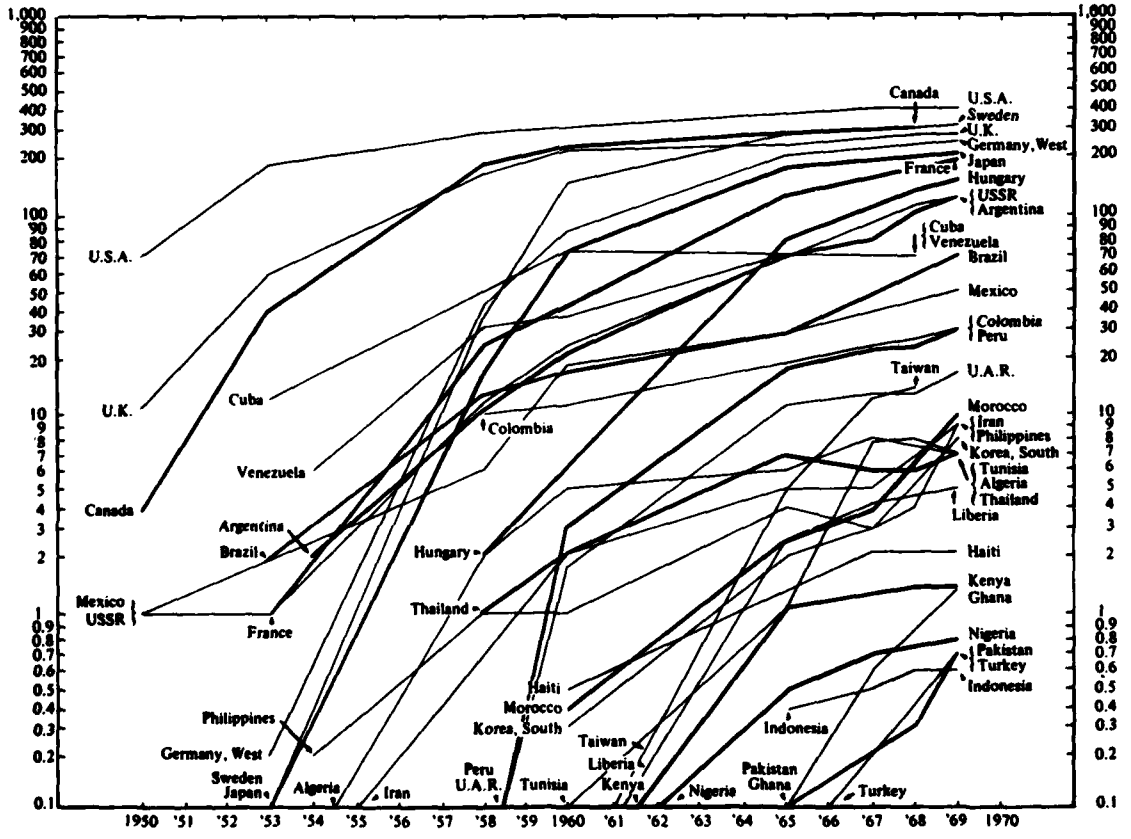


Figure 1—Television receivers per 1000 persons, 1950–1969 [7]

greatest concentration occurs in North America, the second largest in Europe, and the smallest in South Asia. Still, the numbers are large and growing steadily.

The facilities that support these means of telecommunications have expanded accordingly. The first transatlantic telephone cable was laid only 20 years ago, in 1956, with a capacity of 36 telephone circuits. Additional cables were laid in 1959, 1963, 1965, 1968, and 1971, increasing the number of circuits to 2700. Even so, the demand has always seemed to exceed capacity. Today, geostationary satellites, some 37 000 km above the surface of the earth, provide thousands of additional channels and even they are becoming overloaded. And these are in addition to millions of ordinary landline, microwave relay, laser, and radio channels. It appears that modern society has an insatiable appetite for telecommunications.

Technical Complexity and Diversity—The third characteristic of the communication explosion is that the technical equipment supporting telecommunications has become enormously complex and costly and so has greatly increased our requirements for technical training and specialism to maintain it. At the same time, this complexity has greatly increased human factors problems of adapting telecommunication systems to the needs of their users. As an example, in 1972 the Division of Health Care of the National Center for Health Services funded seven experimental telemedicine projects at a cost of nearly a million dollars. One of the main conclusions to emerge from a review of those projects was the seriousness of the mismatches between the equipment and the users of it. The following quotation [8] summarizes succinctly some of the difficulties encountered:

> *All projects* involving interactive television *experienced* continual *aggravation by the complexity*, [low] *reliability*, *lack of ubiquity*, *size*, [insufficient] *mobility*, *personnel support requirements* and *set up time* of the television equipment. Such complaints pervaded the comments and attitudes of the physicians and non-physician staff personnel. (Words in square brackets are our insertions.)

The other side of the coin is the diversity of telecommunication options that technology now offers. Although refinements such as touch-tone and direct long distance dialing represent truly great advances in user convenience and effectiveness, they are minor considerations in comparison with the increase in telecommunication alternatives we have today: telephone, telex, commercial and citizens-band radio, closed circuit television, Picturephone®, and facsimile data transmission. Deciding among these many possibilities is one of the major human considerations in telecommunications today.

TELECOMMUNICATIONS IN THE NAVAL ESTABLISHMENT

The importance of communication for the control of farflung empires and for waging war was recognized even by the ancients in the writings of Herodotus, Xenophon, and Polybius. Naval warfare, especially, has always had its own requirements for communication. Units often operate out of sight of land bases for extended periods of time and, as recently as 1900, a ship was isolated once she had left a port and sailed over the horizon. One of the most important technological innovations in naval communications occurred when the first official radio message from a U.S. naval vessel was transmitted from the U.S.S., *New York* on November 2, 1899. The introduction and use of wireless telegraphy for tactical purposes, less than 70 years ago, completely revolutionized naval warfare. It meant that for the first time, vessels could communicate with each other even though they were out of each other's sight.

Advances since that time have followed in almost bewildering succession. Radioteletypewriters greatly increased the flexibility and usefulness of ordinary radio systems, and radio photo (facsimile) equipment made it possible to transmit maps, charts, photographs, and other printed materials. Telephony appeared in both familiar and less familiar forms, for example, as sound-powered telephony and radiotelephony. Television extended man's sight to distances limited only by his willingness to accept the cost and trouble of installing the necessary equipment and instrumentation.

On Technological Forecasting for the Navy

It is impossible for scientists outside the Naval Establishment to make accurate estimates about the role telecommunications might play in naval operations of the future. For one thing, naval tactics are classified. More important, however, it is virtually impossible to predict the long-term consequences of major technological innovations on human activities in general.

Recall that only a hundred years ago almost no one thought seriously that the telephone had any commercial possibilities. After Alexander Graham Bell demonstrated his new invention at the Centennial Exhibition held in Philadelphia in 1876, the New York *Tribune* commented editorially:

> Of what use is such an invention? Well, there may be occasions of state when it is necessary for officials who are far apart to talk with each other without the interference of an operator. Or some lover may wish to pop the question directly into the ear of a lady and hear for himself her reply, though miles away; it is not for us to guess how courtships will be carried on in the twentieth century.

One of Bell's kindest critics, a friend with some knowledge about scientific matters, explained that, since every spoken word has many delicate vibrations that must be converted into electric waves by the telephone, a message would not be intelligible if any of them got lost. Obviously, any device so liable to error could never have any practical value [9].

We are also reminded of the exercises conducted in 1906 by the U.S. Atlantic Fleet over large areas of the ocean in an attempt to develop the strategic use of radio. These exercises were judged to be a failure by senior naval officers and set back for several years the naval use of radio.

We know of no way to predict reliably on theoretical or on a priori grounds the eventual impact of technical advances on society. Rather, the eventual usefulness of new technologies must be assessed through empirical research and field trials.

Despite these caveats, there is some merit in trying to see possible implications of new technologies on human affairs and activities. Although many of these visions turn out to have little or no merit, they serve to stir us out of our conventional ways of thinking and to stimulate us into looking at bold new ways of doing things. Indeed, without such attempts to change, society would quickly become stagnant. With this in mind, we peer briefly into the clouded crystal ball to discern what role person-to-person telecommunications might play in the Navy of the future. Although some statements we make are based on Navy sources which for a number of reasons must remain uncited, the statements are all ours and are in no way to be interpreted as reflecting official naval doctrine or thinking.

The Role of Communications in the Development of Navy Command Concepts

One of the most significant changes that has occurred in naval operations through the centuries has been the centralization of authority. In ancient times, when communications were primitive, commanders of naval vessels operated under only the most general instructions and made their own decisions locally. The first major change came about 1890 when commercial telegraphic or cable facilities became available in virtually every port used by the Navy. These facilities provided rapid communication between the Navy Department and commanders of naval squadrons, when they were in port. Not only did such communications make it possible for the Navy Department to keep its officers abreast of political and military situations, but it also greatly decreased the amount of discretion that commanders had for their actions. A senior officer in China is said to have commented at about this time, "Now we have become mere messenger boys at the end of the cable" [10].

The situation changed again dramatically when radio made it possible for the Navy Department to communicate with ships at sea. This increased capacity for interactive communication made it much easier to keep tabs on changing local situations and to coordinate the activities of farflung naval operations. It was also a further step in the direction of centralization of authority.

These historical developments have led to the current Navy command concept—a unity of command, with responsibility and authority vested in a single individual who, through a hierarchical structure, sets policies, assigns tasks, and supervises the operations of subordinates. Each commander in the command hierarchy operates in essentially this manner.

Future Trends

Without compromising the basic principle of Navy command, the nature of warfare at sea is certain to become more complex and dependent on the utilization of the most advanced technology. Indeed, some sources have characterized naval warfare of the future as being, first of all, an "information war." The side that is able most quickly to gather, assimilate, and act upon information will have the tactical advantage.

It also seems certain that allowable response times will be greatly decreased in future warfare, requiring the rapid assimilation and integration of information and rapid decisionmaking based on that information. More than ever naval activities will be conducted over vast distances in which it will be impossible for most command elements to conduct their business face to face. Finally, situations in future wars may require task groups and units to shift from one chain of command to another with no appreciable delay.

The Navy as a Business Organization

Entirely aside from its military function, the Navy may be regarded as a very large business organization. It employs nearly 850 000 persons and carries out a great variety of business functions. Thousands of people each year are recruited, screened, selected, trained, evaluated, and promoted in several hundred different occupational specialties. The Navy writes specifications for, orders, procures, builds, maintains, and repairs hundreds of thousands of items, from stationery supplies to enormously complex systems such as nuclear-powered submarines. It operates and staffs hospitals and provides medical services for its personnel. It carries on and supports a diverse program of research and development. All these business activities are carried out in establishments that are almost literally scattered over the face of the globe.

The execution of these functions requires an inordinate number of interpersonal contacts. No one really knows how many conferences go on each year in carrying out the Navy's business, but it must be some astronomically large number. Assuming that the Navy is not substantially different from other large business organizations, we can confidently assume that the typical Navy employee or military person spends at least half of his working time in some form of communication [11]. That represents a considerable amount of communicating.

The Importance of Person-to-Person Telecommunications in the Navy

All the characteristics described previously—the necessity for gathering and assimilating information, for coordinating the activities of widely scattered groups, for arriving at decisions, and for conducting all these activities over vast distances—are precisely the conditions for which person-to-person telecommunications seem to have been made. The Navy is ultimately made up of people and it is people who in the final analysis must assess situations, gather and report information, assimilate that information, coordinate activities, and arrive at decisions. To be sure, the people in the system may be assisted by computers and other technological devices, but the decisions and actions are ultimately humanly derived and humanly based. In this picture, person-to-person telecommunications are vital. How best to select among the various telecommunications options, how best to design and organize them, and how best to use these facilities are, in our opinion, problems of great importance to the Navy. Moreover, these problems are almost certain to increase rather than decrease in the years to come.

PSYCHOLOGICAL PROBLEMS OF TELECONFERENCING

Teleconferencing can be done in a great many ways. Some of these ways—for example, communicating by closed-circuit television—seem

superficially to be quite similar to face-to-face communication. Other ways—for example, audio conferencing or computer conferencing through teletype terminals—seem *quite different from* face-to-face communication. When they are examined critically, however, it turns out that all forms of teleconferencing differ from face-to-face communication in a number of respects, and the differences are often significant for human interaction. Although some designers and users of more complex telecommunication systems may feel that conference television is "just like face-to-face," this opinion, as we shall show, is merely an indication that they have not fully considered or appreciated the many differences among communication modes that have some psychological significance.

In this section, we elaborate on some of the major psychological problems of teleconferencing. However, we are handicapped in this enumeration of problems by the incompleteness of present knowledge regarding face-to-face communication, the standard or criterion against which various kinds of teleconferencing are usually compared. Description of the complex processes of human interaction, both verbal (through language) and nonverbal (through facial expression, gestures, tone of voice, and other cues), is still very far from complete, despite a considerable amount of research activity that has been devoted to it [12, 13].

Some critics of telecommunication systems have contended that introducing any artificial or mediated links between communicators is most likely to disrupt the smooth flow of human communication. However, this is not necessarily so. One can conceive of ways in which telecommunications could have important advantages over face-to-face communication, entirely aside from the obvious and very important advantage that all forms of telecommunication have in allowing us to conquer space. As an example, using the telephone can speed business transactions. The ring of a telephone is so insistent that it is usually given priority over other business, that is, a telephone caller is able to "jump the queue" ahead of other people who are waiting for face-to-face attention. In addition, some social niceties, such as offering refreshments, are completely omitted from telephone conferences. There are thus some business situations in which use of the telephone might have substantial advantages, even if, through a futuristic transport system, one could travel *instantaneously* from anywhere to anywhere.

We turn now to a more detailed discussion of the various human problems of teleconferencing. In many cases, this involves a certain amount of speculation. Although differences between the media exist, it is not always clear what psychological impact, if any, these have.

Visual Cues

The following visual cues about the communicators have been shown to be important in face-to-face communication [12, 13]:

- Direction of gaze, especially eye contact
- Facial expression
- Gestures and other bodily movements
- Body posture and orientation
- Proximity, i.e., physical distance between communicators
- Physical appearance, e.g., attractiveness, hair length.

Moreover, all these cues have been shown to have some communicative value. To varying degrees, all telecommunication systems omit or distort these visual cues. An audio system (e.g., the telephone), or a written system (e.g., telautograph), transmits no visual cues about the communicators. Video systems (e.g., Picturephone®), transmit some visual cues but omit others, such as leg and body position, and distort still others, such as apparent proximity and eye contact. The latter effect is due to the displacement of camera and screen, so that a gaze at the eyes of a person on the screen is a gaze away from the camera, and thus appears as gaze aversion.

What are the effects of the omission or distortion of visual cues on communication? The relevant literature suggests many effects, since visual cues seem to be implicated in the communication of such diverse "messages" as superiority [14], romantic love [15], and persuasiveness [16]. The one safe generalization that emerges from this literature is that nonverbal cues have an important role in forming, building, or maintaining relation-

ships between people. The absence of visual channels seems likely to produce disturbances in the socioemotional aspects of the interaction but will not seriously affect the transmission of cognitive information which is primarily transmitted through the verbal channel. Some such disturbances have been demonstrated in experiments which will be discussed in the following section.

Apart from the transmission of nonverbal cues, the visual channel has two other important functions. First, it helps identify who is speaking. Some audio-only telecommunication systems can be used by a large number of speakers only if each person gives their name before speaking, a procedure that often seems overly formal and disrupts the smooth flow of conversation. Voice characteristics are often inadequate for identification if the group members do not know each other, if large numbers of people are participating, or if the audio link is of poor quality.

Although we usually identify who is speaking through such visual cues as mouth movements and gesticulations, automatic methods of speaker identification using other channels are possible and have been incorporated in some systems, e.g., the Remote Meeting Table [17], used in several parts of the British government. In this system, each of a pair of interconnected tables has six microphones, one for each of the participants seated around the table. A loudspeaker is placed between each pair of microphones, each loudspeaker corresponding to the position of a speaker at the remote table. When a participant speaks he captures and activates his microphone by virtue of the differential loudness of his voice in the several microphones. His voice is then transmitted to his own loudspeaker at the distant location. Each speaker's name appears above his respective loudspeaker and a light is sometimes used to indicate which loudspeaker is carrying a message.

The second important function that the visual channel serves is to permit diagrams, documents, or other graphic material to be shown. The variety of graphics that may be used is enormous. In some meetings or types of business, the liberal use of slides, films, viewgraphs, and blueprints is commonplace, while in others, such aids are never used. In some cases, participants may even want to modify the graphics as they talk, erasing and adding to some display such as a blackboard. Clearly, the design of a single telecommunication system that can accommodate all these kinds of graphic displays is virtually impossible. There are however, a number of systems that do transmit graphic material with varying degrees of adequacy, e.g., facsimile, teletype, Picturephone®, telautograph, electronic blackboard, Scribblephone.

Physical Separation

By its nature, telecommunications implies physical separation between communicators. Although some telecommunication systems can transmit most visual and auditory information, they inevitably omit other cues, such as touch and smell. These latter may seem relatively unimportant to diffident Anglo-Saxons but may not be so in other cultures. Physical contact between Arabs for example, is frequent and of considerable importance in the interaction process [18]. The warmth of a handshake may be an important part of a meeting, particularly between strangers. As a final example, consider the following quote from an interview with a British civil servant, ". . . we had arranged for coffee or tea to be served at our end . . . and he [the person at the far end] didn't have any. We sat there drinking our coffee and passing the biscuits around and he looked increasingly glum" [19]. Since nobody is likely to invent a telecommunication system that will transmit refreshments, such problems of physical separation seem likely to persist for a long time.

Input-Output Problems

Some telecommunication systems, such as the telephone, have few input-output problems. The speaker speaks much as he would face-to-face, and the listener listens, again much as in face-to-face communication. Providing there is not too much distortion on the line, communication should proceed normally. However, things are not that simple in all telecommunication systems. In some cases, input must be in a special form, as with Morse code taps on a telegraph system or

with keyboard typing for computer conferencing. Speaking comes naturally and quickly to most adults, while typing or Morse code are relatively slow and laborious [20]. Thus, telecommunication systems that impose such constraints on inputs may create difficulties for most human users, although some persons, e.g., the deaf, may find them advantageous. Although these inconvenient forms of input have been adopted primarily because of transmission limitations of the relevant telecommunication systems, there may be compensatory advantages on the output side. Morse code is more easily decipherable than speech under noisy conditions, and most adults can read much faster than they can speak.

Another output problem relates to the delivery of messages. In some telecommunication systems, e.g., computer conferencing, what goes in may never come out. This may cause problems for the sender of a message: if he receives no reply, he does not know whether this is due to the nondelivery of his message, the nondelivery of the reply, or deliberate neglect of his message by the other party. Since the best course of action is different according to which of these explanations is correct, the sender may not know how to react and may subsequently avoid a medium that has uncertain delivery. The development of appropriate feedback mechanisms will undoubtedly be a major consideration in all telecommunication systems.

Information Overload and the Organization of Time

In many communication systems the initiative for starting a conversation lies with one party, the caller. This means that there is some probability that a message will come at an inconvenient time for the recipient. Being called out of the bath by a telephone call is an obvious problem, but less obvious is the disruption that incoming messages can cause to other activities such as reading, thinking, or meeting face to face. To quote Meier [21],

> Observation of human interaction suggests that a prime cause of stress in human behavior is the appearance of signals or cues calling for the initiation of a new operation before the current one is completed. A choice must be made as to which is more important. At the present time the telephones and "intercom" systems almost always win, and a flurry of calls leaves behind a debris of incompleted sequences of behavior upon which effort has been expended but for which personal rewards have not yet been realized. Increasing interruptions seem to be associated with increasing stress . . . too much stress is destructive and even deadly. Each system has its outer limits of endurance.

People will often try to avoid such stress. Two major methods of reducing overload are to use an assistant who sequences inputs for the principal (a role often played by receptionists and secretaries) and insistence on the use of media that the recipient can use at times that he chooses (letters, memos, telephone recording devices). Computer conferencing is the newest system to offer the advantage of an input and output that can be asynchronous.

Existence of a Record of Previous Transactions

Some methods of communication, such as telewriters and teletypewriters, leave a permanent record, while others, such as the telephone, do not normally do so. Computer conferencing is particularly good for "on-the-record" discussion, since all material is automatically recorded verbatim, with time and originator of the message automatically attached. A permanent record has both advantages and disadvantages. On the one hand, a permanent record ensures that important statements or decisions can be referred to subsequently, even by those who were not present at the time. For that reason, word-by-word records are kept of the proceedings of important institutions such as the United Nations, Congress, and the courts. However, the completeness and openness of such records often cause the most important business to be off the record in informal meetings, while the on-the-record forum becomes merely a talking shop, where rhetoric is plentiful

but few decisions are made. Sensitive negotiations are also usually done off the record and only after agreement has been reached are the final decisions put in writing. There will always be some question as to whether records should be kept and how extensive those records should be. The answer to that question will help determine the choice and design of telecommunications media.

Group Dynamics

Face-to-face meetings are conducive to various sociopsychological processes, such as the emergence of leadership, conformity, and the formation of subgroups which have been subsumed under title "group dynamics." Extensive research on group dynamics [22] indicates many ways in which such processes affect group productivity and cohesion. It appears likely that the use of new telecommunication systems will affect these group processes.

In normal face-to-face meetings, leaders tend to emerge who then fulfill various functions such as focusing the group's attention on the problem, maintaining group solidarity, and organizing the participation of group members [23]. In telecommunicating groups, the leader's role may be diminished or greatly strengthened. Diminution of the leader's role may occur if he is denied some of his prerogatives. If, for instance, he cannot exert dominance over the other members through nonverbal cues, the leader's position is weakened. The result may be a disorganized group with no clear leader. On the other hand, the leader's position is probably strengthened by some telecommunication systems. For example, the chairman may be able to cut off other group members by controlling whether their microphones are on or off. Such power is likely to alter the relationship of the leader with other participants.

Some telecommunication systems may require or produce more than one leader. Many systems are designed for the use of groups at two or more locations, in which case there may be a leader for each group. Whether there is then a leader for the conference as a whole or whether there is fragmentation into subgroups may depend on other factors. Certainly one could hypothesize that, if a conflict situation arose, the division into subgroups by location could be complicated by coalitions between group members that form in response to the conflict.

Most of the discussion in face-to-face groups is both public (everyone can receive the message) and personalized (one can identify the originator of the messages). However, some communications may be private, by whispered conversations or hurried notes, and in very large groups it may be possible to make anonymous comments because most listeners cannot identify the source of the message.

Telecommunication systems often alter the public-private and personalized-anonymous balances. The problem of identifying the speaker in many audio systems has been previously mentioned. Most audio systems also prevent private messages, everyone is "on line" at all times. This medium thus shifts the conversation to an anonymous, public mode quite unlike anything we normally encounter face to face. Computer conferencing, on the other hand, provides a public, personalized channel, as well as a private channel that is more private than anything encountered face to face. It is impossible even to detect that private messages are being sent by other participants, let alone discover their content. Some computer conferencing systems also provide an anonymous channel that is more anonymous than any commonly encountered in present-day systems. Since anonymity has already been shown to influence social behavior [24] and privacy also seems likely to do so, the potential significance of these differences between face-to-face and some telecommunication media should not be ignored.

Security and Confidentiality

Participants in face-to-face meetings frequently do not want the material they are discussing to be made public. Even though it is rare that the contents of a meeting rate as top secret, lesser degrees of confidentiality are common for most business meetings. Business and military espionage are not unknown and, for this reason, many potential users question the security of new telecommunication systems. Naturally, face-to-face meetings

and telephone calls are not particularly resistant to espionage, but, because these are familiar, users are not as suspicious of them as they are of new media. The use of communication satellites as links in the system gives rise to especially serious security worries, since these broadcast signals, at least with the newer satellites, can be picked up by relatively unsophisticated antennas. Electronic scrambling methods have been developed to help insure privacy in both audio and video links, but these can be quite expensive and so are generally used only for special purposes.

In concluding this section, one can identify on a priori and empirical grounds many psychological problems in the use of new telecommunication media. As we shall see in the next section, some psychological effects have actually been measured. In some cases, it could be reasonably argued that these are not problems but merely differences and that the use of telecommunications could in some circumstances be superior to face-to-face communication. However, in other cases the effects of telecommunications usage are clearly detrimental. How these psychological problems might be solved is a topic to which we now turn.

PREVIOUS RESEARCH ON THE PSYCHOLOGY OF TELECOMMUNICATING

Studies of human performance in telecommunication systems can be grouped into three classes: uncontrolled field trials, controlled field experiments, and laboratory experiments. To demonstrate the strengths and weaknesses of these methods of inquiry and to summarize the most interesting findings from such studies, we shall describe briefly some examples of each type.

Uncontrolled Field Trials

Most investigations of new telecommunication systems have by and large been pragmatic. The aim of implementing the system has been to improve the functioning of an organization or to create a new and profitable service. For this reason most innovators have usually adopted a fairly crude "try it and see" approach. They install some equipment and see if people will use it. It is usually assumed that one can easily judge whether the system is a success or failure, so there is often little monitoring or measurement. As examples, we shall take two very contrasting trials using this approach. The first was a video teleconferencing system (closed-circuit television) within the Department of Environment (DoE) in the United Kingdom. The system had two locations about 2.5 km apart across the River Thames. The link was achieved through line-of-sight microwaves. Each studio could accommodate three people comfortably and more at a squeeze. The system was available free of charge to any of the several thousands of employees in the buildings that contained the studios. However, usage was dismally low. Only two groups of people ever used it (one group doing so on several occasions) and, after a while, usage dropped to zero. The system was proclaimed a failure and dismantled.

Compare this with results at the National Aeronautics and Space Administration (NASA) in the United States which introduced an audio conference system to be used by its own employees and those of associated contractors. The system has been in use for some 8 years and by 1976 had expanded to about 30 studios. Collectively, the studios attract about 30 000 man-meetings per year, resulting in an estimated saving to the organization (travel costs saved minus telecommunication costs expended) of about $500 000 per year. The consensus is that the system has been extremely successful. More complete summaries of both the NASA and DoE systems can be found in Hough [25].

It would be easy to jump to conclusions on the basis of these two field trials. However, there are many reasonable hypotheses that could be advanced to explain the differences in the apparent success of these two systems:

- NASA had a better designed teleconference system than did DoE.
- The NASA conference rooms were more easily accessible than were those in the DoE.
- NASA has more meetings of a type suitable for teleconferencing than does DoE.
- Publicity for the NASA system was better than that for the DoE system which, in fact, seems to have been especially bad.

• The NASA locations were more dispersed (several hundreds or even thousands of kilometers apart) than the DoE locations (only 2.5 km apart) so that the incentive to avoid traveling was much greater in the former case.

The problem is that, given the uncontrolled nature of these field trials, it is virtually impossible to say which of these and still other hypotheses are correct. We thus cannot tell what is and what is not crucial to the success of a teleconference system. We know only that telecommunication systems can be successful and that they can fail, but we don't know why. Although field trials are a good test of the feasibility of a design in the real world, they are inadequate as the sole method of study.

Controlled Field Experiments

Due to the uncertainty of inference from the results of field trials, most researchers prefer more carefully controlled methods of study. In most cases this means laboratory research, such as will be described in the next section. However, in some cases researchers have succeeded in carrying out investigations in a field situation which has at least some of the control of laboratory research.

An example is the study of media differences in telemedicine by Conrath, Dunn, Swanson, and Buckingham [26]. Patients, who had been recently seen for medical problems by their doctor, were asked to return to the clinic to take part in an experiment. When they returned, they were allocated by a random process to a particular sequence of four successive diagnostic consultations, each with a different doctor (and not the same one who originally saw them), in four different media: face to face, two-way color television, two-way black-and-white television, and hands-free telephone. In the three telecommunication conditions, a nurse was in the same room as the patient, helping to transmit information to the doctor who was elsewhere in the building. In all, 32 patients and 8 doctors took part.

The results were unexpected. The face-to-face and the three telecommunication media were equally effective in terms of accuracy with which physicians could diagnose most critical ailments. Average consultation time was also unaffected by medium. Face-to-face consultation was more effective than the telecommunication media for detecting subsidiary ailments, but there were no reliable differences among the telecommunication media in this respect. Thus, although the doctors and patients preferred face to face to telecommunications and preferred television to telephone, the objective data show that performance did not always support their subjective preferences.

Note the experimental controls used in this study. Various media were compared, patients and doctors used the media according to a random schedule rather than according to their own preferences, and systematic outcome and attitude measures were taken. These controls make it possible to draw some fairly positive conclusions from this study, unlike the situation with uncontrolled field trials. Compare the results of this study with those of the seven telemedicine field trials summarized by O'Neill et al. [8], where few positive conclusions can be drawn about the relative effectiveness of different media.

Field experiments, however, do have serious limitations. Some realism was sacrificed in the Conrath et al. [26] study to gain experimental control. Patients had the unusual experience of undergoing multiple consultations and both doctors and patients knew they were part of an experiment. However, the effects of knowing that one is part of an experiment (the so-called Hawthorne effect) are not specific to this method. They plague nearly all uncontrolled studies and laboratory studies as well. In addition, it is often difficult to find participants as cooperative as the doctors and patients in Conrath's study. Furthermore, some interactions, such as business meetings, are less standardized than medical consultations. Finally, there are often ethical problems involved in monitoring people's behavior in the field.

Laboratory Experiments

About 30 laboratory experiments have been reported on the effectiveness of communications

media. These all have the following characteristics in common: participants are invited to a laboratory where they are organized into groups of two to six (according to the study) and randomly allocated to communicate face to face or by some telecommunication medium. They are given a standard task to complete during the meeting. The tasks used have varied widely among the 30 studies. Various dependent measures are taken, including length of time to finish the task, task outcome or solution, verbal processes from tape recordings and transcripts, nonverbal behavior, and participant attitudes and opinions. All these data are then reduced to numerical form and analyzed statistically.

Rather than summarize all these studies (which has been done elsewhere [19] we shall describe three particularly interesting experiments. In the first, by Chapanis et al. [20], 20 pairs of participants communicated to solve a problem. The two problems that were used had objective solutions: for instance, the correct assembly of a trash-can toter. One of the participants, designated the "source," had all the instructions, while the other, the "seeker," had the parts to be assembled. They communicated by one of four media: face to face, audio only, handwritten notes, or teletypewriters. The results show that the time needed to reach a solution was strongly affected by medium (Figure 2). Both face to face and audio only were much faster than handwriting or typewriting in time required to reach a solution, although neither the former two nor the latter two differed from each other in solution time. Interestingly, there was no difference in solution time between experienced and inexperienced typists, a finding confirmed independently by Weeks, Kelly, and Chapanis [27]. The differences that were found could be traced, through observational analysis of the participants' behaviors, both to the slowness of input and output in the "hard-copy" modes and to the difficulty of engaging in other activities (e.g., searching) while communicating by these modes.

The second study, by Short [28], deals with negotiation. Forty-eight pairs of participants communicated with each other face to face, over closed-circuit television, or by an audio-only link. The problem involved bargaining about cuts in budget items of a hypothetical government

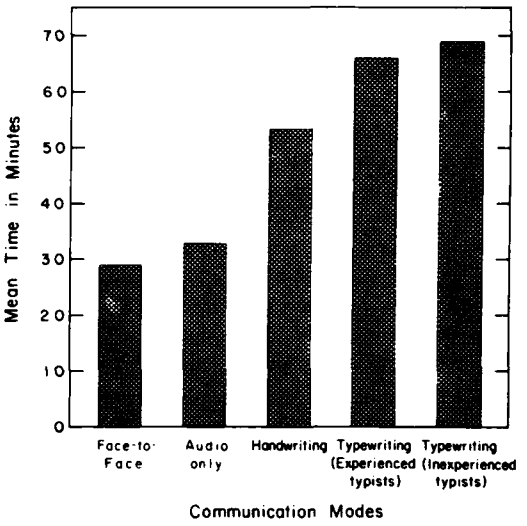


Figure 2—Average times to solve problems in four communication modes [20]

agency. One of each pair argued a case consonant with his own views. The other was given a brief which did not accord with his own private opinion. Bargaining success could be quantified in terms of the final solution. Each item retained in the budget had a numerical payoff for each "player" and his bargaining success was indicated by the size of his total payoff.

Results showed that the person arguing a case consonant with his own view was more successful than the person arguing a brief when the medium was face to face or closed-circuit television. However, when the medium was audio only, the person arguing a brief was more successful than the opponent who believed in his case. This result was explained in terms of the extent to which the media encourage interpersonal as opposed to interparty considerations. Face-to-face communication encourages the intrusion of interpersonal considerations, which benefits the person who is arguing for his personal opinions, as opposed to a brief.

The third study, by Williams [29], dealt with coalition formation over telecommunication media. In many teleconference systems, a larger group is split into smaller groups at the various locations, and Williams hypothesized that this

split might affect the patterns of support and of disagreement. Forty-five groups of four people took part, with equal numbers of groups communicating face to face, by a closed-circuit television link, or by an audio link. In both telecommunication conditions, the groups of four were split into a pair of people at each of the two locations. The task for the groups was to generate ideas on improving transportation in Britain. A secretary noted the names of proposers, seconders, and dissenters for all ideas generated.

Results showed that the medium of communication did not affect the number of ideas generated or their judged quality and originality. However, it did affect patterns of support. For both the television and audio conditions, the proposer and second were more frequently at the same location than would have been expected by chance. Furthermore, in the audio condition, dissenters were more frequently at the opposite end of the link from both proposer and seconder than chance expectation. It seems that the spatial division produced by the use of the medium was producing a division of the group into two opposing subgroups. It was also found that in the audio condition, participants judged their partners in the same room to be more intelligent, constructive, competent, trustworthy, and sensible and less impersonal, boring, and unreasonable than the two members of the subgroup in the far room.

These three examples demonstrate that laboratory experiments can give clear-cut results, indicating differences between communication media without the problems of inference associated with uncontrolled field trials. Admittedly, these benefits are gained at the expense of realism: these are very clearly experiments rather than real life. Some realism can be maintained by using motivated volunteers, representative of real-world populations, as participants and by using problems that are chosen from real-life situations. Perhaps the greatest advantage of laboratory experiments is their cost-effectiveness. A single well-designed laboratory experiment can provide hard data on more variables and more interactions than can be reasonably tested in any field trial. Moreover, in this area a laboratory experiment can often be done for from one-tenth to one-twentieth the cost of a single field experiment. That is not a negligible consideration.

Some Conclusions From Laboratory Research

At present, the only area of research on the effectiveness of telecommunication media that has been comprehensively explored is the suitability of audio and video media for a variety of tasks. Even here, much uncertainty exists, but it is possible to give a first estimate of what types of business meetings identified by Pye, Champness, Collins, and Connell [30] could be transferred from face to face to telecommunication media. Table 1, from Short et al. [19], gives one such first estimate.

SOME RESEARCH NEEDS

Technology for person-to-person telecommunications has advanced to the point where it is no longer important to ask "What can we do?" but rather "What should we do?" The answer to the question of "What should we do?" involves many elements—cost and effectiveness being two of the most important.

The most reasonable prediction one can make at present is for a world in which energy costs will continually increase for at least the next few decades. Since travel is a heavy user of energy and of oil, the most precious form of energy at that, it seems safe to predict that travel costs will increase for some decades to come. Under these circumstances, face to face meetings are certain to become more and more expensive and the choice will not be "teleconference or travel to a face to face meeting" but rather "teleconference or nothing." If this becomes the choice, the relative effectiveness of face to face conferences and teleconferencing will become less important. The important question will not be "Do we teleconference?" but rather "How do we teleconference?"

Critical in all these decisions is the question of effectiveness. Here we must not lose sight of the fact that this criterion is ultimately a human one. Communications, whether they be face to face, telecommunications, or computer mediated, exist for one purpose only and that is to serve man. The engineering options are almost limitless. How to select from among those options is the critical

Table 1

Suitability for Substitution of Various Types of Meeting: An Interim Answer

Substitutability	*Fairly Definitely*	*Tentatively*
Would Have to Remain Face to Face	○ Inspection of Fixed Objects	○ Conflict ○ Negotiation ○ Disciplinary Interview ○ Presentation of Report*
Transferrable to Two-Way Video	○ Forming Impressions of Others	○ Giving Information To Keep People in the Picture ○ Briefing
Transferrable to Two-Way Audio	○ Problem Solving ○ Information Seeking ○ Policy Decisionmaking	○ Discussion of Ideas ○ Delegation of Work*

*Evidence is so scanty that this allocation is virtually pure guesswork.

question. We will sketch briefly some of the research that would help to answer that question.

Studies of Face-to-Face Meetings

In our opinion, research on telecommunications should begin, paradoxically enough, with studies of real face to face meetings. One of the main conclusions that emerges from several questionnaire studies of face to face conferencing is that there is an enormous variety of meetings in business and other organizations. This, in turn, leads to the conclusion that the search for one, ideal teleconferencing system for all situations may be a wasted effort. We may need rather to design many new types of telecommunication devices, each of which is ideally suited to a particular organization or particular situation. That, in turn, raises the question of exactly what is a telecommunication system supposed to be a substitute for. To answer that question, we need observational studies of conferences and meetings to

- Identify and categorize typical conference groups
- Categorize meetings by purpose and function
- Get normative data on the sizes of various conferences, that is, what proportion of all conferences involve two, three, four, and so forth persons
- Identify special communication requirements (blackboards, flip charts, projection equipment) of various kinds of meetings
- Obtain data on the various activities (reading, writing, one-way communication, interactive communication) that transpire during meetings
- Analyze the kinds of telecommunication facilities that might be used as substitutes for typical meeting activities
- Develop methods for evaluating the effectiveness of meetings and conferences of various types.

With such information we will be in a much better position to state more clearly what kinds of meetings and conferences will be or can be served

effectively by what kinds of telecommunication systems.

Studies of Telecommunication Variables

The preceding section and Table 1 show that a number of variables in telecommunications have already been studied systematically. However, other factors have received little or no attention. For example, while the presence of interrupt facilities has received some attention [31, 32, 33], other media variables, such as the existence of privacy channels, of anonymity, or of strong or weak leadership control, have been ignored in favor of the most obvious media differences—the presence or absence of a visual channel. A catalog of all the research that needs to be done on variables of relevance to telecommunications would make a list that is longer than is justified for this paper [34]. We will list a few of the major interesting problems for which empirical data are lacking:

- We still do not know as much as we should about the effectiveness of telewriting and teletypewriting for various communication and conference purposes. How effectively can these media be used either alone or to augment other media? For what kinds of conferencing are these media effective?
- All teleconferencing systems seem to have been built with an implicit, although sometimes peculiar, view of the chairman's role. In conference calls, there is complete laissez-faire—anyone can speak at any time. In computer conferencing, the chairman has more power—he can control who is admitted to the conference and can cut off obstreperous persons. In most video conferencing systems, there are two chairmen, one at each node, who control the video pictures. In some audio-conference systems, the chairman can control the audio circuits and can arbitrarily give the floor to whomever he pleases. What are the effects of giving or not giving the chairman strong systems-based powers?
- In teleconferencing, communication terminals or nodes may be used individually or they may be partially shared. For example, a number of conferees may carry on a teleconference through individual closed-circuit television facilities. Alternatively, two or more conferees in one location might share a camera and monitor. What is the most effective distribution of facilities to conferees?
- Some proponents of computer conferencing argue that the relative anonymity of the participants in computer conferencing constitutes an important advantage of that medium over face to face conferencing. On the other hand, it is possible that people might become more aggressive and less considerate when anonymous. In military situations, the relative anonymity of some forms of telecommunication might change or dilute the effectiveness of military rank per se. What are the facts about anonymity in telecommunication? Is it or isn't it an advantage?
- Practically nothing is known about teleconferencing with groups of different sizes. Face to face conferences can be carried out with very large numbers of people. It is at least conceivable that audio conferencing or conferencing via teletypewriter might become a shambles as the number of conferees increases beyond some number. If there is such a number, what is it? What happens with various telecommunication systems as the number of conferees increases?
- Virtually nothing is known about telecommunications in languages other than English. What are the communication patterns of peoples who speak languages other than English? What special requirements must be met for effective multilanguage teleconferencing?
- How can all these diverse pieces of hardware and equipment be best human engineered to meet the needs of the diverse persons who will use them?

AND WHAT OF THE FUTURE?

Although people do not at all resemble computers physically, some of the things they both do are sufficiently similar that computers have been called "giant brains" [35]. The similarities become even more striking when we compare person-to-person telecommunications with man-computer communications. In the first place, the interactions between man and modern computers may, in a manner of speaking, be thought of as conversations. They are characterized by com-

mands, statements, questions, answers to questions, and sundry other messages that go from man to computer and vice versa. As may be apparent, these exchanges are truly interactive in the sense that we have been using that word here.

Conversations between people and computers are all carried out in one of several different languages which, although they are not exactly colloquial English, are close enough to it so that the language can be recognized and learned more or less easily. To be sure, the input options for communications from man to computer are still limited to typewritten materials, some simple and highly constrained forms of cursor-positioning and handwriting, and a few primitive voice signals. On the other hand, output devices that carry communications from computers to man cover the full range of those that one finds in person-to-person telecommunication systems—printed materials, voice, graphics, and pictures. Most impressive of all, however, is that some computer programs have been made so humanlike that people who have used the system have actually been misled—at least for a time—into believing that they were communicating with another person [36]!

The essential unity of communication problems, whether they be with other people or with computers, is the basis for our belief that the future will see an integration of communication systems that are now seen as separate. Vannevar Bush's visionary article, "As We May Think" [37], first called attention to the extraordinary power that modern computers have to supplement human cognitive functions. Bush saw the computer as providing an enlarged intimate supplement to a user's memory. "Associative trails," much like the associations that characterize human thinking, would make it possible to bring the enormous capacity of modern computers to integrate, file, sort, and compile the contents of encyclopedias, books, newspapers, letters, opinions, and human experiences.

Bush's article was, of course, far ahead of the technology of that time. A similar and more recent endeavor is Licklider's treatment of *Libraries of the Future* [38] which foresaw the revolution in library systems now beginning to appear in such forms as the New York Times Information Bank.

Combine such computer systems with the kinds of telecommunication systems we have been discussing here and the product will be a truly all-purpose information system. With it one will be able to

- Exchange messages and "letters" with other people and with computers
- Hold teleconferences
- Do computations
- Jointly write and edit articles and journals
- Collect files of important documents
- Search files
- Keep personal diaries
- Design and write specifications for equipment and new systems
- Teach classes
- Conduct interviews
- Order equipment.

And the list could go on and on.

One of the most important characteristics of such advanced systems is that all these functions would be independent of time and space. Conferences, interviews, classes, and other interactions could be carried out on opposite sides of the world as easily as they could be conducted next door. Even more important is that such systems would make it possible to draw upon the collective intelligences of man and computer. Indeed, one can easily imagine that the contributions of man and computer would be so commingled that one would never be sure whether a thought, idea, suggestion, or solution came from a man or computer.

To make that dream reality will require a great deal of imaginative and careful research on the ways in which telecommunications and computer technologies can be most effectively married to satisfy their ultimate users. Only after we have done that research will we be able to achieve the complete "man-computer symbiosis" that was so confidently predicted nearly two decades ago [39] but that has remained so elusively and so tantalizingly beyond our grasp.

REFERENCES

1. R. S. Lewis, *The Voyages of Apollo: The Exploration of the Moon*, Quadrangle, the New York Times Book Company, New York, 1974.
2. C. Cherry, *World Communication: Threat or Promise?* Wiley-Interscience, New York, 1971.
3. J. C. Madden, "The Wired World," in *McGraw-Hill Yearbook of Science and Technology*. McGraw-Hill, New York, 1973, pp. 56–65.
4. C. V. Newsom, "Communications Satellites: A New Hazard for World Cultures," *Educ. Broadcast. Rev.* 7, 77–85 (1973).
5. M. Turoff, "Human Communication Via Data Networks," *Comput. Decis.* **5(1)**, 25–29 (1973).
6. R. Amara and J. Vallee, "Forum: A Computer Based System To Support Interaction Among People," *Proc. Int. Fed. Inform. Process. Congr*, 1974, pp. 1052–1056.
7. F. W. Frey, "Communication and Development," Fig. B. 3, p. 445, in I. de Sola Pool, F. W. Frey, W. Schramm, N. Maccoby, and E. B. Parker, eds. *Handbook of Communication*, Rand McNally College Publishing Company, Chicago, 1973.
8. J. J. O'Neill, J. T. Nocerino, and P. Wolcoff, *Benefits and Problems of Seven Exploratory Telemedicine Projects*, Report No. MTR-6787, The MITRE Corporation, Washington, D.C., 1975, p. 15.
9. C. D. Mackenzie, *Alexander Graham Bell: The Man Who Contracted Space*, Houghton Mifflin, Boston, 1928, pp. 143–144.
10. L. S. Howeth, *History of Communications-Electronics in the United States Navy*, U.S. Government Printing Office, Washington, D.C., 1963, p. 11.
11. A. Chapanis, "Prelude to 2001: Explorations in Human Communication," *Amer. Psychol.* **26**, 949–961 (1971)..
12. M. Argyle, *Social Interaction*, Methuen, London, 1969.
13. A. Mehrabian, *Nonverbal Communication*, Aldine-Atherton, Chicago, 1972.
14. M. Argyle, V. Salter, H. Nicholson, M. Williams, and P. Burgess, "The Communication of Inferior and Superior Attitudes by Verbal and Non-Verbal Signals," *Brit. J. Soc. Clin. Psychol.* **9**, 222–231 (1970).
15. Z. Rubin, "Measurement of Romantic Love," *J. Person. Soc. Psychol.* **16**, 265–273 (1970).
16. S. Albert and J. M. Dabbs, Jr., "Physical Distance and Persuasion," *J. Person. Soc. Psychol.* **15**, 265–270 (1970).
17. A. A. L. Reid, *The RMT Teleconference System*, Paper No. P/72024/RD, Communications Studies Group, Joint Unit for Planning Research, University College, London, 1972.
18. O. M. Watson and T. D. Graves, "Quantitative Research in Proxemic Behavior," *Amer. Anthropol.* **68**, 971–985 (1966).
19. J. A. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunication*, Wiley International, Chichester, in press.
20. A. Chapanis, R. B. Ochsman, R. N. Parrish, and G. D. Weeks, "Studies in Interactive Communication: I. The Effects of Four Communication Modes on the Behavior of Teams During Cooperative Problem-Solving," *Hum. Factors* **14**, 487–509 (1972).
21. R. Meier, *The Communications Theory of Urban Growth*, M.I.T. Press, Cambridge, Mass., 1962, pp. 69, 71.
22. D. Cartwright and A. Zander, eds., *Group Dynamics: Research and Theory*, 2d ed., Row, Peterson, Evanston, Ill. (1960).
23. R. F. Bales, "Task Roles and Social Roles in Problem-Solving Groups," in E. E. Maccoby, T. M. Newcomb, and E. L. Harley, eds, *Readings in Social Psychology*, 3d ed., Henry Holt, New York, 1958.
24. P. G. Zimbardo, "The Human Choice: Individuation, Reason and Order Versus Deindividuation, Impulse, and Chaos," in W. J. Arnold and D. Levine, eds., *Nebraska Symposium on Motivation*, Vol. 17, University of Nebraska Press, Lincoln, 1969.
25. R. Hough, *Teleconferencing Systems: The State of the Art and a Preliminary Evaluation*, National Science Foundation, Washington, D.C., 1976.
26. D. W. Conrath, E. V. Dunn, J. N. Swanson, and P. D. Buckingham, "A Preliminary Evaluation of Alternative Telecommunication Systems for the Delivery of Primary Health Care to Remote Areas," *IEEE Trans. Commun.* **Com-23**, 1119–1126 (1975).
27. G. D. Weeks, M. J. Kelly, and A. Chapanis, "Studies in Interactive Communication: V. Cooperative Problem Solving by Skilled and Unskilled Typists in a Teletypewriter Mode," *J. Appl. Psychol.* **59**, 665–674 (1974).
28. J. A. Short, "Effects of Medium of Communication on Experimental Negotiation," *Hum. Relat.* **27**, 225–234 (1974).
29. E. Williams, "Coalition Formation Over Telecommunications Media," *Europe. J. Soc. Psychol.* **5**, 503–507 (1975).

30. R. Pye, B. Champness, H. Collins, and S. Connell, *The Description and Classification of Meetings*, Paper No. P/73160/PY, Communications Studies Group, Joint Unit for Planning Research, University College, London, 1973.
31. I. E. Morley and G. M. Stephenson, "Interpersonal and Inter-Party Exchange: A Laboratory Simulation of an Industrial Negotiation at the Plant Level," *Brit. J. Psychol.* **60**, 543–545 (1969).
32. A. Chapanis and C. M. Overbey, "Studies in Interactive Communication: III. Effects of Similar and Dissimilar Communication Channels and Two Interchange Options on Team Problem Solving," *Percep. Mot. Skills* **38**, 343–374 (Monograph Supplement 2-V38) (1974).
33. R. B. Ochsman and A. Chapanis, "The Effects of 10 Communication Modes on the Behavior of Teams During Co-operative Problem-Solving," *Int. J. Man-Mach. Stud.* **6**, 579–619 (1974).
34. A. E. Casey-Stahmer and M. D. Havron, "Planning Research in Teleconference Systems," Report No. HSR-RR-73/10-St-X, Human Sciences Research, Inc., McLean, Va., Sep. 28, 1973.
35. E. C. Berkeley, *Giant Brains or Machines That Think*, Wiley, New York, 1949.
36. J. Weizenbaum, "Contextual Understanding by Computers," pp. 334–348, in Z. W. Pylyshyn, ed., *Perspectives on the Computer Revolution*, Prentice-Hall, Englewood Cliffs, N.J., 1970.
37. V. Bush, "As We May Think," *Atlantic Monthly* **176**, 101–108 (1945).
38. J. C. R. Licklider, *Libraries of the Future*, M.I.T. Press, Cambridge, Mass., 1965.
39. J. C. R. Licklider, "Man-Computer Symbiosis," *IRE Trans. Hum. Factors Electron.* **HFE-1**, 4–11 (1960).



After twelve years on the staff of the Oregon Research Institute, where he co-ordinated the program in judgment and decision making, Paul Slovic recently became a co-founder of Decision Research. Dr. Slovic is a member of the editorial boards of the *Journal of Experimental Psychology*, the *Journal of Experimental Research in Personality*, and *Organizational Behavior and Human Performance*. He received a B.A. from Stanford University and a Ph.D. from the University of Michigan, where he worked for 2 years at the Engineering Psychology Laboratory.

TOWARDS UNDERSTANDING AND IMPROVING DECISIONS

Paul Slovic

Decision Research
Eugene, Ore.

The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world—or even for a reasonable approximation to such objective rationality.

Herbert Simon [1]

The rise of automation in military and defense contexts and the increased potency of modern weaponry have changed radically the hierarchy of needed human skills. Strength and motor performance have become less important. So have perceptual skills although these will never be unimportant. Modern technology has made intellectual skills, especially those of judgment and decisionmaking, the crucial human elements.

The difficulties of decisionmaking are usually blamed on the inadequacy of the available information; therefore, much technological sophistication has been mobilized to remedy this problem. Devices proliferate to supply the decisionmaker with an abundance of data—consider, for example, the sophisticated electronic sensors in aircraft and satellites that relay great quantities of strategic data for military intelligence.

It has become apparent, however, tha even the best attainable information often leaves us with a mass of uncertainties and doubts. Roberta Wohlstetter's analysis of the crises at Cuba and Pearl Harbor illustrates the problem. She notes

> . . . in both the Pearl Harbor and Cuban crises there was plenty of information. But in both cases, . . . the data were ambiguous and incomplete. There was never a single, definitive signal that said, "Get ready, get set, go!" but rather a number of signals that, when put together, tended to crystallize suspicion. The true signals were always embedded in the noise or irrelevance of false ones. [2]

It has become evident that a key element in decisionmaking is the ability to interpret and integrate information items, the reliability and validity of which are imperfect. Typically, decisionmakers are left to their own devices. More likely than not they will proceed in much the same manner that has been relied upon since antiquity—by following their intuition.

But things have begun to change. Specialists from many disciplines have begun to study information processing and decisionmaking. Their efforts, and mine in this paper, center around two broad questions: "What are decisionmakers doing?" and "What should they be doing?" The first is a psychological problem, one of understanding how people make decisions and relating this knowledge to the mainstream of cognitive psychology. The second problem is a practical one and involves the attempt to make decisionmaking more effective and efficient.

AIMS AND ORGANIZATION OF THE PAPER

Decisionmakers of the future will be supplied with many techniques, simple and complex, to

help them. The purpose of this paper is to preview these decision-aiding technologies and to outline some of the behavioral considerations underlying their development and their potential for successful application.

The paper begins with an overview of research that describes the shortcomings of unaided decisions. This work, much of it sponsored by the Office of Naval Research, has led to the sobering conclusion that, in the face of uncertainty, man may be an intellectual cripple, whose intuitive judgments and decisions violate many of the fundamental principles of optimal behavior. These intellectual deficiencies underscore the need for decision-aiding techniques; the prospects for such techniques are outlined in the second half of the paper.

A NEW IMAGE OF HUMAN CAPABILITIES

The traditional view of human beings' higher mental processes assumes that we are intellectually gifted creatures. Shakespeare referred to man as ". . . *noble in reason, infinite in faculties* . . . the beauty of the world, the paragon of animals." A more recent expression of this esteem was provided by economist Frank Knight: "We are so built that what seems reasonable to us is likely to be confirmed by experience or we could not live in the world at all." [3] Given appropriate information on which to take action, why should such a creature need decision aids?

The answer lies with a rather different picture of human capabilities that has emerged out of the computer era and its concern for information processing by man and machine. Miller [4] in his famous study of classification and coding, showed that there are severe limitations on people's ability to attend to and process sensory signals. About the same time, close observation of performance in concept formation tasks led Bruner, Goodnow, and Austin [5] to conclude that their subjects were experiencing a condition of "cognitive strain" and were trying to reduce it by means of simplification strategies. The processing of conceptual information is currently viewed as a serial process that is constrained by limited short-term memory and a slow storage in long-term memory [6].

In the study of decisionmaking, too, the classic view of behavioral adequacy, or rationality, has been challenged on psychological grounds. For example, Simon's theory [1] of "bounded rationality" asserts that cognitive limitations force decisionmakers to construct simplified models in order to cope with their problems. Simon argued that the decisionmaker

> . . . behaves rationally with respect to this [simplified] model, and such behavior is not even approximately optimal with respect to the real world. To predict his behavior, we must understand the way in which this simplified model is constructed, and its construction will certainly be related to his psychological properties as a perceiving, thinking, and learning animal. [1]

Recent laboratory experiments have provided dramatic support for the concept of bounded rationality and have demonstrated its impact in a variety of judgmental and decisionmaking situations. This research, to be reviewed below, is organized around several basic problems of concern to decisionmakers. First, they need to know what will happen or how likely it is to happen, and their use of information to answer these questions gets them involved in processes of inference, prediction, subjective probability, and diagnosis. They must also evaluate the worth of objects, and this often requires them to combine information from several component attributes of the object into an overall judgment. Finally, they are called upon to integrate their opinions about probabilities and values into the selection of some course of action. What is referred to as "weighing risks against benefits" is an example of the latter combinatorial process.

Studies of Probabilistic Information Processing

Because of the importance of probabilistic reasoning to decisionmaking, a great deal of recent experimental effort has been devoted to understanding how people perceive and use the probabilities of uncertain events. By and large, this research indicates that people systematically

violate the principles of rational decisionmaking when judging probabilities, making predictions, or otherwise attempting to cope with probabilistic tasks. Frequently these violations can be traced to the use of judgmental heuristics or simplification strategies [7]. These heuristics may be valid in some circumstances but in others they lead to biases that are large, persistent, and serious in their implications for decisionmaking.

Misjudging Sample Implications—One example of the errors people make when dealing intuitively with probabilistic phenomena comes from a study by Tversky and Kahneman [8] who analyzed the kinds of decisions psychologists make when planning scientific experiments and interpreting their results. Despite extensive formal training in statistics, psychologists usually rely on their educated intuitions when they make decisions about how large a sample of data to collect or whether they should repeat an experiment to make sure their results are reliable. After questioning a large number of psychologists about their research practices and studying the designs of experiments reported in psychological journals, Tversky and Kahneman concluded that these scientists seriously underestimated the error and unreliability inherent in small samples of data. As a result, they (1) had unreasonably high expectations about the replicability of results from a single sample, (2) had undue confidence in early results from a few subjects, (3) gambled their research hypotheses on small samples without realizing the extremely high odds against detecting the effects being studied, and (4) rarely attributed any unexpected results to sampling variability because they found a causal explanation for every observed effect.

Tversky and Kahneman summarized these results by asserting that people's intuitions seemed to satisfy a "law of small numbers," which means that the "law of large numbers" applies to small samples as well as to large ones. The "law of large numbers" says that very large samples will be highly representative of the population from which they are drawn. For the scientists in this study, small samples were also expected to be highly representative of the population. Since knowledge of logic or probability theory did not make the scientist any less susceptible to these cognitive biases, Tversky and Kahneman concluded that the only effective precaution is the use of formal statistical procedures, rather than intuition, to design experiments and evaluate data.

In a related study using Stanford undergraduates as subjects, Kahneman and Tversky [9] found that many of these individuals did not understand the fundamental principle of sampling—that the variance of a sample decreases as the sample size gets larger. They concluded that "For anyone who would wish to view man as a reasonable intuitive statistician, such results are discouraging."

Errors of Prediction—Kahneman and Tversky [10] contrasted the rules that determined people's intuitive predictions with the normative principles of statistical prediction. Normatively, the prior probabilities or base rates, which summarize what we knew before receiving evidence specific to the case at hand, are relevant even after specific evidence is obtained. In fact, however, people seem to rely almost exclusively on specific information and neglect prior probabilities.

For example, Kahneman and Tversky asked subjects to judge the likelihood that an individual, Tom W., is a graduate student in a particular field of specialization. The judges in this study were all graduate students in psychology. The only information they had available to them was the following brief description written several years earlier by a psychologist on the basis of some projective tests:

> Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people, and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.
>
> Tom W. is currently a graduate student. Please rank the following nine fields of graduate specialization in order of the likelihood that Tom W. is now a student in that field. Let rank 1 be the most probable choice.

____ Business Administration
____ Computer Sciences
____ Engineering
____ Humanities and Education
____ Law
____ Library Sciences
____ Medicine
____ Physical and Life Sciences
____ Social Science and Social Work

In this study, people ranked the graduate programs on the basis of the similarity between the brief description and typical student in each program. What was remarkable was that the prior probabilities, as determined by the base rates for these graduate programs, had no influence whatsoever upon the judgments. Computer Sciences and Engineering were judged to be the most probable fields for Tom W., even though these fields have relatively few students in them. This is especially surprising considering the fact that the judges recognized the thumbnail personality sketch as having little or no validity. In addition, all of these judges had been exposed to the notion of base-rate prediction in their statistical training, and they used the base rate in a condition where no other information was provided. The important result here is the apparent inability of the judges to integrate the similarity ordering with the base-rate information in a situation where base rate should have been predominant. In other words, the judges knew the description was of low validity and they knew that base rates differed, yet they were unable to put this knowledge into practice. As a result, their judgments did not properly reflect their underlying beliefs.

Another normative principle is that the variance of one's predictions should be sensitive to the validity of the information on which the predictions are based. If validity is not perfect, predictions should be regressed toward some central value. Furthermore, the lower the validity of the information on which predictions are based, the greater the regression should be. Kahneman and Tversky [10] observed that otherwise intelligent people have little or no intuitive understanding of the concept of regression. They fail to expect regression in many situations when it is bound to occur and, when they observe it, they typically invent complex but spurious explanations. People fail to regress their predictions towards a central value even when they are using information that they themselves consider of low validity.

A third principle of prediction asserts that, given input variables of stated validity, accuracy of prediction decreases as redundancy increases. Kahneman and Tversky [10] found, however, that people have greater confidence in predictions based on highly redundant or correlated predictor variables, since these tend to agree with one another in their implications. Thus, the effect of redundancy on confidence is opposite what it should be.

Availability Bias—Another form of judgmental bias can be traced to the use of the "availability heuristic" [11] whereby an event is judged likely or frequent if it is easy to imagine or recall relevant instances. In life, instances of frequent events are typically easier to recall than instances of less frequent events, and likely occurrences are usually easier to imagine than unlikely ones. Thus, availability is often a valid cue for judging frequency and probability. However, since availability is also affected by subtle factors unrelated to likelihood, reliance on it may result in systematic overestimation of probabilities for familiar, recent, emotionally salient, or otherwise memorable or imaginable events. Evidence to support this contention comes from a study by Slovic, Fischhoff, and Lichtenstein [12] which found that (1) the probabilities of dramatic, well-publicized events such as botulism, tornadoes, motor vehicle accidents, homicides, and cancer were overestimated and (2) unremarkable or less dramatic events such as asthma, diabetes, and emphysema were underestimated. In addition to demonstrating availability bias, this study shows that intelligent individuals do not have valid perceptions about the frequency of hazardous events to which they are exposed.

Anchoring Bias—Bias also occurs when a judge attempts to ease the strain of processing information by following the heuristic device of "anchoring and adjustment." In this process, a natural starting point or anchor is used as a first approximation to the judgment. This anchor is then adjusted to accommodate the implications of additional information. Typically, the adjustment is crude and imprecise and fails to do justice to the importance of additional information. Recent

work by Tversky and Kahneman [7] demonstrates the tendency for adjustments to be insufficient. They asked subjects questions such as "What is the percentage of people in the U.S. today who are age 55 or older?" They gave the subjects starting percentages that were randomly chosen and asked them to adjust these percentages until they reached their best estimate. Because of insufficient adjustment, subjects whose starting points were high ended up with higher estimates than those who started with low values.

Application of the anchoring and adjustment heuristic is hypothesized to produce a bias that occurs when people attempt to calibrate the degree to which they are uncertain about an estimate or prediction. Specifically, in a number of studies subjects were given almanac questions such as the following:

How many foreign cars were imported into the United States in 1968?

(a) Make a high estimate such that you feel there is only a 1% probability the true answer would exceed your estimate.

(b) Make a low estimate such that you feel there is only a 1% probability the true answer would be below this estimate.

In essence, the person is being asked to estimate an interval such that there is a 98% chance that the true answer will fall within the interval. The spacing between the high and low estimates is an expression of the person's uncertainty about the quantity in question. We cannot say that this single pair of estimates is right or wrong. However, if the person were to make many such estimates or if a large number of persons were to answer this question, we should expect the range between upper and lower estimates to include the truth about 98% of the time—if the subjective probabilities were unbiased. What is typically found, however, is that the 98% confidence range fails to include the true value from 25 to 40% of the time, across many subjects answering many kinds of almanac questions [13]. In other words, subjects' confidence bands are much too narrow, given their state of knowledge. This bias persists even when subjects are given feedback about their overly narrow confidence bands and are urged to widen the bands on a new set of estimation problems.

These studies indicate that people believe they have a much better picture of the truth than they really do. Why this happens is not entirely clear. It has been hypothesized [14] that people approach these problems by searching for a calculational scheme or algorithm by which to make a best estimate. They may then adjust this estimate up and down to get a 98% confidence range. For example, in answering the above question, one might proceed as follows:

> I think there were about 180 million people in the U.S. in 1968; there is about one car for every three people thus there would have been about 60 million cars; the lifetime of a car is about 10 years, this suggests that there should be about 6 million new cars in a year but since the population and the number of cars is increasing let's make that 9 million for 1968; foreign cars make up about 10% of the U.S. market, thus there were probably about 900,000 foreign imports; to set my 98% confidence band, I'll add and subtract a few hundred thousand cars from my estimate of 900,000.

People's estimates seem to assume that their computational algorithms are 100% correct. However, there are two sources of uncertainty that plague these algorithms. First, there is uncertainty associated with every step in the algorithm and there is uncertainty about the algorithm itself. That is, the whole calculational scheme may be incorrect. It is apparently quite difficult to carry along these several sources of uncertainty and translate them intuitively into a confidence band. Once the "best guess" is arrived at as an anchor (e.g., the 900,000 figure above), the adjustments are insufficient in magnitude, failing to do justice to the many ways in which the estimate can be in error.

The research just described implies that our estimates may be grossly in error—even when we attempt to acknowledge our uncertainty. This may have profound implications for many important judgments.

Hindsight Bias—A series of experiments by Fischhoff [15, 16, 17] has examined the phenomenon of hindsight. Fischhoff found that being told some event has happened increases our feeling that it was inevitable. We are unaware of this

effect, however, and tend to believe that this inevitability was apparent in foresight, before we knew what happened. In retrospect, we tend to believe that we (and others) had a much better idea of what was going to happen than we actually did have. Fischhoff [16] shows how such misplaced belief that we "knew it all along" can seriously prejudice the evaluations of decisions made in the past and limit our ability to learn by experience. Hindsight bias may also lead us to underestimate the informativeness of facts gleaned from intelligence operations [18] and research studies [19].

Overconfidence—An important criterion for evaluating judgments of probability is their degree of calibration. A probability assessor is well calibrated if, for all statements assigned a given probability (e.g., the probability is 0.65 that "Rumania will maintain current relations with the People's Republic of China"), the proportion that is true is equal to the probability assigned. For example, if you are well calibrated, then across the many statements to which you assign a probability of 0.80, 80% of them should turn out to be true. In the past few years, numerous laboratory and real-world experiments have studied calibration [13, 20]. Across a wide variety of tasks and subjects, one finding has consistently occurred. People are overconfident; they tend to estimate much higher probabilities than are warranted. Slovic, Fischhoff, and Lichtenstein [12] studied cases of extreme overconfidence in a task in which people judged the odds that their answers to general knowledge questions were correct. Subjects were wrong frequently on answers they judged almost certain (odds of 50:1 or greater) to be correct. Feelings of certainty were so strong that subjects were willing to bet on the correctness of their knowledge. Because of their great overconfidence, the bets they accepted were disadvantageous to them and they lost considerable money. The psychological basis for unwarranted certainty seems to derive from the fact that people reach conclusions about answers by reconstructing their knowledge from fragments of information, much as a paleontologist infers the appearance of a dinosaur from fragments of bone. For example, a person who is "absolutely certain" that the potato is native to Ireland and not Peru may base this judgment on the ready association "Irish potato" and the knowledge that a great potato famine caused mass emigration from Ireland to America. Unfortunately, we appear to be insufficiently critical of the assumptions and reasoning on which our opinions are based—indeed we typically feel that we have direct access to our knowledge and thus we are unaware that we are making inferences. The potato, by the way, is native to Peru.

Problems of Decisionmaking

Consider next the integration of information from diverse sources into an overall judgment of value or a decision about a course of action. Here, too, we observe that cognitive limitations lead people to take actions that are inconsistent with their underlying values and opinions.

The failure of one's decisions to reflect personal opinions can be considered one of the most fundamental aspects of nonoptimal decisionmaking. One example of this comes from an experiment by Lichtenstein and Slovic [21] conducted on the floor of the Four Queens Casino in Las Vegas. Consider the following pair of gambles used in the experiment:

Bet A
11/12 change to win 12 chips
1/12 chance to win 24 chips

Bet B
2/12 chance to win 79 chips
10/12 chance to lose 5 chips

where the value of each chip has been previously fixed at, say, 25¢. Notice that bet A has a much better chance of winning, but bet B offers a higher winning payoff. Subjects were shown many such pairs of bets. They were asked to indicate, in two ways, how much they would like to play each bet in a pair. First they made a simple choice, A or B. Later they were asked to assume they owned a ticket to play each bet, and they were to state the lowest price for which they would sell this ticket.

Presumably, these selling prices and choices are both governed by the same underlying quality, the subjective attractiveness of each gamble. Therefore, people should state a higher selling price for the gamble that they prefer in the choice

situation. However, the results indicated that subjects often chose one gamble, yet stated a higher selling price for the other gamble. For the particular pair of gambles shown above, bets A and B were chosen about equally often. However, bet B received a higher selling price about 88% of the time. Of the subjects who chose bet A, 87% gave a higher selling price to bet B, thus exhibiting an inconsistent preference pattern.

What accounts for the inconsistent pattern of preferences? Lichtenstein and Slovic conclude that people use different cognitive strategies for *setting prices than for making* choices. People choose bet A because of its good odds, but they set a higher price for B because of its large winning payoff. Specifically, it was found that, when making pricing judgments, people who find a gamble basically attractive use the amount to win as a natural starting point. They then adjust the amount to win downward to take into account the less-than-perfect chance of winning and the fact that there is some amount to lose as well. Typically, this adjustment is insufficient and that is *why winning payoffs lead people to set prices* that are inconsistent with their choices. Because the pricing and choice responses are inconsistent, it is obvious that at least one of these responses does not accurately reflect what the decisionmaker believes to be the most important attribute in a gamble.

A "compatibility" effect seems to be operating here. Since a selling price is expressed in terms of monetary units, subjects apparently found it easier to use the monetary aspects of the gamble to *produce this type of response. Such a bias did* not exist with the choices, since each attribute of one gamble could be directly compared with the same attribute of the other gamble. With no reason to use payoffs as a starting point, subjects were free to use any number of strategies to determine their choices. The overdependence on payoff cues when pricing a gamble suggests a general hypothesis to the effect that the compatibility or commensurability between a dimension of information and the required response affects the ease with which that information can be used and, ultimately, its importance in determining the response. This hypothesis received support in an experiment by Slovic and MacPhillamy [22], who found that dimensions common to each alternative in a choice situation had greater influence on decisions than did dimensions that were unique to a particular alternative. Interrogation of the subjects after the experiment indicated that most did not wish to give more weight to the common dimension and were unaware that they had done so.

The message in these experiments is that the amalgamation of different types of information and different types of values into an overall judgment or decision is a difficult cognitive process and, in our attempts to ease the strain of processing information, we often resort to judgmental strategies that may do an injustice to our underlying values. In other words, even when all the relevant events, probabilities, and outcomes are known and made explicit, as in the gambling situation, subtle aspects of the decision we have to make, acting in combination with our intellectual limitations, may bias the balance we strike among the attributes.

When the decision is not well structured, that is, when all the relevant aspects are not explicitly specified, further difficulties arise. Foremost among these is the neglect of one or more crucial factors whose relevance only becomes apparent, sadly, after the decision has been made. An example of this is provided by Birkin and Ford [23] who examined the after-effects of the "Zero Defects" program. This program, adopted by more than 12 000 industrial firms, attempted to attack the problem of defective workmanship by motivating employees to do the job right the first time. The program was based on the following rationale: "Because of the complexity of today's products and because of the drastic consequences of product failure, management should use all means possible to meet customers' specifications. Human error on the job is not inevitable and employees, if properly motivated, could maintain a desire to get a job done right the first time." Once the program was implemented, many firms discovered they could not live with the consequences of making quality a primary goal. As quality rose, productivity declined, production deadlines were missed, and amounts of spoiled and scrapped goods increased. A high percentage of firms dropped the program.

Random Error—We're all familiar with the effects of random error in activities that involve motor skills—playing golf is one such activity that

comes to mind. Random error is the mysterious lack of control that causes two drives, seemingly executed the same way, to end up in different parts of the fairway. We're less aware that similar lack of control affects our decisionmaking behaviors as well as our golf games. In fact, it's only quite recently that decisions have been studied in a way that illustrates this problem.

Goldberg [24] described the problem of error and unreliability by noting that

> He [the judge] "has his days": Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configuration are not identical. He is subject to all those human frailties which lower the reliability of his judgments below unity.

There are a number of studies demonstrating the presence of random error in the judgments of experts. One of the most significant of these studies was done by Garland [25], who measured the reliability of radiologists as they attempted to detect the presence of lung disease on X-ray films. Garland found that radiologists changed their minds in about 20% of the cases when reading the same film on two separate occasions.

Another example of inconsistency comes from a study of expert horserace handicappers, which Bernard Corrigan and I conducted at the Oregon Research Institute. We were interested, not in horserace predictions but in the stresses caused by information overload. Horseracing provided an appropriate context in which to study this. We expect that the results will generalize to any domain in which the integration of large masses of quantitative information is performed by means of skilled human judgment.

Our judges in this study were eight individuals, carefully selected for their expertise as handicappers. Each judge was presented with a list of 88 variables taken from the horses' past-performance charts. The judges were asked to indicate which five variables out of the 88 they would wish to use when handicapping a race, if they were limited to just five variables. They were then asked to indicate which 10, which 20, and which 40 they would use if 10, 20, or 40 items of information were available.

All the handicappers judged each of 45 races under all four information conditions. First they saw five variables and ranked the top five horses in the race in the order they thought the horses would finish. They then received their preselected 10-variable set and reranked the horses. They then ranked them again using 20 and finally 40 variables. All handicappers had their own personalized set of 5, 10, 20, and 40 variables. Five of the races were repeated at the end of the experiment. By examining a handicapper's two rankings for the same race, we were able to assess the degree of inconsistency in that person's judgment policy.

The results indicated that, on the average, accuracy of prediction was as good with five variables as it was with 10, 20, or 40. However, every handicapper became more confident in the accuracy of the judgments as amount of information increased. Examination of judgments for the repeated races showed that inconsistency increased sharply as the amount of available information increased. With 5 predictors, 22% of the first-place choices were changed on the second replication; with 40 predictors, 39% of the judgments changed. These results should give pause to those who believe they are better off getting as much information as possible prior to making a decision.

Are Important Decisions Biased?

Since the results described previously contradict our traditional image of the human intellect, it is reasonable to ask whether these inadequacies in decisionmaking exist outside the laboratory in situations where experts use familiar sources of information to make decisions that are important to themselves and others.

Much evidence suggests that the laboratory results will generalize. Cognitive biases appear to pervade a wide variety of socially important judgments in which intelligent individuals serve as decisionmakers, often under conditions that maximize motivation and involvement. For example, the subjects studied by Tversky and Kahneman [8] were scientists, highly trained in statistics, evaluating problems similar to those

they faced in their own research. The overdependence on specific evidence and neglect of base rates observed in laboratory studies have also been found among psychometricians responsible for the development and use of psychological tests [26] and among intelligence officers evaluating military information reports [27]. The latter based their evaluations primarily on a report's content, neglecting the base-rate reliability of the report's source. Flood-plain residents misjudge the probability of floods in ways readily explained in terms of availability bias [28, 29]. Roberta Wohlstetter's study [30] of American unpreparedness at Pearl Harbor found the U.S. Congress and military investigators guilty of hindsight bias in their judgment of the Pearl Harbor command staff's negligence. A classic case of the "law of small numbers" is Berkson, Magath, and Hurn's discovery [31] that aspiring lab technicians were expected by their instructors to show greater accuracy in performing blood cell counts than was possible given sampling variation. These instructors marveled that the best students (those who would not cheat) had the greatest difficulty in producing acceptable counts. *Overconfidence* has been observed in intelligence analysts' probability estimates for such events as a coup in a particular country, the shooting down of a reconnaissance plane, or an arms shipment from one country to another [32].

The anchoring and insufficient adjustment that Tversky and Kahneman observed with their almanac questions could well contribute to errors that plague projected cost estimates. For example, one congressional study noted that the cost of major weapon systems was running nearly 50% ahead of original estimates. In one case where the original estimate for six submarine rescue vehicles was $18 million, the actual cost was close to $460 million—a value that most certainly would have been viewed as impossible when the original estimates were made. This gigantic overrun, like many others, was blamed on a failure to foresee development problems. The moral seems to be that there are many ways our estimates can go wrong, and it is difficult to incorporate our uncertainty about these possible sources of error into our judgments.

In case studies of policy analyses, Albert Wohlstetter [33] found that American intelligence analysts consistently underestimated Soviet missile strength, a bias possibly due to anchoring.

Finally, I'd like to point out a particularly painful example of anchoring and insufficient adjustment from my own experience. A few years ago a colleague and I agreed to write a chapter for a book. After the project was completed, we were rummaging through our correspondence with the book's editor and were rather dismayed to note the string of optimistic projections and broken promises that is illustrated as follows:

History of the Chapter

On this date	We promised it for this date
Sept. 16, 1968	June 1969
May 1969	End of July 1969
Dec. 1969	End of Jan. 1970
Jan. 1970	Apr. 1970
Apr. 1970	End of June 1970
But we finally sent the first draft	July 24, 1970.

Many of you probably have had the same experience, and we can take some small comfort in a study by Kidd [34] showing that a similar thing happens when the Central Electricity Generating Board in England and Wales attempts to estimate how long it will take to overhaul its equipment.

Comment

One additional implication of the research on people's limited ability to process probabilistic information deserves comment. Most of the discussions of "cognitive strain" and "limited capacity" that are derived from the study of problem solving and concept formation depict a person as a computer that has the right programs but cannot execute them properly because its central processor is too small. The biases due to availability and anchoring certainly are congruent with this analogy. But the misjudgment of sampling variability and the errors of prediction illustrate more serious deficiencies. Here we see that peoples' judgments of important probabilistic phenomena

are not merely biased but are in violation of fundamental normative rules. Returning to the computer analogy, it appears that people lack the correct programs for many important judgmental tasks.

How could it be that we lack adequate programs for probabilistic thinking? Sinsheimer [35] argues that the human brain has evolved to cope with certain very real problems in the immediate, external world and thus lacks the proper framework with which to encompass many conceptual phenomena. Following Sinsheimer's reasoning, it might be argued that we have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty. We are essentially trial-and-error learners, who ignore uncertainty and rely predominantly on habit or simple deterministic rules. When we can afford to learn from our mistakes, this may be a satisfactory way to behave. When we cannot, we must look toward decision aids to help minimize errors of judgment.

DECISION AIDS

Research in both laboratory and field settings strongly supports the view of decision processes as boundedly rational. Given this awareness of our cognitive limitations, what sort of techniques will enhance our capacity for making intelligent decisions?

I have found it useful to consider the repeatability of the task when characterizing decision aids. Near one end of what is really a continuum of repeatability are tasks such as selection or rejection of applicants for jobs. The essential structure of each application (e.g., the types of information available) remains nearly the same from case to case, although the specific details of each application will, of course, change. Toward the other end of the continuum are more unique decisions. The decision to build a supersonic commercial airliner exemplifies this type of problem.

Figure 1 depicts my conception of the relationship between decision repeatability and decision-aiding techniques. When decisions are repeatable they can be handled quite effectively by precise rules or standard operating procedures (SOPs). Although SOPs, such as rules for reordering supplies in an office, have been around for a long time, there are new and powerful variants, bootstrapping and multiattribute utility analysis, that merit discussion here. When predesignated rules are insufficient, computerized information management systems and realistic experience in a simulated decision environment serve as aids. If the decision task is unique, I believe it is important to consider the time available for deliberations prior to action. If the leadtime is long and the decision is important enough, then decision analysis is the relevant aiding technology. If the leadtime is short, I see no recourse other than to rely on educated intuition. These various types of aids will be discussed at length.

	TYPE OF DECISION	
	Unique	*Repeated*
	Long lead time: Decision analysis	Rule-based systems: bootstrapping multiattribute utilities
	Short lead time: Educated intuition	Computer information systems Simulation

Figure 1—Aids for major decisions

Aids for Unique Decision Situations

Decision Analysis—Decision analysis is a general-purpose technology for making decisions when the stakes are high and both time and resources are ample. The roots of decision analysis can be traced to World War II and the need to solve strategic problems in situations in which experience was either costly or impossible to acquire. The technique developed then was labeled "operations analysis" and later became known as "operations research."

During recent years, a number of closely related offshoots of operations research have been applied to decision problems. These include systems analysis and cost-benefit analysis. Systems

analysis is a branch of engineering, whose objective is capturing the interactions and dynamic behavior of complex systems. Cost-benefit analysis attempts to quantify the prospective gains and losses from some proposed action, usually in terms of dollars. If the calculated gain from an act or project is positive, it is said that the benefits outweigh the costs, and its acceptance is recommended (see, for example, the application of cost-benefit analysis to the study of auto safety features by Lave and Weber [36].)

What systems analysis and operations research approaches lacked for many years was an effective normative framework for dealing either with the uncertainty in the world or with the subjectivity of decisionmakers' values and expectations. The emergence of decision theory provided the general normative rationale missing from these early analytic approaches.

The objective of decision theory is to provide a rationale for making wise decisions under conditions of risk and uncertainty. It is concerned with prescribing the course of action that will conform most fully to the decisionmaker's own goals, expectations, and values.

Decisions under uncertainty are typically represented by a payoff matrix, in which the rows correspond to alternative acts that the decisionmaker can select and the columns correspond to possible states of nature. In the cells of the payoff matrix are one set of consequences contingent on the joint occurrence of a decision and a state of nature.

Since it is impossible to make a decision that will turn out best in any eventuality, decision theorists view choice alternatives as gambles and try to choose according to the "best bet." In 1738 Bernoulli defined the notion of a best bet as one that maximizes the "expected utility" of the decision. That is, it maximizes the quantity

$$EU(A) = \sum_{i=1}^{n} P(E_i)U(X_i) \qquad (1)$$

where EU(A) represents the expected utility of a course of action which has consequences $X_1, X_2, \ldots, X_n$ depending on events $E_1, E_2, \ldots, E_n$, $P(E_i)$ represents the probability of the ith outcome of that action, and $U(X_i)$ represents the subjective value or utility of that outcome.

A major advance in decision theory came when von Neumann and Morgenstern [37] developed a formal justification for the expected utility criterion. They showed that, if an individual's preferences satisfied certain basic axioms of rational behavior, then that person's decisions could be described as the maximization of expected utility. Savage [38] later generalized the theory to allow the $P(E_i)$ values to represent subjective or personal probabilities.

Maximization of expected utility commands respect as a guideline for wise behavior because it is deduced from axiomatic principles that presumably would be accepted by any rational person. One such principle, that of transitivity, asserts that, if a decisionmaker prefers outcome A to outcome B and outcome B to outcome C, it would be irrational for that person to prefer outcome C to outcome A. Persons who are deliberately and systematically intransitive can be used as "money pumps." You can say to them, "I'll give you C. Now, for a penny, I'll take back C and give you B." Since they prefer B to C, they accept. Next you offer to replace B with A for another penny and again they accept. The cycle is completed by offering to replace A by C for another penny; they accept and are 3¢ poorer, back where they started, and ready for another round.

Applied decision theory assumes that the rational decisionmaker wishes to select an action that is logically consistent with his or her basic preferences for outcomes and feelings about the likelihoods of the events on which those outcomes depend. Given this assumption, the practical problem becomes one of structuring the alternatives and scaling the subjective values of outcomes and their likelihoods so that subjective expected utility can be calculated for each alternative. Another problem in application arises from the fact that the range of possible alternatives is often quite large. Also, each outcome may have multiple facets that must be combined into an overall estimate of worth.

Decision analysis is the result of the merger of decision theory and the sophisticated modeling of decision situations provided by systems analysis. A key element of decision analysis is its emphasis on structuring the decision problem and decomposing it into a number of more elementary problems. In this sense, it attempts a simplification

process that, unlike the potentially detrimental simplifications the unaided decisionmaker might employ, maintains all the essential ingredients that are necessary to make the decision and ensures that they are used in a manner logically consistent with the decisionmaker's basic preferences. Raiffa [39] expresses this attitude well in the following statement:

> The spirit of decision analysis is divide and conquer: Decompose a complex problem into simpler problems, get your thinking straight in these simpler problems, paste these analyses together with a logical glue, and come out with a program for action for the complex problem. Experts are not asked complicated, fuzzy questions, but crystal clear, unambiguous, elemental hypothetical questions.

Decision analysis assumes that all relevant considerations in a decision can be assigned to one or another of four components: initial options, possible consequences, values, and uncertainties. In addition, they can, in principle, be represented in a decision tree. Figure 2 shows one such tree; much simplified, it should be viewed merely as illustrative.

In Figure 2, the United States is represented as considering four courses of action: aiding both Israel and the Arabs, aiding one but not the other, or aiding neither. Depending on what the United States does, the Soviets may or may not choose to aid the Arabs; they are considered most likely (probability of 0.80) to aid them if we aid only the Israelis and least likely to do so (probability of 0.30) if we aid only the Arabs. In any of these eight possible situations, three outcomes are considered to be possible: A Mideast settlement, a continuation of the status quo, or an Arab-Israeli war. Basically, we regard these three outcomes as having values of +27, −12, and −119, respectively. But the cost of materials, transport, and the like to aid either side is −2, which must be added to the values of the outcomes if an aid strategy is adopted.

Of course, the probabilities of the various outcomes depend on the patterns of U.S. and Soviet decisions about aid. For example, a Mideast war is most likely (0.75) if we aid the Israelis and the U.S.S.R. aids the Arabs; it is least likely (0.15) if we aid the Arabs but not the Israelis and the U.S.S.R. aids no one.

The arithmetic is straightforward. Suppose no one provides aid to either side (the bottom branch of the tree). Then the expected or average value of the possible outcomes is calculated as follows:

$$0.35(+27) + 0.35(-12) + 0.30(-119) = -30.5$$

The expected value of not aiding either side regardless of what the U.S.S.R. does combines the weighted expected values of the possible Soviet actions as follows:

$$0.50(-68.4) + 0.50(-30.5) = -49.5$$

All other numbers in Figure 2 are calculated in analogous ways.

The decision rule suggested by Figure 2 is: from the available acts, choose the one that on the average is most desirable (or, as in this example, least undesirable). In the example, the proper choice would be aiding only the Arabs. How much confidence you should put in this conclusion depends, obviously, on the confidence you have in the options and relevant numbers that went into it. The numbers presented here are illustrative only and do not represent any serious attempts at realistic modeling of the options, the probabilities, or the utilities on either side. Serious attempts to model this problem would involve thousands of possible outcomes and would require a computer program for their storage and manipulation.

Beyond its primary role of serving as a method for the logical solution of complex decision problems, decision analysis has additional advantages as well. The formal structure of decision analysis makes clear all the elements, their relationships and their associated weights that have been considered in a decision problem. Because the model is explicit, it can serve an important role in facilitating communication among those involved in the decision process. With a decision problem structured in a decision analytic framework, it is an easy matter to identify the location, extent, and importance of any areas of disagreement and to determine whether such disagreements have any material impact on the indicated decision. In addition, should there be any change in the cir-

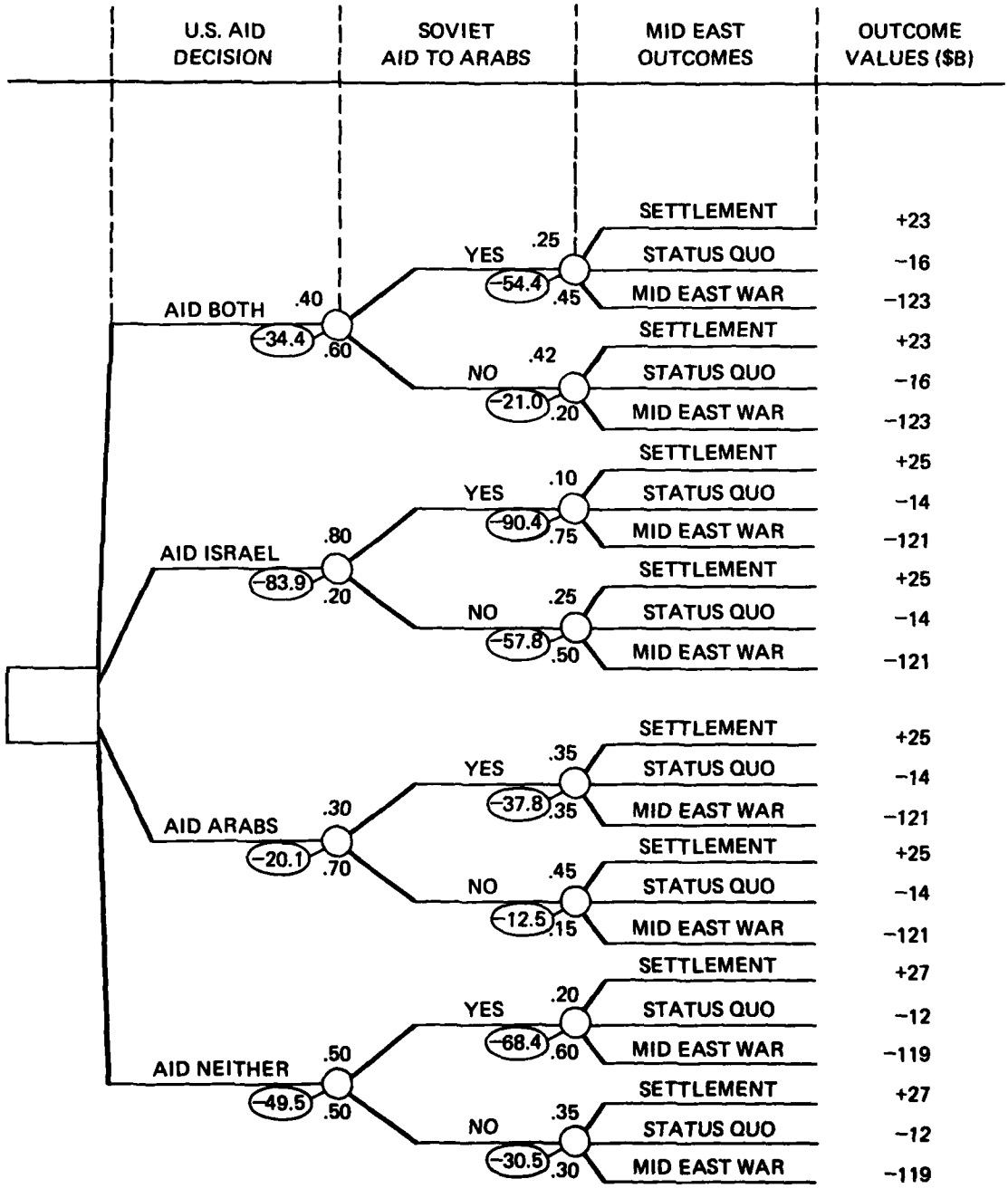


Figure 2—A portion of the decision tree for the U.S. decision on aid in the Middle East

cumstances bearing on a given decision problem, it is fairly straightforward to reenter the existing problem structure to change values or to add or remove problem dimensions as may be indicated.

It should be emphasized that in no sense does decision analysis replace decisionmakers with arithmetic or change the role of wise human judgment in decisionmaking. Rather, it provides an orderly and more easily understood structure that helps to aggregate the wisdom of experts on the many topics that may be needed to make a decision, and it supports skilled decisionmakers by providing them with mathematical techniques to support, supplement, and ensure the internal consistency of their judgments.

Kelly, Peterson, Brown, and Barclay [40] describe a number of applications of decision analysis to military and political problems including decisions about the level of embargo for high-powered computers sold to the Soviet bloc; analyses of U.S. treaty negotiation positions; evaluation of foreign policy strategies aimed at ensuring a stable, expanding supply of oil from Saudi Arabia; and selection among defense contractors proposing to deliver the best system for a fixed price. Other instructive applications include an analysis of whether cloud seeding programs to modify hurricanes should be made operational [41] and a study to determine what type of scientific experiments should be carried out by the first spacecraft on Mars [42]. It is difficult to convey in a summary such as this the depth of thinking and the logic underlying decision analysis. Any brief description necessarily simplifies the analysis and highlights a chief objection to decision analysis in general—the claim that it oversimplifies the situation and thus misleads. Nevertheless, even those who read a complete analysis may have concerns over its validity. Critics argue that such analyses are inevitably constrained by time, effort, and imagination and must systematically exclude many considerations.

A second major objection to decision analysis is the possibility that it may be used to justify and give a gloss of respectability to decisions made on other and perhaps less rational grounds.

Decision analysts counter these attacks by invoking one of their basic tenets—namely, that any alternative must be considered in the context of other alternatives. What, they ask, are the alternatives to decision analysis, and are they any more immune to the criticisms raised above? The analysts point out that traditional modes of decisionmaking are equally constrained by limits of time, effort, and imagination and are even more likely to induce systematic biases (as illustrated previously). Such biases are much harder to detect and minimize than the deficiencies in the explicit inputs to decision analysis. Furthermore, they argue, if some factors are unknown or poorly understood, can traditional methods deal with them more adequately than decision analysis does? Traditional methods also are susceptible to the "gloss of respectability" criticism noted above. We often resort to expertise to buttress our decisions without really knowing the assumptions and logic underlying the experts' judgments. Decision analysis makes these assumptions explicit. Such explicit data are easy for knowledgeable persons to criticize and the explicitness thus focuses debate on the right issues.

Decision analysts would agree that their craft is no panacea, that incomplete or poorly designed analyses may be worse than no analyses at all, and that analysis may be used to "overwhelm the opposition." It seems clear, however, that the main task for the future is not so much to criticize decision analysis but rather to see how it can be used most appropriately.

Educated Intuition

Decision analysis will require extensive further development before it is ready for use in situations in which unique decisions must be taken with little time for deliberation. Thus, the standard method of decision in these situations will continue to be intuition. Given the pitfalls to which intuitive decisions are susceptible, we have little reason to feel comfortable with this prospect. It would seem desirable to prevent such situations from occurring, whenever possible. Every attempt should be made to foresee contingencies and plan for them in advance. Failing that, conservative decisions, which permit one to take fast corrective action to recover from the inevitable mistakes, would seem advisable.

Since we cannot avoid the necessity of making some important decisions intuitively, we should

at least educate decisionmakers to the pitfalls that await the unwary. For example, one should realize the difficulties of using case-specific information to predict low-base-rate (rare) phenomena and, therefore, should take special precautions to ensure adequate consideration of the base rate. When action is contingent on quantitative estimates that may be susceptible to anchoring bias, the wise decisionmaker will obtain multiple estimates, based on differing methods, to allow biases to "cancel out." Since feelings of certainty often lead to bold, decisive action, it is important to alert decisionmakers to the kinds of situations that foster unwarranted confidence. Before taking action in these situations, decisionmakers should scrutinize the assumptions on which their confidence is based and force themselves to consider scenarios that might make their actions look bad (see, for example, Howard, Merkhofer, Miller, and Tani. [43]).

Aids for Repeated Decisions

Bootstrapping—Judgment and decisionmaking have traditionally been viewed as mysterious phenomena, incapable of being described precisely. However, considerable research over the past 15 years has demonstrated that this traditional view is incorrect. The hidden cognitive processes of the judge can be modeled, made explicit, and programmed so that a computer can make judgments that correlate highly with those made by the human. The ability to construct models has important practical consequences. In repeatable decision situations, judges can be replaced by their own models. The benefit from doing this is not merely increased efficiency or freeing the judge for more creative activity. In many cases, the model of the judge makes better predictions than the judge himself! Dawes [44] has termed this phenomenon "bootstrapping."

Before discussing bootstrapping in more detail, let's first consider the sorts of models that might be used to simulate the decisionmaker. These models take two forms, simple and complex. An example of the latter is Clarkson's simulation [45] of the portfolio selection process of a bank's trust investment officer. Clarkson followed the officer around for several months and studied his verbalized reflections as he was asked to think aloud while reviewing past and present decisions. Using these verbal descriptions as a guide, the investment process was translated into a sequentially branching computer program. When the validity of the model was tested by comparing its selections with future portfolios selected by the trust officer, the correspondence between actual and simulated portfolios was found to be remarkably good.

Clarkson's work shows that, given patient and intelligent effort, many of the expert's cognitions can be distilled into a form capable of being simulated by a computer. One application of Clarkson-type modeling has been proposed but not yet implemented by researchers at the department of clinical medicine at a leading medical school. These researchers are concerned with the difficulty of making decisions with regard to medical tests. In addition to being expensive, the tests are sometimes painful and dangerous. The interpretation of the test results is hindered because they are affected by treatment variables and other aspects of the patient's condition. New tests are continually being developed. As a result of these factors, the average physician often does a poor job of selecting and evaluating tests. It has been proposed that sequential decision trees or flow chart models be developed for the world's leading experts on various sorts of tests—tests for thyroid disorder, liver disease, and so forth. These models can then be programmed into a computer and made accessible to practitioners.

There is yet another approach to modeling—a simpler one that provides less of a sequential analysis and more of a quantified descriptive summary of the way that a decisionmaker weights and combines information from diverse sources. This approach aims to develop a mathematical model of the decisionmaker and requires less time and effort on the part of investigator, subject, and computer. It forms a nice compromise between Clarkson's complex, sequentially branching model and the relatively naive approaches of the precomputer era—such as simply asking decisionmakers how they make their judgments. The rationale behind these mathematical models and techniques for building them are reviewed by Slovic and Lichtenstein [46].

The basic approach requires the decisionmaker to make quantitative evaluations of a fairly large

number of cases, each of which is defined by a number of quantified cue dimensions or characteristics. A financial analyst, for example, could be asked to predict the long-term price appreciation for each of 50 securities, the securities being defined in terms of cue factors such as their P/E ratios, corporate earnings growth trend, dividend yield, and so forth. The manner in which the analyst weights these various factors can then be described by fitting a linear equation to the judgments.

The resultant equation would be

$$\hat{J}_{pa} = b_1X_1 + b_2X_2 + \ldots b_kX_k \qquad (4)$$

where J_{pa} = predicted judgment of price appreciation; $X_1, X_2 \ldots X_k$ are the quantitative values of the defining cue factors (i.e., P/E ratios, earnings, and so forth); and $b_1, b_2 \ldots b_k$ are the weights given to the various factors in order to maximize the multiple correlation between the predicted judgments and the actual judgments. These weights are assumed to reflect the relative importance of the factors for the analyst. Eq. (4) is known as the linear model.

Psychologists have found linear equations to be remarkably successful in modeling such diverse phenomena as psychiatric and medical diagnoses, and judgments of job performance, graduate school applicants, suicide risk, financial soundness of businesses, price increases of stocks, Air Force cadets, theatrical plays, and trout streams; political scientists have found linear models useful for describing judicial decision processes in workman's compensation and civil liberties court cases [46, 47]. Even U.S. senators have been modeled and their roll-call votes predicted [48].

More complex, nonlinear, judgmental processes can be modeled by including exponential terms (x^2, x^3, etc.) or cross product terms (e.g., $x_1 x_2$) into the judge's equation. However, nonlinear processing typically accounts for only a small fraction of the predictable variance in human judgments. Most of the variance is accounted for by linear equations, whose coefficients have provided useful descriptions of the judges cue-weighing policies and have pinpointed the source of inter-judge disagreement and nonoptimal cue use [49].

Why do linear models do so well? Dawes and Corrigan [50] have observed that in most judgment situations (a) the predictor variables are monotonically related to the criterion being judged (or can easily be rescaled to be monotonic) and (b) there is error in the predictors and the judgments. They demonstrated that these conditions practically ensure good fits by linear models.

Now that we've examined the ways that decisionmakers can be modeled, let's look again at bootstrapping. The rationale behind it is quite simple. As noted earlier in the discussion of random error, human judgment often lacks reliability. Goldberg [24] observed:

> . . . if the judge's reliability is less than unity, there must be error in his judgments—error which can serve no other purpose than to attenuate his accuracy. If we could . . . [eliminate] the random error in his judgments, we should thereby increase the validity of the resulting predictions.

A model captures the judge's weighting policy and applies it consistently. If there is some validity to this policy to begin with, filtering out the error via the model should increase accuracy. Of course, bootstrapping preserves and reinforces any misconceptions or biases that the judge may have. Implicit in the use of bootstrapping is the assumption that these biases will be less detrimental to performance than the inconsistencies of unaided human judgment.

Bootstrapping has been explored independently by a number of different investigators [46]. One particularly noteworthy demonstration comes from a study of a graduate student admissions committee by Dawes [44]. Dawes built a regression equation to model the average judgment of the four-man committee. The predictors in the equation were overall undergraduate grade point average, quality of the undergraduate school, and a score from the Graduate Record Examination. To evaluate the validity of the model and the possibility of bootstrapping, Dawes used it to predict the average committee rating for his sample of 384 applicants. He found that it was possible to find a cutting point on the distribution of predicted scores such that no one who scored below that point was invited by the admissions committee. Fifty-five percent of the

applicants scored below this point, and thus could have been eliminated by a preliminary screening without doing any injustice to the committee's actual judgments. Furthermore, the weights used to predict the committee's behavior were better than the committee itself in predicting later faculty ratings of the selected students. In a cost-benefit analysis, Dawes estimated that the use of such a linear model to screen applicants to the nation's graduate schools could result in an annual saving of about $18 million worth of professional time.

The potential of judgment modeling for facilitating military and defense decisions is unlimited. One such application has been described by Kelly and Peterson [51] who were concerned with assessing the value and expense of the information collected by the various offices of the Defense Attaché System. Applications to selection decisions are obvious and a little thought turns up many other possibilities. Consider, for example, the task of improving a submarine commander's ability to know when he had been detected by the enemy. It may be possible to model experienced commanders who are expert at this judgmental task. The essence of their model can be communicated to trainees or used as the basis for constructing detection aids.

Other Decision Rules—The linear regression model describes the weighting system implicit in the decisionmaker's behavior. One disadvantage to this approach is that the decisionmaker, perhaps because of cognitive limitations, may not be weighting information in the desired way. Another disadvantage is that it is not always feasible to obtain the large number of judgments necessary for building the model. These difficulties can be overcome by the use of a multiattribute utility (MAU) model that explicitly states the desired weights for each factor in order to produce some overall judgment. For example, one might wish to define the relative importance of variable X to variable Y as 2:1 rather than inferring the values from someone's judgments. MAU procedures are gaining widespread acceptance as rule-based methods for combining component dimensions into an overall evaluation. For a more detailed discussion of this methodology see Fischer [52], von Winterfeldt and Fischer [53], or Slovic, Fischhoff, and Lichtenstein [47].

Multiattribute utility procedures employ predetermined rules to integrate various value components. Another form of predetermined rule employs the notion of a threshold. Of particular importance is the probability threshold whereby a prior decision analysis of the sort described earlier determines that action X should be taken if the probability of event E is less than some threshold value but action Y should be taken otherwise. A detailed report of the use of probability thresholds for naval command decisions is presented by Brown, Peterson, Shawcross, and Ulvila [54].

Information Control Systems—Of course, not all repeatable decisions can be handled by rules. When the human element is necessary, performance can be facilitated by computer-based information systems for storing, modifying, retrieving, and displaying data and for performing various sorts of symbolic and arithmetic manipulations. One such system called AESOP (An Evolutionary System for On-line Planning) is described by Doughty and Feehrer [55].

In one experimental test of AESOP involving allocation of tactical aircraft to various missions, planners were required to make decisions which represented an optimal tradeoff between several criteria, including time over target, minimization of use of recycled aircraft, and minimization of total flying time. Performance of planners assisted by AESOP was superior to that of those who were unassisted. AESOP provided no formal procedures or rules to aid the decisionmaker. However, its concise displays appeared to help planners comprehend the extent to which their resources would be strained and, therefore, enabled them to develop a better "feel" for their plans.

Simulation—One of the most extensively developed methods for sharpening decision performance is that of simulation. Simulation places the decisionmaker in situations that are similar in certain important aspects to those they are likely to encounter in the real world. Simulation has the advantage of exposing the decisionmaker to a rich variety of situations in which the consequences of error are not catastrophic. Performance can be evaluated and immediate feedback provided. On the negative side, simulations must be carefully designed to present the critical aspects of the real decision if proper transfer is to be obtained. For

further discussion of simulation approaches see Abt [56], Driver and Hunsaker [57], and a review by Nickerson and Feehrer [58].

Future Work

Decision-aiding technologies are still in an early stage of development. Thus, although decision analysis is undoubtedly the wave of the future, many problems need to be resolved before we can reap its full benefits.

First of all we need to develop techniques for structuring the decision problem. The logic of decision theory cannot be applied until the alternatives, critical events, and outcomes are specified. We need algorithms for accomplishing this and for simplifying the large, complex decision trees that may result. Crisis situations, where stakes are high, time is short, and the alternatives and information continually changing, pose particularly difficult structuring problems.

Subjective judgments of probability and value are essential inputs to decision analyses. We still do not know the best ways to elicit these judgments. Now that we understand many of the biases to which judgments are susceptible, we need to develop debiasing techniques to minimize their destructive effects. Simply warning a judge about a bias may prove ineffective. Like perceptual illusions, many biases do not disappear upon being identified. It may be necessary to (a) restructure the judgment task in ways that circumvent the bias, (b) use several different methods allowing opposing biases to cancel one another, or (c) correct the judgments externally, based on an estimate of the direction and strength of the bias.

Decision aids should be easy to use. Development of computer graphics techniques is needed to accomplish this goal. Aids also need to be evaluated to determine whether they really are improving quality.

Much progress has been made recently toward understanding judgmental and decisionmaking processes. We need to continue this pursuit of basic knowledge. Simon [59], outlining the historical development of writing, the number system, calculus, and other major aids to thought, provided what seems to me a fitting observation with which to conclude this article:

> All of these aids to human thinking, and many others, were devised without understanding the process they aided—the thought process itself. The prospect before us is that we shall understand that process. We shall be able to diagnose the difficulties of a . . . decision maker . . . and we shall be able to help him modify his problem solving strategies in specific ways.
>
> We have no experience yet that would allow us to judge what improvement in human decision making we might expect from the application of this new and growing knowledge. . . . Nonetheless, we have reason, I think, to be sanguine at the prospect.

REFERENCES

1. H. A. Simon, *Models of Man: Social and Rational*, Wiley, New York, 1957, p. 198.
2. R. Wohlstetter, "Cuba and Pearl Harbor: Hindsight and Foresight," Memo RM 4328-ISA, RAND Corporation, Santa Monica, Calif., 1965, p. 36.
3. F. H. Knight, *Risk, Uncertainty, and Profit*, Houghton-Mifflin, Boston and New York, 1921, p. 227.
4. G. A. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychol. Rev.* 63:81–97 (1956).
5. J. S. Bruner, J. J. Goodnow, and G. A. Austin, *A Study of Thinking*, Wiley, New York, 1956.
6. A. Newell and H. A. Simon, *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, N.J., 1972.

7. A. Tversky and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science* **185**:1124–1131 (1974).
8. A. Tversky and D. Kahneman, "The Belief in the 'Law of Small Numbers'," *Psychol. Bull.* **76**: 105–110 (1971).
9. D. Kahneman and A. Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychol.* **3**:430–454 (1972).
10. D. Kahneman and A. Tversky, "On the Psychology of Prediction," *Psychol. Rev.* **80**:237–251 (1973).
11. A. Tversky and D. Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychol.* **5**:207–232 (1973).
12. P. Slovic, B. Fischhoff, and S. Lichtenstein, *The Certainty Illusion*, Res. Bull. 16–4, Oregon Research Institute, Eugene, Ore., 1976.
13. S. Lichtenstein, B. Fischhoff, and L. Phillips, "Calibration of Probabilities: The State of the Art," in H. Jungermann and G. de Zeeuw, eds., *Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making*, in press.
14. P. Slovic, *From Shakespeare to Simon: Speculations—and Some Evidence—About Man's Ability to Process Information*, Res. Monogr. 12–2, Oregon Research Institute, Eugene, Ore., 1972.
15. B. Fischhoff, "Hindsight-Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty," *J. Exp. Psychol.: Hum. Perception and Performance* **1**:288–299 (1975).
16. B. Fischhoff, "Hindsight: Thinking Backward?" *Psychol. Today* **8**:70–76 (Apr. 1975).
17. B. Fischhoff and R. Beyth, "'I Knew It Would Happen'—Remembered Probabilities of Once-Future Things," *Organ. Behav. and Hum. Performance* **13**:1–16 (1975).
18. B. Fischhoff, "Perceived Informativeness of Facts," *J. Exp. Psychol.: Hum. Perception and Performance* **3**:349–358 (1977).
19. B. Fischhoff and P. Slovic, *On the Psychology of Experimental Surprises*, Res. Bull. 16–2, Oregon Research Institute, Eugene, Ore., 1976.
20. S. Lichtenstein and B. Fischhoff, *Do Those Who Know More Also Know More About How Much They Know?* Res. Bull. 16–1, Oregon Research Institute, Eugene, Ore., 1976.
21. S. Lichtenstein and P. Slovic, "Response-induced Reversals of Preference in Gambling: An Extended Replication in Las Vegas," *J. Exp. Psychol.* **101**:16–20 (1973).
22. P. Slovic and D. J. MacPhillamy, "Dimensional Commensurability and Cue Utilization in Comparative Judgment," *Organ. Behav. and Hum. Performance* **11**:172–194 (1974).
23. S. J. Birkin and J. S. Ford, "The Quantity/Quality Dilemma: The Impact of a Zero Defects Program," pp. 517–529 in J. L. Cochrane and M. Zeleny, eds., *Multiple Criteria Decision Making*, Univ. of South Carolina Press, Columbia, S.C., 1973.
24. L. R. Goldberg, "Man Versus Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences," *Psychol. Bull.* **73**:422–432 (1970).
25. L. H. Garland, "The Problem of Observer Error," *Bull. N.Y. Acad. Med.* **36**:569–584 (1960).
26. P. E. Meehl and A. Rosen, "Antecedent Probability and the Efficacy of Psychometric Signs, Patterns, or Cutting Scores," *Psychol. Bull.* **52**:194–216 (1955).
27. M. G. Samet, "Quantitative Interpretation of Two Qualitative Scales Used to Rate Military Intelligence," *Hum. Factors*, **17**:192–202 (1975).
28. R. W. Kates, "Hazard and Choice Perception in Flood Plain Management," Res. Pap. 78, Dep. of Geography, University of Chicago, Chicago, Ill., 1962.
29. P. Slovic, H. Kunreuther, and G. F. White, "Decision Processes, Rationality, and Adjustment to Natural Hazards," in G. F. White, ed., *Natural Hazards: Local, National, and Global*, Oxford Univ. Press, New York, 1974.
30. R. Wohlstetter, *Pearl Harbor: Warning and Decision*, Stanford Univ. Press, Stanford, Calif., 1962.
31. J. Berkson, T. B. Magath, and M. Hurn, "The Error of Estimate of the Blood Cell Count as Made With the Hemocytometer," *Amer. J. Physiol.* **128**:309–323 (1940).
32. R. V. Brown, A. S. Kahr, and C. Peterson, *Decision Analysis for the Manager*, Holt, Rinehart & Winston, New York, 1974.
33. A. Wohlstetter, "Legends of the Strategic Arms Race, Part I: The Driving Machine," *Strategic Rev.*, 1974, p. 67–92.
34. J. B. Kidd, "The Utilization of Subjective Probabilities in Production Planning," *Acta Psychol.* **34**:338–347 (1970).
35. R. F. Sinsheimer, "The Brain of Pooh: An Essay on the Limits of Mind," *Amer. Sci.* **59**:20–28 (1971).
36. L. B. Lave and W. E. Weber, "A Benefit-Cost Analysis of Auto Safety Features," *Appl. Econ.* **2**:265–275 (1970).

37. J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 3rd ed., Princeton Univ. Press, Princeton, N.J., 1953.
38. L. J. Savage, *The Foundations of Statistics*, Wiley, New York, 1954.
39. H. Raiffa, *Decision Analysis: Introductory Lectures on Choice Under Uncertainty*, Addison-Wesley, Reading, Mass., 1968, p. 271.
40. C. W. Kelly, C. R. Peterson, R. V. Brown, and S. Barclay, "Decision Theory Research," Tech. Prog. Rep. 4, Decisions and Designs, Inc., McLean, Va., 1975.
41. R. A. Howard, J. E. Matheson, and D. W. North, "The Decision to Seed Hurricanes," *Science* 176:1191–1202 (1972).
42. J. E. Matheson and W. J. Roths, "Decision Analysis of Space Projects: Voyager Mars," in R. Howard, J. E. Matheson, and K. L. Miller, eds., *Readings in Decision Analysis*, Stanford Research Institute, Palo Alto, Calif., 1976.
43. R. A. Howard, M. W. Merkhofer, A. C. Miller, and S. N. Tani, "A Preliminary Characterization of a Decision Structuring Process for the Task Force Commander and His Staff," Tech. Rep. MSC-4030, Stanford Research Institute, Palo Alto, Calif., 1975.
44. R. M. Dawes, "A Case Study of Graduate Admissions: Applications of Three Principles of Human Decision Making," *Amer. Psychol.* 26:180–188 (1971).
45. G. P. E. Clarkson, *Portfolio Selection: A Simulation of Trust Investment*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
46. P. Slovic and S. Lichtenstein, "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," *Organ. Behav. and Hum. Performance* 6:649–744 (1971).
47. P. Slovic, B. Fischhoff, and S. Lichtenstein, "Behavioral Decision Theory," *Annu. Rev. Psychol.*, 1977–28, 1–39.
48. H. Wainer, N. Zill, and G. Gruvaeus, "Senatorial Decision Making: II. Prediction," *Behav. Sci.* 18:20–26 (1973).
49. K. R. Hammond, T. R. Stewart, B. Brehmer, and D. O. Steinmann, "Social Judgment Theory," pp. 271-312 in M. F. Kaplan and S. Schwartz, eds., *Human Judgment and Decision Processes*, Academic Press, New York, 1975.
50. R. M. Dawes and B. Corrigan, "Linear Models in Decision Making," *Psychol. Bull.* 81:95–106 (1974).
51. C. W. Kelly and C. R. Peterson, "Decision Theory Research," Tech. Rep. DT/TR 75-5, Decisions and Designs, Inc., McLean, Va., 1975.
52. G. W. Fischer, "Experimental Applications of Multi-attribute Utility Models," in D. Wendt and C. A. J. Vlek, eds., *Utility, Probability, and Human Decision Making*, Reidel, Dordrecht, The Netherlands, 1975.
53. D. von Winterfeldt and G. W. Fischer, "Multi-attribute Utility Theory: Models and Assessment Procedures," in D. Wendt and C. A. J. Vlek, eds., *Utility, Probability, and Human Decision Making*, Reidel, Dordrecht, The Netherlands, 1975.
54. R. V. Brown, C. R. Peterson, W. H. Shawcross, and J. W. Ulvila, "Decision Analysis as an Element in an Operational Decision Aiding System," Tech. Rep. 75–13, Decisions and Designs, Inc., McLean, Va., 1975.
55. J. M. Doughty and C. E. Feehrer, "The AESOP Testbed. Test Series I/2: Summary Report," Tech. Rep. 848, MITRE Corporation, Bedford, Mass., 1969.
56. C. C. Abt, *Serious Games*, The Viking Press, New York, 1970.
57. M. J. Driver and P. L. Hunsaker, "The Luna I Moon Colony: A Programmed Simulation for the Analysis of Individual and Group Decision Making," *Psychol. Rep.* 31:879–888 (1972).
58. R. S. Nickerson and C. E. Feehrer, "Decision Making and Training: A Review of Theoretical and Empirical Studies of Decision Making and Their Implications for the Training of Decision Makers," Tech. Rep. 73-C-0128-1, Naval Training Equipment Center, Orlando, Fla., 1975.
59. H. A. Simon, *The Shape of Automation for Men and Management*, Harper & Row, New York, 1965.

S. B. Sells is Director of Texas Christian University's Institute of Behavioral Research, which he founded in 1962. Dr. Sells has been associated with the university since 1958, when he joined the faculty as a Professor of Psychology. From 1948 to 1957 he was on the faculty of the Air Force School of Aviation Medicine, where he rose to Professor and Head of the Department of Medical Psychology. He is managing editor and associate editor of *Multivariate Behavioral Research* and has written or edited 20 books and about 300 papers, technical reports, and monographs. Dr. Sells earned an A.B. from Brooklyn College and a Ph.D. in Experimental Psychology from Columbia University. He is past President of the Society for Multivariate Experimental Psychology, the Division of Military Psychology of the American Psychological Association, the Southwestern Psychological Association, and the Texas Psychological Association. He is a Fellow of the American Psychological Association and of the Aerospace Medical Association. He received the Raymond F. Longacre Award of the Aerospace Medical Association in 1956 and the Air Force Commendation for Meritorious Civilian Service in 1957.



ORGANIZATIONAL CLIMATE AS A MEDIATOR OF ORGANIZATIONAL PERFORMANCE: THEORETICAL PERSPECTIVE AND APPLICATIONS TO NAVY SHIPS

S. B. Sells

Institute of Behavioral Research
Texas Christian University
Fort Worth, Tex.

The scientific study of organizational behavior seeks to formulate general principles to explain the factors and processes that account for various facets of organizational performance. To achieve generality it is necessary to follow methodological strategies that employ well-defined and replicable measures and organizational situations that reflect the major sources of variance in the phenomena studied. In the case of complex organizations, for example naval ships or business corporations, the total organization is usually too heterogeneous to serve as the primary unit of analysis, and the selection of appropriate organizational units is an important issue.

The developments reported here are results of a collaborative program involving the organizational psychology research group of the Institute of Behavioral Research (IBR) of Texas Christian University and the environmental and social medicine division of the Navy Health Research Center at San Diego. The theoretical orientation of this research reflects in part the influence of interactional theory in psychology and in part the growing ecological and environmental awareness in the social sciences in recent years. Interactional theory views behavior as determined by the transactional interplay of internal dispositional factors and external contextual and stimulus factors [1–4]. In this framework, the environmental context is regarded as a major source of influence in the interactions of individuals with various organizational settings. Organizational psychology was at one time preoccupied primarily with personnel; it is now concerned with organizations as social systems in which persons interact with the total environment. Social systems are interdependent wholes composed of social settings, hardware, traditional and prescribed role requirements and practices, and people.

Since most human organizations in the real world of industrial, institutional, and governmental affairs operate in an indefinite time frame, they tend to be viewed by participants as well as outsiders as "permanent." As a result, goal definition necessarily includes growth and maintenance over time, as well as specific short-term and longer term task objectives. These, in turn, require provisions for organizational and facilities maintenance, for the participation of members over time, and for the concerns of external individuals and organizations responsible for the continued support of the organization or dependent on it for expected products or outcomes. Such considerations led Parsons [5] to define four principal types of exigencies in organizational life: adaptation to environmental pressures, goal (task) attainment, maintenance of organizational patterns, and integration of the total organization.

Parsons pointed out that efforts directed to any of these are frequently carried out at the expense of one or more of the others.

Organizational analysis requires identification of components that have consistent meaning across organizational units and also identification of boundaries that separate each organizational unit from the surrounding environment. Historically, the features of *task, technology, personnel,* and *structure* [6], have received major attention. James and Jones [7], Payne and Pugh [8], and others have also included *context*, which refers to values, policies, traditions, and other normative or prescriptive characteristics of an organization that influence the structure of the organizational units as well as the behaviors of participants at all levels. All of these are not only relevant, but also critically interdependent to the degree that they are best represented as components of social systems. This approach was exemplified in the social system model proposed by Sells [4] as a basis for taxonomic study of organizations. Sells' model involves eight major components, each a cluster of organizational elements, as enumerated in Table 1. The behavioral implications of these components and elements were discussed in relation to isolated small organizations by Sells [4] to organizations in general by Sells [10] and in relation to aspects of long-duration space flight by Sells and Gunderson [11].

The extent to which some of the system characteristics enumerated are interrelated and interdependent with others is obvious, as for example in the case of personnel, whose suitability in a system depends on their compatibility with organizational values, structural requirements (e.g., role requirements, autonomy required), technology, physical requirements, culture patterns, and extent of participation over time. Other important dependencies are less obvious, as between decision time for key decisions and relevant organizational structure. The concept of *system compatibility* is thus a basis for powerful analytic strategies and is undoubtedly used implicitly by planners and organizational analysts in the study of individual systems and subsystems. Nevertheless, the methodology of system analysis has not been embraced extensively in organizational research. This approach, which requires the maximization of information for a single system or unit, may prove more viable when taxonomic research is further advanced than at present and generalization from a typical case becomes a reasonable possibility.

Systems concepts have nevertheless proven helpful in the planning and interpretation of research involving relationships of person and organizational variables to dependent variables representing individual and organizational performance. Such research is closely related to the investigation of environmental influences on behavior and involves issues of representing organizational (environmental) variables in the data matrix [1–4, 9].

When viewed as sources of environmental and organizational influence on behavior, the social system components represent objective reality data at a concrete level. Such specific data, as well as other nonorganizational environmental descriptors (e.g., the temperature extremes of the surrounding area, characteristics of the labor pool, or the population of the city in which a plant is located), have potential importance in organizational research. However, their utility in studies focused on behavioral relations and outcomes may well be limited by their specificity of reference. Many such specific factors nevertheless have related behavioral implications and can be grouped into broad composites. These may also be more meaningfully studied in terms of derived variables conceptualized on the basis of behavioral implications rather than organizational description. Examples of such variables are (a) role ambiguity and (b) role conflict. The first reflects a scale of the extent to which role-prescribed tasks are unclear in their demands, criteria, or relationships with other tasks; the second involves scaling of role-related pressures for conflicting or mutually incompatible behaviors. This class of variables, which reflects common processes, problems, and arrangements that are observable in varying degree in all organizations, represents the basis for definition of organizational climate. Abstractions derived from patterns of such variables are considered to constitute organizational climate factors or dimensions. Combinations of such dimensions are viewed as describing *climates* of particular organizations or groupings of organizations, which could be represented by profiles of scores on scales of the major climate dimensions.

Table 1

Outline of Social System Model for Taxonomic Study of Organizations, Enumerating Major System Components and Elements (Based on Sells [4])

1. *Characteristics of Objectives and Goals*

   Formally prescribed vs informal
   Mandatory vs permissive, voluntary
   Degree of support by superior authority
   Degree of polarization of organization toward their attainment
   Degree of remoteness from current activities
   Existence of criteria of successful attainment
   Degree of certainty of successful attainment
   Number and diversity of priority goals
   Competition with other organizations
   Emphasis on growth

2. *Philosophy and Value Systems*

   Dominant political, religious, social, ethical, economic, and other relevant traditions and values

3. *Personnel*

   Psychological profiles (intellectual, personality, character, attitude)
   Physical profiles (stature, idiosyncratic aspects)
   Demographic characteristics (race, age, ethnic, socioeconomic, other)
   Social status
   Education, experience
   Knowledge and skill profiles

4. *Organizational Structure*

   Size
   Hierarchical organization, centralization-decentralization, autonomy, locus of control
   Differentiation of role and status
   Authority structure, chain of command, succession
   Role structure, communication network

5. *Technology*

   Functions performed—products and services
   Equipment used
   Complexity of theory, knowledge, and training required to perform tasks
   Special requirements involved

6. *Physical Environment*

   Weather, terrain, distances
   Ruggedness: remoteness, hazards, special requirements (e.g., life support), isolation, confinement, endurance demands, embedded stresses, sensitivities
   Mobility permitted
   Structures, furnishings, effects on comfort, health, work efficiency

7. *Social-Cultural Environment*

   Ethnic profile, life styles, living standards, value systems
   Social stratification
   Language, communication, records, forms
   Customs, traditions

8. *Temporal Characteristics*

   Overall duration of system
   Major operational cycles, decision times
   Extent of day-to-day and daily participation required
   Remoteness of goals

DEFINITION AND MEASUREMENT OF ORGANIZATIONAL CLIMATE

In the context of the preceding discussion, the term "climate" is used as a theoretical construct to describe abstractions conceptualized as properties of organizations that have the potential to influence the experience and behavior of their members. These properties are assumed to derive from patterns of concrete elements of social systems in the same manner as meteorologic climate derives from patterns of physical phenomena characteristic of the atmosphere and geography of an area. The effects of organizational climate are various generalized orientations of organizational members that are both shared by a majority of members of an organizational unit and acquired in relation to factors specific to the organizational situation [10].

Jones and James [12] suggested that the following assumptions underlie much of the organizational climate research and theorizing. According to their formulation, organizational climate (a) describes situational characteristics in terms of their influences on individuals and groups; (b) is a multidimensional domain with a common core of dimensions, although some dimensions may vary in relevance in particular situations and populations; (c) is based primarily on those aspects of the environment that have direct and immediate ties to individual experience and behavior; and (d) occupies an intervening role in a model of organizational functioning such that the point of intervention is between the situation and the individual, reflecting a transformation of situational characteristics into situational influence.

Definition of the Domain Universe

In common with many other areas of social science research, the initial systematic efforts to formulate the salient dimensions of organizational climate have emerged from exploratory empirical studies. Undoubtedly global concepts guided the empirical investigations, but most of them focused on prediction of various criteria in particular organizations and were sensitive to the situational features of the respective organizations. Reviews of such research by Sells [10], Indik [13], Hellriegel and Slocum [14], James and Jones [15], Schneider [16], and Payne and Pugh [8] have identified four major areas of principal concern. As summarized by Jones and James [12], these are

1. Job or role aspects, such as variety, challenge, job pressures, and role ambiguity
2. Leadership style and behavior, such as initiation of structure, goal emphasis, and consideration and support of subordinates
3. Characteristics of work groups, such as friendliness and warmth, cooperation and mutual help, and formation of cliques
4. General system and subsystem attributes, such as interdepartmental conflict, provision of career opportunities, fairness of the reward system, and clarity of communication structure.

Table 2 includes a list of 35 variables that constitute a working definition of the domain of processes, problems, and arrangements embraced by organizational climate, as formulated in the joint IBR—Naval Health Research Center (NHRC) report on organizational and environmental factors in health and personnel effectiveness aboard Navy ships [12]. These represent the major discrete facets of organizational climate represented in prior research at IBR and in the literature. In Table 2 these are grouped in the four categories enumerated above. This is not a comprehensive definition of the universe, but a useful working approximation, as discussed below.

Approach to Measurement

Quantitative information concerning the variables indicated in Table 2 can be obtained by several different strategies, such as objective measurement of selected relevant indices, recording of selected behaviors by covert observers, ratings by participant observers, and questionnaires administered to members. However, most investigators of organizational climate have chosen the last-mentioned approach. This has several obvious advantages, but also some potential disadvantages, and has resulted in another major development, discussed below.

The rationale of the questionnaire approach to organizational climate measurement is best explained as that of a group reality test. Most

Table 2

Thirty-five Organizational Climate Variables [12]

Characteristics of Job Task-Role

1. Role ambiguity	Degree of ambiguity in demands, criteria, interface with other jobs-tasks-roles
2. Role conflict	Degree to which role performance is affected by pressures to engage in conflicting or mutually exclusive behaviors
3. Job autonomy	Degree of information and opportunity to analyze tasks or problems and to act without consultation or permission
4. Job variety	Range of types of tasks, equipment, and behaviors involved in jobs
5. Job importance	Degree of importance of job to the organization
6. Job feedback	Degree to which individuals receive information on progress and effectiveness of their work and behaviors
7. Job challenge	Degree to which individuals receive opportunities to make full use of their abilities, skills, and knowledge
8. Job pressure	Adequacy of time, information, resources to complete assignments, and degree of threat implied for substandard performance
9. Job design efficiency	Degree to which job information, procedures, equipment, and arrangements permit effective performance and lead to valued organizational results
10. Job standards	Degree to which exacting standards of quality and accuracy are required in job performance
11. Job isolation	Degree to which job restricts opportunities to interact with other persons

Characteristics of Leadership Style and Performance

12. Leader support	Degree to which leaders are aware of and responsive to needs of subordinates and show consideration for their feelings of personal worth
13. Goal emphasis	Degree to which leaders stimulate subordinates' involvement in meeting organizational goals
14. Work facilitation	Degree to which leaders provide resources, guidance, problem solutions, and aid subordinates in achieving planned goals
15. Interaction facilitation	Degree to which leaders encourage development of close, cohesive work groups
16. Planning and coordination	Degree to which leaders plan effectively and coordinate work group activities to facilitate optimal performance

17. Interaction upward	Degree to which leaders represent their work groups effectively in interactions with higher levels of management
18. Confidence and trust—up	Degree of confidence and trust of members in their superiors
19. Confidence and trust—down	Degree of confidence and trust of superiors in their subordinates

Characteristics of Work Group

20. Cooperation	Existence of an atmosphere of cooperation to carry out difficult tasks; evidence of mutuality of goals and sharing of reward for success
21. Reputation for effectiveness	Degree to which work group enjoys a record of effective performance and is expected to perform well by peers as well as superiors
22. Esprit	Degree to which members show pride in their group, their fellow members, and their record as a group
23. Friendliness and warmth	Degree to which warm, friendly relations, trust, and mutual liking prevail

Management Policies and Postures

24. Openness of expression	Degree to which organizational atmosphere fosters expression of ideas, dissent, criticism, opinions, suggestions, and other information upward
25. Communication—down	Degree to which information is communicated to subordinates on matters affecting their work, status, and feelings of well-being, including advance knowledge of impending changes in procedures, policies, etc.
26. Interdepartmental cooperation	Degree of cooperative action, communication, and mutual help among departments
27. Subsystem conflict	Degree to which subsystem goals, policies, and actions conflict
28. Ambiguity of structure	Degree to which role definition, lines of authority, responsibility, and communication channels are unclear or undefined
29. Management consistency	Degree of consistency and fairness in administration of organizational policies and rules
30. Organizational esprit	Degree to which individuals believe that the organization performs an important function and offers them opportunities for growth and reward
31. Planning effectiveness	Degree to which planning results in effective scheduling and coordination of personnel, materiel, and information

32. Fairness and objectivity of the award system	Degree to which merit rather than favoritism and bias determine the award of recognition, promotion, and other types of reward
33. Opportunities for growth and advancement	Degree to which the organization provides career paths, training, and recognition to afford growth in responsibility and advancement in job status over time
34. Management consideration	Degree to which the organization provides means to understand employee needs and problems and is responsible to them
35. Professional esprit	Degree to which individuals believe that their profession has a good image to outsiders and provides opportunities for growth and advancement

responsible individuals, when confronted with testimony that involves perceptual data, tend to use their own perceptions as criteria of veridicality. For example, if a subordinate were to complain that a room is too hot, his supervisor would generally be more willing to adjust the temperature if he, too, perceived it as too hot. In the present instance, group consensus on questions related to the issues listed as variables in Table 2 implies (a) informed evaluation by populations of participants (group members) used as observers and (b) dependence on consensus as the test of reality. The advantages of this approach are mainly in convenience, cost, and time savings compared to other methods, and in the potential applications of the questionnaire results to organizational development efforts through feedback of tabulated data to participants. The problems, which are not always disadvantages, are related to the feasibility of obtaining frank, objective responses, particularly in situations in which employees may be afraid to report information that they view as critical of the organization or of their superiors. While this can be controlled to a degree by anonymous reply formats and procedures to protect confidentiality, such measures are often only partially effective. In the questionnaire methods, organizational climate measures are usually represented as scaled aggregated scores for organizational units.

Psychological Climate as Distinguished from Organizational Climate

A consequence of the questionnaire method of measurement of organizational climate is that it yields individual scores as well as scores for organizational units. Without violation of confidentiality requirements in many cases, or with informed consent in others, the "climate" scores of individuals can be analyzed in relation to a wide range of person and organization data with highly productive results, as demonstrated subsequently. However, such individual scores are not measures of organizational climate, but rather measures of perceptions of organizational climates, which may vary among members of the same unit as filtered through different idiosyncratic sensitivities. Jones et al. [17] have used the term *psychological climate* to distinguish the individual perceptual measures from the organizational consensus measures and defined psychological climate [12] as referring to "the individual's internalized representations of organizational conditions" and as reflecting "a cognitive transformation and structuring into perceived situational influences."

In contrast to the position taken here, some theorists, such as Schneider [16] have conceptualized organizational climate in phenomenological terms and treat the entire topic in perceptual terms. While this may identify a source of controversy at the theoretical level, it does not affect the uses or interpretation of organizational climate data, since it appears likely that the questionnaire approach will continue to be the measurement method of choice. Whether an organizational climate index is in truth an estimate of a reality situation obtained by consensus of participants used as observers or an aggregated expression of individual perceptions may only reflect the orientations and preferences of different

theorists. In either case, we are dealing with abstractions conceptualized as intervening variables representing organizational influence on member behavior, and these abstractions can be measured and studied in relation to organizational and individual behaviors.

Dimensions of Psychological Climate

In view of the system character of organizations it is reasonable to assume that the 35 variables listed in Table 2 are intercorrelated to some extent and that the true number of discrete climate dimensions is considerably smaller. The IBR-NHRC research referred to above included an organizational climate questionnaire of 145 items in which each of the 35 variables was represented by 2 to 7 items. This questionnaire was part of a larger survey instrument, which also included inquiries concerning the physical environment of work, dining, recreational, berthing, and common areas, job satisfaction, and other information. It was administered to a sample of 4315 Navy enlisted men on 20 ships operating in the Atlantic and Pacific Oceans during the latter half of 1973 and, for comparison and testing of generality, to one sample of 398 male firemen in two municipal fire departments in a large metropolitan area in North Central Texas and a second civilian sample of 504 managerial employess of a nonprofit health care program in Southern California.

For each of the three samples component analyses [18] were made of the intercorrelations among cluster scores, and each analysis yielded six components, which accounted for 59% of the total variance in the Navy sample, 63% in the civilian firemen, and 67% in the health care management employees [12]. The loading patterns of the six components are shown for the Navy sample in Table 3. These were very similar for the first five components in the other two samples, as shown by the coefficients of congruence* in Table 4. Thus, there were five dimensions of climate that were replicated exceptionally well in three samples from dissimilar types of organizations.

As shown in Table 3, the first five components, reflecting *Conflict and Ambiguity in the organizational situations, Job Challenge, Importance, and Variety, Leader Facilitation and Support, Workgroup Cooperation, Friendliness and Warmth,* and *Professional and Organizational Esprit*, were sharply defined by patterns of from 4 to 10 item clusters with moderate to high loadings (over 0.40). The sixth component was more specific with respect to the individual samples; this component was retained for subsequent analysis, with the label *Job Standards*.

Homogeneity of Climate Within Organizations

Although literary references such as "a tight ship" or "Theory X type of management" are common, experienced managers as well as experienced professionals in organizational research recognize that such generalities are not useful indicators of organizational climate, especially of organizational subsystems. By nature, complex organizations with wide variations in personnel training and skills, technology, and types of responsibility among subsystem units are inherently heterogeneous. As a consequence, it is reasonable to consider the implications of the heterogeneity that exists within various types of organizations for organizational climate and also for effective management strategies.

The 20 ships in the IBR-NHRC research were organized, on the average, into 4 or 5 departments each (a total of 91 departments) and 2 to 3 divisions per department (a total of 223 divisions). Relationships were studied among department and division measures of organizational context, structure, and climate across all departments and divisions [12]. *Context* variables included a measure of technology (the degree of nonroutineness and complexity of work, difficulty of evaluation, and uncertainty of success), emphasis on morale (by the officer in charge), and emphasis on following standardized procedures. *Structure* variables included size, complexity of role structure (number of separate occupational titles), number of rank levels, and span of control, as well as

*The coefficient of congruence (C with subscripts to denote the variables compared) was named by Tucker [19] but developed by Burt [20]. It is an index of the relationships among the loadings on any pair of components and can be interpreted as indicating that the respective components are congruent with each when the coefficients are high.

Table 3

Six Climate Components Derived From a Study of 4315 Navy Enlisted Men in 1973 [12]

Defining Variable	*Loading*
I. Conflict and Ambiguity	
27. Subsystem Conflict	0.66
28. Ambiguity of Structure	0.66
26. Interdepartmental Cooperation	−0.57
25. Communication—down	−0.55
31. Planning Effectiveness	−0.53
32. Fairness and Objectivity of the Reward System	−0.51
*1. Role Ambiguity	−0.47
*29. Management Consistency	−0.47
9. Job Design Efficiency	−0.46
*34. Awareness of Employees' Needs and Problems	−0.41
II. Job Challenge, Importance, and Variety	
7. Job Challenge	0.75
5. Job Importance	0.68
4. Job Variety	0.67
11. Job Isolation	−0.54
3. Job Autonomy	0.52
6. Job Feedback	0.46
*1. Role Ambiguity	−0.44
*10. Job Standards	0.42
III. Leader Facilitation and Support	
14. Work Facilitation	0.80
15. Interaction Facilitation	0.77
62. Leader Support	0.72
13. Goal Emphasis	0.72
16. Planning and Coordination	0.61
6. Job Feedback	0.51
*17. Interaction Upward	0.50
IV. Workgroup Cooperation, Friendliness, and Warmth	
20. Cooperation	0.75
21. Reputation for Effectiveness	0.72
23. Friendliness and Warmth	0.64
22. Esprit	0.59
V. Professional and Organizational Esprit	
35. Professional Esprit	0.79
30. Organizational Esprit	0.66
24. Openness of Expression	0.64

theorists. In either case, we are dealing with abstractions conceptualized as intervening variables representing organizational influence on member behavior, and these abstractions can be measured and studied in relation to organizational and individual behaviors.

Dimensions of Psychological Climate

In view of the system character of organizations it is reasonable to assume that the 35 variables listed in Table 2 are intercorrelated to some extent and that the true number of discrete climate dimensions is considerably smaller. The IBR-NHRC research referred to above included an organizational climate questionnaire of 145 items in which each of the 35 variables was represented by 2 to 7 items. This questionnaire was part of a larger survey instrument, which also included inquiries concerning the physical environment of work, dining, recreational, berthing, and common areas, job satisfaction, and other information. It was administered to a sample of 4315 Navy enlisted men on 20 ships operating in the Atlantic and Pacific Oceans during the latter half of 1973 and, for comparison and testing of generality, to one sample of 398 male firemen in two municipal fire departments in a large metropolitan area in North Central Texas and a second civilian sample of 504 managerial employess of a nonprofit health care program in Southern California.

For each of the three samples component analyses [18] were made of the intercorrelations among cluster scores, and each analysis yielded six components, which accounted for 59% of the total variance in the Navy sample, 63% in the civilian firemen, and 67% in the health care management employees [12]. The loading patterns of the six components are shown for the Navy sample in Table 3. These were very similar for the first five components in the other two samples, as shown by the coefficients of congruence* in Table 4. Thus, there were five dimensions of climate that were replicated exceptionally well in three samples from dissimilar types of organizations.

As shown in Table 3, the first five components, reflecting *Conflict and Ambiguity in the organizational situations, Job Challenge, Importance, and Variety, Leader Facilitation and Support, Workgroup Cooperation, Friendliness and Warmth,* and *Professional and Organizational Esprit*, were sharply defined by patterns of from 4 to 10 item clusters with moderate to high loadings (over 0.40). The sixth component was more specific with respect to the individual samples; this component was retained for subsequent analysis, with the label *Job Standards*.

Homogeneity of Climate Within Organizations

Although literary references such as "a tight ship" or "Theory X type of management" are common, experienced managers as well as experienced professionals in organizational research recognize that such generalities are not useful indicators of organizational climate, especially of organizational subsystems. By nature, complex organizations with wide variations in personnel training and skills, technology, and types of responsibility among subsystem units are inherently heterogeneous. As a consequence, it is reasonable to consider the implications of the heterogeneity that exists within various types of organizations for organizational climate and also for effective management strategies.

The 20 ships in the IBR-NHRC research were organized, on the average, into 4 or 5 departments each (a total of 91 departments) and 2 to 3 divisions per department (a total of 223 divisions). Relationships were studied among department and division measures of organizational context, structure, and climate across all departments and divisions [12]. *Context* variables included a measure of technology (the degree of nonroutineness and complexity of work, difficulty of evaluation, and uncertainty of success), emphasis on morale (by the officer in charge), and emphasis on following standardized procedures. *Structure* variables included size, complexity of role structure (number of separate occupational titles), number of rank levels, and span of control, as well as

*The coefficient of congruence (C with subscripts to denote the variables compared) was named by Tucker [19] but developed by Burt [20]. It is an index of the relationships among the loadings on any pair of components and can be interpreted as indicating that the respective components are congruent with each when the coefficients are high.

18. Confidence and Trust—up	0.61
33. Opportunities for Growth and Advancement	0.57
8. Job Pressure	−0.53
*34. Awareness of Employees' Needs and Problems	0.52
2. Role Conflict	−0.49
*17. Interaction Upward	0.48
*29. Management Consistency	0.45
VI. Job Standards	
*10. Job Standards	0.54
8. Job Pressure	0.40
19. Confidence and Trust—down	−0.40

*The following variables had two significant loadings:

Item 1, on Components I and II
Item 29, on Components I and V
Item 34, on Components I and V
Item 10, on Components II and VI
Item 17, on Components III and V.

Table 4

Coefficients of Congruence of the Five Well-Defined Climate Components Among Three Dissimilar Organizational Samples: 1. Navy Enlisted Men (N = 4315), 2. Civilian Firemen (N = 398), and 3. Civilian Health Care Program Management Employees (N = 504)*

Climate Components	*Congruence Coefficients*		
	C12	*C13*	*C23*
I. Conflict and Ambiguity	0.75	0.93	0.74
II. Job Challenge, Importance, and Variety	0.77	0.89	0.89
III. Leader Facilitation and Support	0.97	0.96	0.96
IV. Workgroup Cooperation, Friendliness, and Warmth	0.91	0.87	0.90
V. Professional and Organizational Esprit	0.83	0.90	0.77

*Based on Jones and James [12].

operational measures, such as centralization of decisionmaking, centralization of work allocation and scheduling, interdependence of work units, formalization of role and communication structures, and standardization of procedures. *Organizational climate* was represented by aggregate scores on the six psychological climate dimensions.

With few exceptions, the correlations obtained, which were generally low and insignificant, reflected the substantial degree of heterogeneity on the variables studied that existed among divisions within departments. The context, structure, and climate scores for departments were too global to represent meaningfully the conditions in their respective divisions. For the Navy ships it appeared that the *division level* was the highest level at which productive organizational analyses were warranted. In addition, the consensus (or within-group agreement) on psychological climate measures was high enough for divisions to justify the aggregation of division climate scores. The practical implication of these results was that they pointed toward the *division* as the appropriate level of organizational analysis for Navy ships. The subsequent analyses of organizations were focused on the variations of division climates and also on the predictive power of division climates. Other analyses, at the individual level, involved the psychological climate measures.

TYPES OF ORGANIZATIONAL CLIMATE

The identification of major dimensions of organizational climate that were replicated over widely different types of organizations contributed to the generalizability of research using these measures. A strongly indicated next step was to address the feasibility of describing the *climates of various organizations and types of* organizations in terms of profiles of scores on the climate dimensions. This was approached first by examining the correlates of organizational climate across all divisions in the ship sample.

Correlates of Organizational Climate

A meaningful set of relationships between organizational climate measures and measures of organizational context and structure would be valuable both as an indication of system congruence and as evidence of the consistency with theoretical expectation of the independent measures used in the analysis. James et al. [21] computed correlations among 6 climate dimensions, 7 context measures, 11 structure measures, and also 5 additional personnel measures for the Navy sample of 223 divisions of the 20 ships included in the study. In this analysis the division climate and personnel scores were the mean values for the respective divisions; the corresponding context and structure scores were direct measures of the variables listed. The results are shown for significant correlations only in Table 5. Although the correlation coefficients in Table 5 appear generally low, 33% of the 138 correlations computed were significant to at least the 0.05 level, and 22% of the total number were significant at the 0.01 level.

The ship sample included 18 destroyer types and 2 aircraft carriers and averaged between 12 and 15 divisions per ship. The division types vary widely in equipment, technology, functions, structure, and staffing. They include 12 principal functional types, as follows:

- Deck divisions—maintenance, paint, boat handling, lines
- Engineering, boiler—operation, maintenance, and repair
- Engineering, machinery and engines
- Engineering—auxiliary, repair and damage control, and electrical
- Supply—ship's stores, food service, stewards, cooks
- Navigation and administration—ship administration, personnel, and navigation
- Guns—gunnery and ordnance
- Antisubmarine warfare
- Sophisticated weapons—nuclear, missiles, and fire control
- Operations, communications
- Operations, intelligence—combat intelligence centers
- Operations, electronics.

The correlations in Table 5 were computed across all divisions for all 20 ships. The correlates of each of the six climate dimensions, discussed

Table 5

Correlates of Division Climate: Significant Correlations Between Division Climate Mean Scores and Division Scores on Context, Structure and Personnel Variables (N = 223 Divisions of 20 Navy Ships) [21]*

	Conflict and Ambiguity	Job Challenge, Importance and Variety	Leader Facilitation and Support	Workgroup Cooperation, Friendliness, and Warmth	Professional and Organizational Esprit	Job Standards
Division Context						
1. Emphasis on Morale	—	—	0.17	—	0.17	—
2. Emphasis on Standards and Procedures	—	—	—	—	—	0.18
3. Level of Technology	—	0.21	—	0.19	–0.19	–0.20
4. Funds for Habitability	—	—	—	—	—	—
5. Condition of Equipment	–0.19	—	—	0.31	—	—
6. Rating of Personnel	–0.13	—	0.23	0.43	—	—
7. Funds and Supplies for Work	–0.22	—	—	0.20	—	—
Division Structure						
8. Size	—	–0.24	—	0.23	—	—
9. Specialization—Number of Jobs	—	—	—	—	0.16	—
10. Span of Control	—	0.36	—	0.20	—	—
11. Number of Rank Levels	—	—	—	–0.24	0.13	0.14
12. General Centralization	0.13	—	—	—	–0.14	—
13. General Standardization	—	—	—	—	—	—
14. Interdependence	—	0.17	—	—	—	–0.13
15. Formalization of Roles	—	—	—	—	—	0.18
16. Centralization of Work	—	—	—	—	–0.13	—
17. Formalization of Communication	—	—	0.15	—	0.21	—
18. Standardization of Expenditures	—	—	—	—	—	—
Division Personnel Mean Scores						
19. Time in Navy	—	0.49	0.16	0.18	0.19	—
20. Number of Advanced Train. Schools	—	0.24	—	0.42	—	—
21. Number of Other Training Schs.	—	0.35	—	0.44	–0.15	—
22. Years of Formal Education	—	0.30	—	0.38	—	—
23. Intelligence and Aptitude	–0.13	0.32	—	0.44	–0.23	—

*Correlations over 0.17 are significant at the 0.01 level; all others at the 0.05 level.

below, indicate considerable consistency across dimensions.

Conflict and Ambiguity—Three context variables were negatively associated with this dimension, suggesting that conflict and ambiguity tended to be perceived as low in divisions in which the reliability of equipment was high, the evaluation of division personnel by the officer in charge was favorable, and funds and supplies were adequate for accomplishing the required work. In addition, the structure variable, *general centralization* (of authority and information), was positively associated with conflict and ambiguity—contrary to the apparent opinions of many authoritarian managers—and intellectual aptitude was negatively associated with this dimension in that divisions with higher mean aptitude tended to have less perceived conflict and ambiguity. The keys to minimization of conflict and ambiguity in general indicated by these results were thus

Maintaining equipment
Leader attitudes of approval of personnel
Planning and logistic support to provide adequate funds and supplies
Decentralization of authority and information to the extent possible consistent with work effectiveness.

Conflict and ambiguity were found to be lower in divisions with higher mean aptitude levels, in which the processes mentioned above could presumably be observed to a greater extent than in divisions staffed with personnel of lower aptitude.

Job Challenge, Importance, and Variety—The significant correlates of this dimension included all five personnel variables, three structure variables, and one context variable. Together, these suggest that divisions perceived as high in job challenge, importance, and variety are high in technology, are small in size, have low spans of control, are highly interdependent with other units aboard ship, and are staffed with intellectually able, better educated, well-trained, and experienced personnel. This is a consistent set of correlates that link an important dimension of organizational climate to significant aspects of divisions viewed as social systems.

Leader Facilitation and Support—This dimension of climate reflects the two historically significant aspects of leadership: initiation of structure and consideration [22]. Although correlated with only four variables, these form a meaningful and consistent pattern. According to the results obtained, leadership was viewed most favorably in divisions in which the officer in charge emphasized morale in his actions, evidenced favorable evaluation of his crew, and provided clearly indicated formal communication channels. In such divisions, average service time tended to be high.

Work Group Cooperation, Friendliness, and Warmth—There were more significant correlations with this dimension than with any other. As with job challenge, the highest correlations were obtained with personnel variables; high intelligence and high scores on advanced training in the Navy were associated with cohesive, friendly work group climate, as were the structure variables indicating small work group size, low span of control, and flat organizational configuration (that is, few rank levels in the division). Four context variables were also associated with this dimension: favorable evaluation of the crew by the officer in charge, good condition of equipment, adequate funds and supplies for work needs, and high level of technology. These associations form a highly consistent network identifying divisions that perform high technology jobs; those divisions, which tend to be small, staffed by technically trained and intellectually superior specialists, and somewhat informal in supervisory style, were the most cohesive and friendly.

Professional and Organizational Esprit—Many of the factors that correlated with low conflict and ambiguity; high job challenge, importance, and variety; high leadership; and high work group cooperation, friendliness, and warmth were associated with professional and organizational esprit in the opposite direction. Thus high esprit tends to be perceived in divisions with low mean intellectual aptitude and few training schools attended, but with high mean service time. Other correlates of high esprit are low technology, high emphasis on morale, many job specialties (as in Supply divisions), many rank levels, and formalization of communication channels. Two structure variables, general centralization and centralization of work scheduling, have negative correla-

tions, opposed in sign to expectancy consistent with the pattern described. This is probably best explained by the fact that the diversity of specialties and number of rank levels, together with the relatively low level of jobs and personnel in these divisions, makes centralization a necessity. In sum, high esprit probably reflects identification with job situations in the Navy that offer better opportunities to the personnel described than they could find in more competitive civilian situations.

Job Standards—As expected, high standardization was associated with low-technology, less interdependent, more highly formalized and stratified divisions. There were no significant personnel correlates of this climate dimension in the present study.

A Typology of Climates for Ship Divisions

Divisions of the Navy ships in the IBR-NHRC study sample were judged to be fully represented by the 12 functional types enumerated above. Division climate profiles for 223 divisions so classified were compared [21] by using the method of discriminant analysis which yields composite scores (discriminant functions) that maximize differences between groups in comparison to variance within groups. The significant differences obtained suggested that average profiles of climate scores for these types of divisions could meaningfully represent types of division climate. There were however, similarities among several of the average profiles, and the typology was reduced to seven by means of a hierarchical grouping analysis of the 12 division type profiles [23]. This method of cluster analysis separates a sample of profiles into homogeneous clusters and classifies every profile in the cluster that it resembles most in terms of profile distance.

The seven types of division climate are described below. It is of interest that they reflect certain similarities among all types of divisions as well as a number of characteristic differences in salient dimensions. The similarities are observable in the comparisons of mean scores in Table 6. The mean dimension scores were computed on individuals; they correspond approximately to a mean of 50 and standard deviations from 3.7 to 5.2. Since the total range of mean scores, across all climate dimensions and types, is only from a high of 55 (dimension IV, type VI) to a low of 44 (dimensions II and IV, type V), it is apparent that variations from the grand means were rarely more than one standard deviation. This is important since in an effective Navy all units must function within an optimal range. Differences among types of divisions must be considered within this range, but they are nevertheless interesting and have instructive implications in relation not only to the Navy, but also to problems of organizational management and development in general.

The types were named for the salient variables in each cluster profile in Table 6 that differed more than one-half of a standard deviation from the actual grand means of the respective dimensions. They were as follows:

1. Cooperative and Friendly Division Climate—This climate profile was characteristic of a cluster comprised of three functional types of divisions: guns, antisubmarine warfare, and navigation. The cluster profile can be expressed three ways: as a profile of means, by rank order of dimensions among types, and by rank order of dimensions within types, as shown immediately below:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	49	50	51	53	50	49
2. Rank of Each Dimension Among Division Types	5	4	3	3	3	4
3. Rank of Each Dimension Within Cluster Profile	5.5	3.5	2	1	3.5	5.5

Table 6

Mean Dimension Score Climate Profiles Representing Seven Types of Division Climate [23]

		Climate Profiles					
		I	*II*	*III*	*IV*	*V*	*VI*
Division Climate Type	*Division Types Included*	*Conflict and Ambiguity*	*Job Challenge Importance & Variety*	*Leader Facilitation and Support*	*Work Group Cooperation, Friendliness, & Warmth*	*Professional and Organizational Esprit*	*Job Standards*
I	Guns Antisubmarine Warfare Navigation	49	50	51	53	50	49
II	Missiles Fire Control Nuclear Weapons Engineering—Auxiliary, Repair-damage Control, Electrical	52	51	49	53	50	46
III	Operations—Communications Operations—Intelligence	48	50	52	50	47	54
IV	Engineering—Boiler Engineering—Machinery	51	50	50	47	49	52
V	Deck	50	44	48	44	51	49
VI	Operations—Electronics, Radar	48	54	49	55	47	47
VII	Supply	50	49	51	48	53	51

Mean scores that deviate more than one-half a standard deviation from the grand mean of a dimension are underlined.

The underlining of the mean score for Work Group Cooperation indicates that it was more than one-half a standard deviation above the grand mean for this dimension. This profile is highest on cooperation and friendliness in the work group and next highest on the leadership dimension. It is quite low on conflict and ambiguity. The mean scores on the remaining dimensions have in-between ranks. In view of the rank orders of the dimensions within the profile, a more detailed interpretation of this climate cluster can be given. This should mention high leadership and absence of conflict in the work environment, in addition to friendliness, cooperation, and warmth in the work group. On the basis of the correlates of the climate dimensions (discussed earlier), divisions of this kind tend to have, as associated characteristics, high technology and aptitude levels of personnel, high evaluation of personnel by their leaders, high planning effectiveness, much decentralization of control, and good condition of equipment. Overall the impression is one of an elite group.

II. Conflicting and Ambiguous Division Climate—This cluster included the following types of divisions: missiles, fire control, nuclear weapons, and three types of engineering divisions—auxiliary, repair-damage control, and electrical. The cluster profiles were as follows:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	52	51	49	53	50	46
2. Rank Among Types	1	2	5	2	4	7
3. Rank Within Profile	2	3	5	1	4	6

Although the mean score for workgroup cooperation is higher than that for conflict, the conflict mean ranks highest among all clusters and is also salient in that it is more than half a standard deviation above the grand mean; at the same time the mean for job standards is lowest in the profile and also salient. Looking at ranks among types, this cluster is high on conflict, job challenge, and work group cooperation as well as low on job standards. These are in themselves conflicting indications and suggest that the conflict dimension characterizes this cluster very well. Perhaps the problems involving the responsibility and critical importance of nuclear, missiles, and associated engineering functions on one hand, and the restrictions and frustrations associated with them on the other, are the major contributors to the high level of conflict and ambiguity in these divisions.

III. Alienating and Restrictive Division Climate—This cluster included two types of operations divisions: communications and intelligence. The cluster profiles were as follows:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	48	50	52	50	47	54
2. Rank Among Types	6	3	1	4	6	1
3. Rank Within Profile	5	3.5	2	3.5	6	1

The high mean score and ranks on job standards, together with the low mean score and ranks on esprit are strongly indicative of alienation from the environment aboard ship and also of the restrictiveness caused undoubtedly by the high security and confidentiality usually associated with the communications and intelligence functions aboard ship. The exacting job standards epitomized here fit well with the correlates of high standards discussed earlier—namely, low interdependence with other units, high formalization of structure, and high stratification (many levels of rank)—although not necessarily with low technology in all respects. Most typically, the climate pattern of this cluster reflects the generality of the alienating effect of responsible work in highly secure and confidential areas.

IV. Unfriendly Division Climate—Two types of engineering divisions (boiler and machinery) make up this cluster. The cluster profiles were as follows:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	51	50	50	47	49	52
2 Rank Among Types	2	5	4	6	5	2
3. Rank Within Profile	2	3.5	3.5	6	5	1

The most salient feature of this profile is the low score on work group cooperation. However, the high ranks on job standards (associated with low interdependence, high formalization, and high stratification) and on conflict and ambiguity, elaborate the impression of a type of work situation lacking friendliness, cooperation, and interpersonal warmth. This is also consistent with the low rank on esprit, which suggests tendencies of group members not to identify with their organizations. Together these factors describe the work environment of boiler and machinery activities aboard ship as unfriendly and uncooperative.

V. Monotonous, Impersonal, and Unsupportive Division Climate—This cluster was composed entirely of deck divisions, which are generally manned by greater proportions of men with low aptitude, low rank, and low service time than other types of divisions, and whose functions involve many unskilled and semiskilled tasks. The cluster profiles were as follows:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	50	44	48	44	51	49
2. Rank Among Types	3	7	7	7	2	5
3 Rank Within Profile	2	5.5	4	5.5	1	3

The low mean scores on job challenge, leadership, and work group cooperation represent the lowest cluster ranks on these dimensions; in addition, the rank on job standards tends toward the low extreme. Together, these provide a clear indication of a monotonous and unchallenging, cold

and unfriendly "stepchild" type of work situation, in which leadership is perceived as unsupportive, but which in reality claims a disproportionate amount of management and leadership attention compared to other types of divisions.

VI. Enriched and Cohesive Work Environment, But Organizationally Uninvolving—Composed of highly skilled and trained electronics technicians, the divisions in this cluster provide challenging and intrinsically satisfying work experience, but have difficulty in retaining their superior personnel in Navy careers, mainly because of competition by more attractive civilian alternatives. The profiles for this cluster were as follows:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	48	54	49	55	47	47
2. Rank Among Types	7	1	6	1	7	6
3. Rank Within Profile	4	2	3	1	5.5	5.5

The mean score profile can be seen to be a combination of extremes, with the highest rank among the seven clusters on job challenge and work group cooperation and the lowest on all the others. The cluster title describes the situation clearly and also provides an unequivocal diagnosis for organizational development in a division type of critical importance in the Navy.

VII. Organizationally Involving Division Climate—This cluster was composed exclusively of supply divisions, which handle ships' stores and food service and employ clerks, cooks, food handlers, and stewards, many of whom consist of foreign-born career men who view their Navy jobs as providing superior opportunities to those available in their native land. The cluster profiles were as follows:

	Conflict	*Job Challenge*	*Leadership*	*Work Group Cooperation*	*Esprit*	*Job Standards*
1. Profile of Means	50	49	51	48	53	51
2. Rank Among Types	4	6	2	5	1	3
3. Rank Within Profile	4	5	2.5	6	1	2.5

The most distinctive feature of this profile is the high score on esprit, on which it ranks first both among the seven climate types and within this profile. Although this undoubtedly reflects the influence of the mess stewards, as noted above, the supply divisions also include a number of other job specialties in which perceptions of the work environment and of the Navy are about average on the other dimensions. The high mean score on the Esprit dimension among men in Supply divisions reflects pride in organization as well as identification with the Navy.

Summary Comment—This brief survey of the seven types of organizational climate experienced by the enlisted crews of Navy ships provides a panorama of some of the salient, systemwide or-

ganizational problems that were identified in the IBR-NHRC research program. The problems mentioned are based on the ship sample studied, but are presumed to be generalizable to the entire fleet, assuming that contextual Navy and world conditions have not changed significantly since 1973, when the data were collected. From the standpoint of practical implications these results describe Navy ships as complex and heterogeneous organizations composed of major subsystems that have quite different characteristic problems requiring command and organizational development attention.

One of the contributions of the analysis of climate types is that it highlights salient system characteristics of the respective divisions that have implications for supervisory and organizational development strategies. Examples of these are (a) the frustration and conflict associated with the job and organization in Climate Type II (missiles, fire control, nuclear divisions), (b) the alienation associated with highly classified communications and intelligence work in Climate Type III, and (c) the competition of Navy electronics jobs with more attractive civilian opportunities in Climate Type VI. These should be interpreted as illustrative of individual -environment interactions that must be taken into account in formulating plans for supervisor training, job redesign, or other interventions to achieve higher reenlistment rates, increased unit effectiveness, or other specific goals.

PREDICTION OF ORGANIZATIONAL CRITERIA

The utility of climate classification implies the assumption that organizational climate is related to organizational performance. This assumption was tested in the Navy ship sample by correlating measures representing the seven division climate types (each scored 0 or 1 for this purpose) with an experimental measure of division performance [23].

Division performance for a subset of 160 divisions of 19 ships in the study sample was estimated in a multistage process involving interviews with naval officers and ship commanders. In successive stages, critical dimensions of performance were identified, then these were evaluated for each division, and finally a composite score was developed for each division, incorporating the major dimensions that correlated significantly with the composite. The final composite consisted of a unit-weighted composite of the following nine dimensions of division effectiveness: (a) Quality of Work Performed, (b) Adherence to Planned Maintenance Schedules, (c) Operational Readiness to Fulfill Commitments, (d) Performance Under Pressure, (e) Efficiency, (f) Cooperation with Other Divisions, (g) Leadership Ability of Enlisted Supervisors, (h) Requests for Transfer to Other Divisions or Departments, and (i) Use of Drugs and Alcohol. A tenth dimension, Safety, was excluded from the composite after it was found to have low correlations with the other nine dimensions.

The analytic procedure employed is described in detail by Jones and James [23]. The subsample of 160 divisions was divided into two equivalent groups representing different ships and results for each group were cross-validated on the other. The cross-validity correlations of division climate with division performance were 0.41 in the first group and 0.39 in the second. These correlations were significant beyond the 0.01 probability level. Although the climate dimensions were correlated with the division context, structure, and personnel variables, as shown in Table 5, the combined correlation of these variables and division climate with division performance rose to 0.60.

CONCLUDING COMMENT

Organizational climate represents a domain of organizational description that translates observable features of organization-social systems into variables that have implications for organizational behavior. This discussion has presented the social system concept and developed the theoretical foundations of organizational climate. The research described, representing a joint effort of the IBR and NHRC, has resulted in the development of measures of organizational climate that function consistently in diverse types of organizations and that can provide useful guides to organizational development for Navy ships.

The present research is limited, however, by the fact that it focused on variables describing organizational structure and context as antecedents of climate rather than on related variables that reflect the operationalization of the conditions they represent. In further studies it is planned to include measures of representative behaviors of leaders (e.g. information giving), use of rewards, role-related behaviors (e.g. reaction to role conflict), socialization and acclimatization to group norms, and the like, which are expected to correlate more highly with climate and also to provide more direct indications for remedial action. If the research is extended to include behavioral measures that influence climate more directly, it is believed that the implications for diagnosis and remediation in organizational settings will be greatly enhanced.

REFERENCES

1. S. B. Sells, "An Interactionist Looks at the Environment," *Amer. Psychol.* **18**, 696–702 (1963).
2. S. B. Sells, ed., *Stimulus Determinants of Behavior*, Ronald Press Co., New York, 1963.
3. S. B. Sells, "Ecology and the Science of Psychology," *Multivariate Behav. Res.* **1**, 131–144 (1966).
4. S. B. Sells. "A Model for the Social System for the Multiman, Extended Duration Space Ship. *Aerospace Medicine*, **37**, 1130–1135 (1966).
5. T. Parsons, "An Approach to Psychological Theory in Terms of the Theory of Action," in *Psychology: A Study of a Science*, Vol. 3, S. Koch, ed., McGraw-Hill, New York, p. 612–712, 1959.
6. H. Leavitt, "Applied Organizational Change in Industry: Structural, Technological and Humanistic Approaches," in *Handbook of Organizations*, J. March, ed., Rand McNally, Chicago, p. 1144–1170, 1965.
7. L. R. James and A. P. Jones, "Organizational Structure: A Review of Structural Dimensions and Their Conceptual Relationships with Individual Attitudes and Behavior," *Organ. Behav. Hum. Performance* **16**, 74–113 (1976).
8. R. L. Payne and D. S. Pugh, "Organizational Structure and Climate," in *Handbook of Industrial and Organizational Psychology*, M. D. Dunnette, ed., Rand McNally, Chicago, 1976.
9. S. B. Sells, "Prescriptions for a Multivariate Model in Personality and Psychological Theory: Ecological Considerations," in *Multivariate Analysis and Psychological Theory*, J. R. Royce, ed., Academic Press, New York, pp. 103–122, 1973.
10. S. B. Sells, "An Approach to the Nature of Organizational Climate," in *Organizational Climate*, R. Tagiuri and C. H. Litwin, eds., Harvard University Press, Cambridge, Mass., p. 85–106, 1968.
11. S. B. Sells and E. K. Gunderson, "A Social System Approach to Long-Duration Missions," in *Human Factors in Long-Duration Spaceflight*, D. B. Lindsley, ed., National Academy of Sciences, Washington, D.C., pp. 179–208, 1972.
12. A. P. Jones and L. R. James, "Psychological and Organizational Climate: Dimensions and Relationships," IBR Tech Rep. #76–4, Texas Christian University, Forth Worth Tex., Sept. 1976.
13. B. P. Indik, "The Scope of the Problem and Some Suggestions Toward a Solution," in *People, Groups and Organizations*, B. P. Indik and F. W. Berrien, eds., Teachers College Press, New York, 1968.
14. D. Hellriegel and J. W. Slocum, Jr., "Organizational Climate: Measures, Research, and Contingencies," *Acad. Manage. J.* **17**, 255–280 (1974).
15. L. R. James and A. P. Jones, "Organizational Climate: A Review of Theory and Research," *Psychol. Bull.* **81**, 1096–1112 (1974).
16. B. Schneider, "Organizational Climates: An Essay," *Personnel Psychol.* **28**, 447–479 (1975).
17. A. P. Jones et al., "Psychological Climate: Dimensions and Relationships," IBR Tech. Rep. #75–3, Texas Christian University, Fort Worth, Tex., Dec. 1975.
18. Harry H. Harman, *Modern Factor Analysis*, rev. ed., University of Chicago Press, Chicago, Ill., 1967.
19. Ledyard R. Tucker, "A Method for Synthesis of Factor Analysis Studies," Dep. of the Army, Personnel Res. Sect., Rep. 984, 1951.
20. C. Burt, "The Factorial Study of Temperamental Traits," *Brit. J. Psychol., Statistical Section* **1**, 3–26 (1947).
21. L. R. James et al., "Relationships among Subsystem Context, Structure, Climate, and Performance from the Perspective of an Integrating Model," IBR Tech. Rep. #75–4, Texas Christian University, Fort Worth, Tex., Dec. 1975.
22. A. W. Halpin, "The Leadership Ideology of Aircraft Commanders," Lackland AFB, Tex., A. F. Personnel Training Research Center, Res. Rep. AFPTRC-TN 55–57, 1955.
23. A. P. Jones and L. R. James. Psychological and Organizational Subsystem Climate: Dimensions and Relationships. IBR Tech. Rep. #77–1. Texas Christian University, Fort Worth, Tex., Jan. 1977.

EARTH SCIENCES



O. G. (Mike) Villard, Jr., is a Professor of Electrical Engineering at Stanford University and a Senior Scientific Advisor at the Stanford Research Institute. Dr. Villard has been an ONR contractor for 25 years, working in the fields of meteor burst and other communications, radar and radar countermeasures, applications of ionospheric knowledge to problems in geophysics, space research, and defense electronics. He was a member of the Naval Research Advisory Council from 1967 to 1975, and Chairman from 1973 to 1975. He is also a former member of the Air Force Scientific Advisory Board. Dr. Villard did undergraduate work at Yale University and graduate work at Stanford University, where he received a Ph.D. in Electrical Engineering in 1949. He is a member of Union Radio Scientifique Internationale (Past Chairman of USA Commission III), the National Academy of Sciences, the National Academy of Engineering, Phi Bet Kappa, and Sigma Xi. In 1957 he received the Morris N. Liebmann Memorial Prize Award of the Institute of Radio Engineers, and in 1955 he was elected the "Outstanding Bay Area Engineer."

RADIO WAVE PROPAGATION IN THE SOLAR-TERRESTRIAL ENVIRONMENT: PERSPECTIVES FOR THE FUTURE

O. G. Villard, Jr.

Stanford University
Stanford, Calif.

Even in the daytime man's ability to see objects at a distance is variable, since it can be strongly affected by smoke, mist, and mirages. At such times, as well as at night, the Navy depends on radio waves in situations where the eye would otherwise be used. But radio waves, like light waves, are not immune to time and space variations imposed by the environment, and these restrictions must be understood and if possible avoided if the Navy is to use electromagnetic radiation as efficiently as it must.

Habituation makes it easy to overlook the enormous contributions made to the quality of civilian life by radio waves in their various forms. Broadcasts bring us news at breakfast or on the way to work; long-distance telephony (much of it handled by satellite or microwave repeater) helps transact the day's business; data networks using similar routes facilitate banking and virtually every aspect of commerce, and TV broadcasts provide the evening amusement or edification.

All of these functions (with the exception of the last) are required by our naval forces afloat, and a great many more beside. Since a moving ship is a self-contained unit and has no umbilical cord in the form of a bundle of wires connected to the Bell System, radio waves must be used for navigation, detection (radar), fire control, remote control (of bombs and RPV's), and of course communication. In addition, if the ship is a carrier, it is a moving platform from which other self-contained moving platforms (aircraft) operate.

Great as the benefits of using radio waves in naval operations are, their Achilles' heel must never be forgotten; transmissions of the normal kind inevitably betray the location, the type, and sometimes even the identity of the source. A naval combatant, unless it is groping along "blindfolded" with everything shut off, is roughly as conspicuous as a floating lighthouse. Thus each side in a naval engagement must be prepared to spoof or blind the other. The use of radio waves for intelligence and counteraction is an application whose importance easily equals that of the ones mentioned above. For every use of radio there is now a corresponding scheme or device capable of degrading effectiveness. It is no wonder that the art of detecting, locating, and then deceiving or otherwise effectively neutralizing an opponent's electronic assets has been dignified by the name "electronic warfare" (EW).

Since we live in air rather than in a perfect vacuum, all those systems and antisystems that depend on radio waves are affected to a greater or lesser extent by the atmosphere. Normally the air can be ignored, but there are many situations in which it cannot.

Radio waves have been in naval use for over 70 years, and it might seem surprising that the details of their propagation cannot be said to be perfectly

understood even now. This expectation would be reasonable if requirements for propagation knowledge remained static. But military technology is constantly expanding in complexity and sophistication; new uses and new precision of older uses require constant improvement of our understanding of the way in which radio waves interact with our surroundings.

Our "surroundings," in this instance, can be divided into two major categories, in addition to the neutral gas (consisting of air, water vapor, etc.) with which we are concerned at low altitudes, there is the invisible ionized component higher up. The energy of sunlight knocks electrons out of gas atoms or molecules to produce ions; the free electrons, being charged and light in weight, have a surprisingly strong effect on radio waves. The effect is most profound at the lower radio frequencies, but it is noticeable even at microwave frequencies. Even the positively charged ions can affect the longest radio waves of interest to the Navy. Ionization of our atmosphere is significant at heights from 60 to tens of thousands of kilometers; it is strongest in the 200-400 km interval, where about 1 atom out of 1000 is ionized.

SOME CURRENT RESEARCH

Some randomly selected examples of recent research results may help set the scene for comments on future trends and possibilities.

Improving Communication with Submerged Submarines

No one will dispute the necessity for the national command authorities to be in contact with attack- or ballistic-missile submarines at all times. Modern nuclear submersibles of either type are able to operate for long periods at very great depths.

Their commanders understandably prefer to stay as far down as possible, since the safety of a submarine depends on concealment and concealment is best when there is plenty of water between the submarine and any possible attacker.

It is well known that the longer the radio wavelength, the deeper the wave can penetrate into saltwater. The present-day standard system for submarine communication uses Very Low Frequency (VLF) transmission (roughly 20 kHz,) which can be received at depths on the order of 100-200 ft (30-60m). The Navy would like to supplement VLF with the—alas!—controversial Seafarer (formerly Sanguine) system whose waves, some 400 times as long, suffer far less attenuation in seawater.

Since the frequency interval to be used for Seafarer is not far from the world powerline frequencies of 50 and 60 Hz, it may come as no surprise that the "antenna" takes the form of a *buried* wire, rather than one suspended from a mast.

Communication of sorts could in principle be maintained at even greater depths if still longer waves could be employed. Serious consideration is, in fact, being given to the use of frequencies in the range from 0.5 to 2 Hz, where the free-space wavelength would be on the order of 10 times the circumference of the Earth. Of course, enormously long waves such as these carry information at a tortoiselike pace, but this is not so serious a disadvantage as it might seem. The transmissions could still perform an alerting function, effectively advising a submarine to come closer to the surface, where it can receive more detailed instructions on a different waveband.

The essential problem in any of the systems using very long waves is the problem of launching them efficiently from structures of affordable size. As a result, a number of ingenious schemes are being explored. In one approach, plain old-fashioned induction fields, such as were tried and discarded in the earliest days of radio, are being considered. Another suggestion is to radiate from an electrically conducting column of gas in the sky, using electrons knocked temporarily free by collisions with particles beamed vertically from a high-energy accelerator. Such a column would represent an essentially massless and practically indestructible antenna. Somewhere in the collection of possibilities lies the practical answer.

Propagation research in support of the submarine communication mission takes many forms. For example, in the 50 Hz-to-100 kHz part of the radio wave spectrum the waves are deflected downward and thus prevented from escaping into space by the lowermost part of the iono-

sphere, where the gas is comparatively stable owing to its relatively high density. (This is why VLF is so effective for time signals and navigation.) Even here, though, the reflection height changes appreciably from day to night, thus changing the mode structure and leading to wave interference. As a result, signals crossing the sunset or sunrise lines may undergo an undesirable amount of strength variation.

Furthermore, particle bombardment such as would accompany high-altitude nuclear explosions can also cause signal strength changes. Although treaty limitations of course prevent use of nuclear devices for testing, nevertheless infrequent bursts of natural radiation of various kinds can give rise to rather similar disturbances. By judicious extrapolation these natural events can be used to verify theoretical models from which the nuclear environment can be predicted.

It was once thought that the only source of significant incident radiation was the sun, which is characterized by occasional flarelike outbursts. We now know that the Earth carries around with it its own store of radiation. This takes the form of high-energy particles trapped by the terrestrial magnetic field in what is left of the earth's atmosphere in the height range from 500 km to several tens of thousands of kilometers. At such heights, most of the gas is ionized by incident solar radiation, and the particles are few enough to be contained by the magnetic "pressure." This extremely tenuous "magnetosphere," of global dimensions, where electron mean free lifetimes are measured in hours, has provided as rich a hunting ground for new physical effects as Africa provided for wild animals in the days of the early explorers. For example, in the magnetosphere radio waves can either add energy to or abstract energy from the particles, in a manner reminiscent of, but only distantly related to, the processes of maser or laser amplification. This interaction between waves and particles makes the region surprisingly dynamic; the distribution of particle and wave energy is constantly changing.

Interestingly, the radio waves responsible for all this activity can be either natural or manmade. If strong enough, they can cause some of the trapped particles to be released into the lower ionosphere in quantities sufficient to perceptibly affect propagation of waves of interest to the Navy traveling in the earth-ionosphere "waveguide." (See Figure 1.)

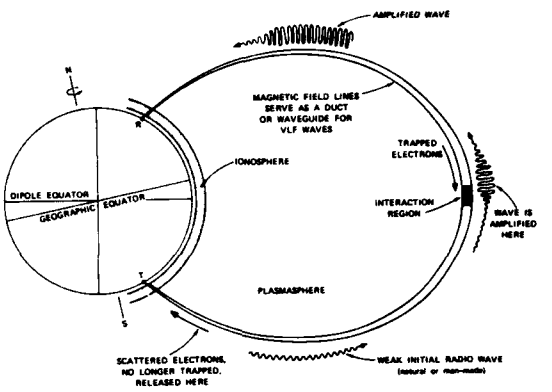


Figure 1–Far above the equator, energetic electrons from the solar wind are trapped by the earth's magnetic field. Spiraling around magnetic lines of force, they travel from hemisphere to hemisphere, reversing direction after each transit at "mirror points" well above the ionosphere.

Natural or manmade VHF signals (for example, "atmospherics" caused by lightning flashes) follow similar paths, except that they travel essentially from surface to surface. Within the "interaction region" shown, some electrons give up energy to waves, which are thereby amplified, but in the process the interacting electrons become untrapped. Such spilled electrons penetrate lower into the ionosphere before giving up the rest of their energy in collisions. They thereby change the electron density of the radio-wave reflecting layers and give rise to signal-fading effects.

Efforts are underway to measure the space and time variation of the streams of energetic particles by means of instruments carried in satellites. The aim is to predict the effect of charged-particle spills on propagation at Seafarer and other frequencies.

The complexity of the various wave-particle interactions is fascinating to contemplate. For example, it now appears that a burst of particles from the Sun (effectively a gust in the solar wind) can impart energy to the radio noise background, and effectively amplify it, thus giving rise to a noise emission at VLF (5-15 kHz). Not surprisingly, manmade signals in this band, such as the navigational service Omega, may also be amplified. At the same time, energetic particles spilled from their magnetic-field "traps" by such a disturbance [1] can cause a change in received signal strength at 100 kHz, and the basic disturbance itself can additionally give rise to a spontaneous emission in the micropulsation band, from 0.1 to 10 Hz. The relationships among these

various events are just now being perceived, and their implications in possible Navy communication systems of the future are beginning to come into focus.

Reversibly Remodeling the Ionosphere For Communication Purposes

The following research was motivated by the perennial need for beyond-line-of-sight communication at VHF. Although some remarkable capabilities were uncovered, the attendant cost proved to be not inconsiderable, so that at the moment other options seem more attractive. Nevertheless, like all good research, this opens vistas whose extent we cannot at the moment fully perceive; for example, it is possible that knowledge of ionospheric plasma behavior gained by this means may help us understand and explain ionospheric characteristics of immediate importance, such as the unexpected scintillations that affect satellite radio transmissions under certain conditions.

By way of background, we may recall that Kennelley and Heaviside in 1902 postulated the existence of an electrically conducting region in the upper atmosphere, to explain Marconi's success in communicating across the Atlantic. Until very recently, users of the ionosphere have had to be content with whatever reflections nature happened to provide. Therefore it can be said that something of a landmark in the history of man's control of his environment was passed in April 1970, when a team directed by W. F. Utlaut of the Department of Commerce at Boulder, Colorado, succeeded in causing a substantial (but, happily, reversible) change in the reflecting properties of the principal ionospheric layer by heating it with the aid of a very high power radio transmission [2]. The antennas they used are shown in Figure 2.

The underlying principle is analogous to heating foodstuffs containing moisture in a microwave oven. Both water and the electron "gas" of the ionosphere are imperfect—and therefore lossy—dielectrics. But there is this difference: foodstuffs are confined in an enclosed cavity, so that energy piped in has essentially no place to go but into the water. In the case of the ionosphere, radio waves tend to either travel right through or

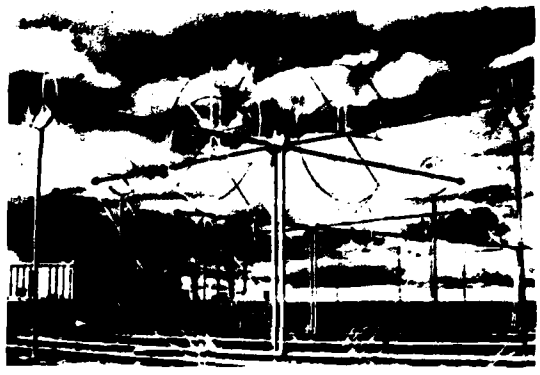


Figure 2–From this bizarre collection of wires and aluminum irrigation pipes, 10 million watts of power are radiated straight up. When the right radio frequency is chosen for ionospheric heating, "plasma" 300 km overhead is modified sufficiently to produce a tenfold increase in its normal radio-wave reflecting power. Effects disappear shortly after the heater is turned off.

be completely reflected, in either case losing little strength. Fortunately there proves to be a wonderfully simple trick that can be played on the waves, and that is to make the frequency of the heating transmission very close to the so-called "plasma" frequency, or, as radio engineers would say, the "critical" frequency of the layer at its densest part. As this frequency is approached, the heating waves slow down drastically in their speed of travel, with the result that there is ample time for them to lose a substantial fraction of their energy to dielectric losses during their passage. The idea of heating the ionosphere in this manner goes back a long time, but modern interest in the matter was sparked by calculations that suggested that measurable effects could be achieved with an affordable investment in equipment. The Soviet scientist A. V. Gurevich, who made such a prediction in 1962, deserves the credit [3].

It was originally thought that heating would result in expansion of the affected region, giving rise to a dome or incipient bubble roughly 160 km in diameter. This does, in fact, take place. But while observing the magnitude of this effect with vertical-incidence sounders, the Boulder group discovered to their astonishment that the heating was also causing to appear a condition known as "spread F," in which a clearly defined layer echo becomes extended (either in frequency or in slant

range), as if the otherwise homogeneous region had become corrugated.

Such spreading occurs sporadically in Colorado as a natural event. At more northerly latitudes it is seen much more frequently, usually in association with auroral disturbance. Since the "artificial spread F" appeared when the heating transmissions were turned on and disappeared when they were shut off, the Boulder investigators received the impression that they had at least one facet of auroral disturbance under human control! (See Figure 3.)

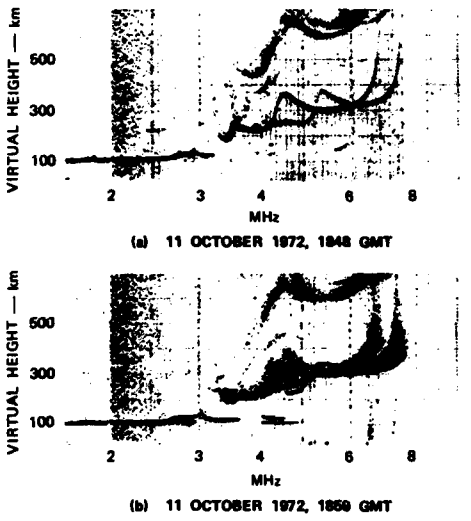


Figure 3–An example of artificial "spread F." The lower echo-sounder trace in part (a) is a first-order reflection from the unmodified ionosphere and shows a typical variation of time delay with radio frequency, plus reflection from a smooth surface. The spread appearance in part (b), characteristic of considerable layer roughness, was caused by several minutes' operation of the heater in Figure 2.

Even more unexpected was the discovery, by a group studying the effect of artificial layer tilts on radio direction finding, that the heating was also rearranging the electrons of the affected layer in such a way as to permit VHF signal transmission at distances far beyond the line of sight, provided that certain geometrical requirements were satisfied. The practical effect of this rearrangement was as if there had been created, 200 to 400 km above the Earth, a large number of evanescent thin columnar reflectors, each with its major axis aligned in the direction of the Earth's magnetic field. (See Figure 4.)

Now the highest frequency that is (on rare occasions) returned to Earth by the normal ionosphere is roughly 40 MHz. The highest frequency returned to Earth by the heating-associated reflectors in usual strength is in the order of 10 times that value. Thus the additional channel width thereby opened up is impressive [4].

To generate these reflectors requires heating power on the order of 100 kW if a large antenna system is used, or 1 MW if a simpler array is employed. Either way, the capital investment is not inconsiderable.

The fact that the ionospheric reflectors are directional imposes constraints on the choice of transmission paths, but jamming and intercept of circuits thus established becomes proportionately more difficult.

Heated ionospheric gas is not vulnerable to physical attack in the same sense as is, say, an orbiting satellite. To the author's knowledge, the effect of nuclear explosions on artificial spread-F communication has not yet been considered. It is known that high-altitude nuclear explosions generate effects rather similar to the natural aurora. Therefore it seems that nuclear events would be more likely to enhance, rather than diminish, artificial propagation. After the explosion, the heater presumably could be turned off until normal conditions returned.

While layer profile changes associated with ionospheric heating can be large enough to degrade the accuracy of present-day direction finders (which of course depend on the tacit assumption that the reflecting layers are for all practical purposes concentric with the earth), the logistic problems attendant on attempting to modify an area the size, say, of the Mediterranean Sea, turn out to make the scheme relatively unattractive.

Outwitting Satellite Signal Scintillations

Heating the ionosphere has uncovered some interesting new possibilities for future naval communication. It has also had an indirect payoff because it has brought to light unexpected new

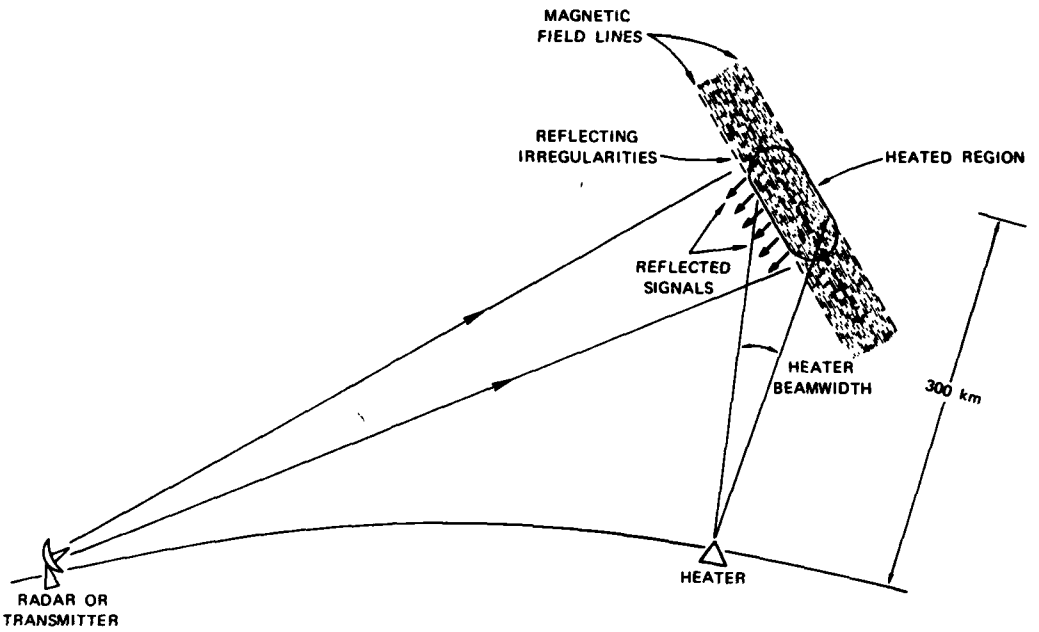


Figure 4—Heating creates the equivalent of reflecting irregularities in the ionosphere, elongated in the direction of the earth's magnetic field. These form powerful, if highly directional radio-wave reflectors.

properties of plasma, which may help in understanding the surprising fading (or scintillation) observed on microwave transmissions to and from satellites. This is observed when the line of sight to the satellite passes through the equatorial (and, occasionally, the auroral) ionosphere. Fading ranges of 7 and 3 dB, peak to peak, have been measured at 4 and 6 GHz, respectively. An amount of fading as small as this might not seem serious, but satellite circuits tend to be operated with very low signal-to-noise ratio margins, so that even a small degradation has a noticeable effect. (See Figure 5.)

The story of the discovery of microwave ionospheric scintillation is interesting. The possibility that waves of 5-cm length could be affected appreciably by passage through the upper ionosphere, where electron mean free paths exceed 1 km, was once regarded as wildly improbable.

Scintillations can be caused only by irregularities in refractive index along the line of sight, either moving or time-varying. To have a strong effect, such irregularities must be at least roughly comparable with the wavelength in size. But a mean free path of given length tends to smooth out variations in electron density between any two points spaced closer than that length. Therefore it was difficult to imagine any arrangement of electrons either physically small enough or dense enough (or both) to interact significantly with such shortwaves.

Before the space age, knowledge of the extent to which the ionosphere refracts or perturbs radio signals passing completely through it was derived almost entirely from measurements using the so-called radio stars. These represent essentially pointlike signal sources superimposed on a background continuum. Both signals and continuum are time-varying and noiselike. Although easy to pick out at VHF, the radio stars are progressively more difficult to identify against the background as microwave frequencies are approached. Most star measurements, therefore, were at VHF and led investigators to deduce values on the order of 1 km for the crossfield scale size of the scintillation-producing irregularities. This deduction was quite correct for the radio frequencies employed. However, in interpreting

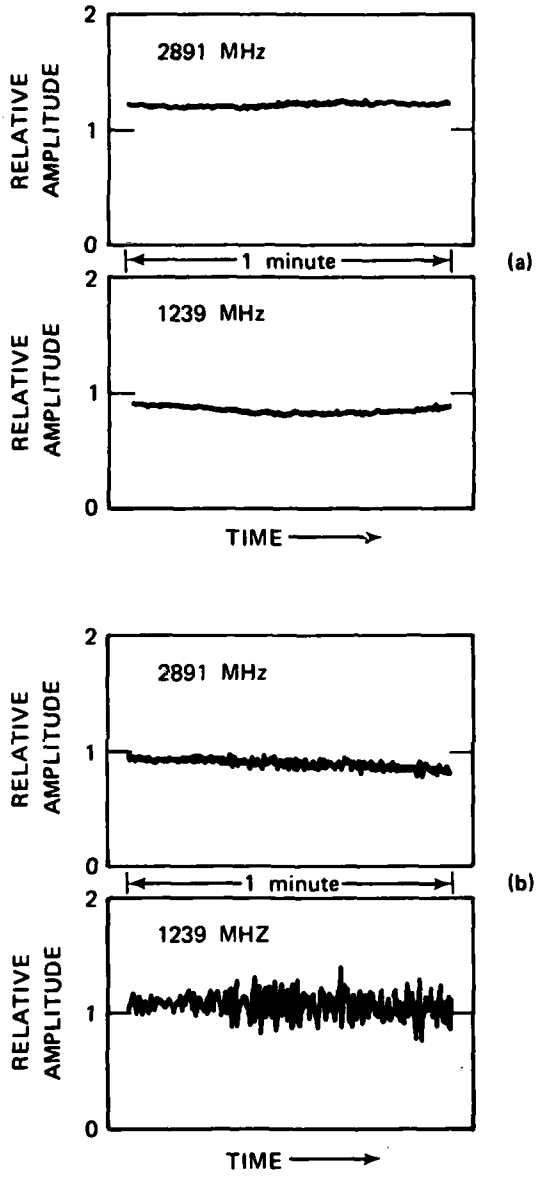


Figure 5–An example of fading imposed by the ionosphere on a 25-cm signal from a satellite. (A longer wavelength is shown as reference.) Although far above the radio-reflecting layers, the source (P76-5, portrayed in Figure 7) is nevertheless moving with respect to the receiver at Ancon, Peru. Part (a), at 0344 GMT, shows the normal condition; 4 min later the line of sight is passing through electron-density irregularities, as shown in part (b), at 0348 GMT.

In this instance most of the fluctuations are due to motion of the satellite. However, since the irregularities are also drifting in position, even transmission from nonmoving (geostationary) satellites show similar fading when irregularities are present. (Record courtesy of the Defense Nuclear Agency.)

these measurements a Gaussian form for the spatial distribution of the irregularities had been postulated. Such an assumption, together with the above deductions derived from observations, led to a considerable underestimation of the magnitude of scintillation at frequencies much higher than the original observing frequency, which is why microwave effects over the equator were so unexpected.

In the years since the discovery of microwave scintillation, two important revisions of the original interpretation have come forward. First, it has been found that a power-law spectrum is a much more realistic approximation than its Gaussian counterpart. It is, in fact, the spectrum shape that describes the way turbulence breaks down into eddies of ever-diminishing size. In addition, it is now appreciated that scintillation measurements are subject to an effect called Fresnel filtering; a measurement at a particular radio frequency, when a low-pass spatial spectrum is present, tends to be most sensitive to electron-density fluctuations comparable to the size of a Fresnel zone. (See Figure 6.) (For a distant transmitter, the radius of a Fresnel zone at a distance z from the receiver for a signal wavelength λ is $\sqrt{\lambda z}$. Thus, the radius depends on both λ and z.) Out of a

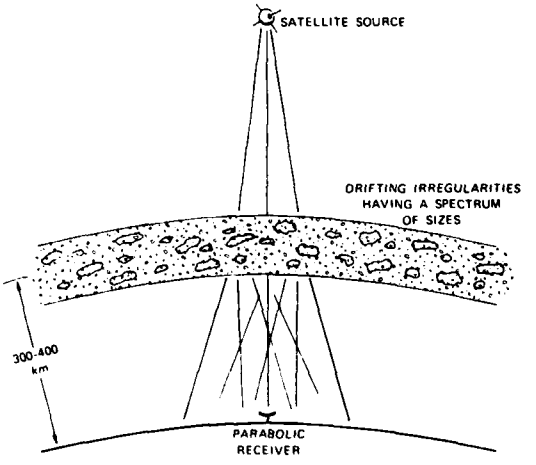


Figure 6–The geometry of "Fresnel filtering." At any given radio frequency, when a normal "low-pass" spatial spectrum of irregularities is present, the received signal is most disturbed by those irregularities whose size is comparable with that of a Fresnel zone at the irregularity height.

low-pass spectrum of irregularities of different sizes, measurement of amplitude fluctuations (or scintillation) at a single radio frequency will tend to favor—to a surprising degree—those irregularities having a size close to that of the Fresnel zone. Correct extrapolation of scintillation data to predict effects at other radio frequencies or distances unfortunately requires accurate knowledge of the irregularity spatial spectrum. A straightforward way to make such measurements calls for data from one source at a variety of radio frequencies; this was not feasible prior to satellites and cannot be done very readily even now.

Once the full ramifications of Fresnel filtering were appreciated and initial direct measurements of the ionospheric spatial spectrum by satellite-borne probes became available, more accurate predictions became feasible. The more plausible choice of a turbulencelike power-law spectrum (a three-dimensional index approximately equal to 4 is reasonable) certainly falls off less rapidly with increasing spatial frequency that does the Gaussian. But even the above power-law spectrum, by itself, is not sufficient to account for the observed levels of scintillation at GHz frequencies.

Two main lines of thought have arisen in attempts to explain the observations. Both are alternatives to postulating unrealistically high electron densities or implausibly strong spatial modulation of the density. One is the suggestion that the region of structured plasma, where scattering occurs, may encompass not only the equatorial F layer but also an appreciable fraction of the magnetosphere, possibly out to several earth radii [5]. Such a thick region would enhance scintillation at all frequencies by virtue of the very long raypaths through the structure region.

The second explanation takes into account the concept of Fresnel filtering plus the best available estimates of the underlying spatial spectrum of the scattering irregularities in deriving a frequency dependence for the scintillation. It then proposes that localized nonmonotonic features in the spatial spectrum (which might be called "spatially resonant plasma instabilities" are responsible for GHz scintillation [6].

Although the vast majority of measured ionospheric spatial spectra show monotonically decreasing (turbulencelike) behavior, there are some interesting exceptions. Certain measurements actually made in the topside equatorial ionosphere show distinct spatial resonances (i.e., regularities) at wavelengths between 1 and 10 km [7]. If similar events also occur in the wavelength regime between 0.1 and 1 km, which seems plausible, they could greatly enhance the magnitude of GHz scintillation for a given level of VHF-UHF scintillation.

Since the above spatial regularities were observed within a few degrees of the magnetic equator, and in the same local-time hours in which GHz scintillation occurs, there is clearly reason to suspect that similar resonances at somewhat smaller scales might be responsible for the surprisingly strong GHz effect.

There turns out to be no reason to view the two leading hypotheses about the origin of equatorial GHz scintillation—an extended plasmaspheric layer and nonmonotonic spatial spectra—as mutually exclusive. What is needed is better understanding of the way irregularities are distributed in height and of the circumstances under which they can have size distributions that differ from that characteristic of the decay of turbulence.

An experiment potentially able to provide valuable new information was begun on May 22, 1976, with the launching of the P76-5 satellite, carrying the Defense Nuclear Agency (DNA) 002 coherent beacon. Orbiting at a height of 1000 km, in a nearly circular but highly inclined orbit, this probe



Figure 7–Artist's conception of the P76-5 payload, launched in May 1976. It carries the most comprehensive experiment yet devised for measuring ionospheric irregularities, including those that produce fading and scintillation of signals from geostationary satellites. A comb of coherent radio frequencies from 147 through 2891 MHz is radiated.

will radiate a "comb" of radio frequencies from VHF to SHF, all coherent in phase. In the past, space probes have provided only one or two frequencies for study, typically noncoherent and only available incidentally to another mission. (An exception is the ATS-6 satellite, but its highest coherent frequency is 360 MHz.) In the DNA experiment, mapping the ionosphere is the central theme; the spread of frequencies is wide, and the fact that the phase is coherent permits collecting considerably more information (such as total electron content and its variations, and data from which crude images can be constructed) that would otherwise be possible. Figure 7 is an artist's conception of the satellite and its orbit.

FUTURE POSSIBILITIES

Avoiding the Radio Mirages

As our radio "vision" becomes progressively sharper, there is a continuing need to fashion lenses (so to speak) to correct deficiencies when that is possible. A major thrust at the present time is on improving knowledge of weather in the troposphere, where invisible water vapor can and does strongly interact with radar beams. For example, electronically steered ballistic-missile-warning and satellite-tracking radars, such as the AN-FPS 85, occasionally encounter beam bending and distortion when looking for unknown targets close to the horizon. This is a result of a disturbance of the normal distribution of water vapor with height, and is closely related to the conditions that sometimes cause FM and TV signals to span unusually long ranges. Continuing research on the lower atmosphere with acoustic and radio sounders has given remarkable new insights into the details of these exceptional refractive-index conditions, and satellite photography has made it possible to determine remotely, and in real time, the areas affected by a given event [8].

Very long range radars can now correct (at least to some extent) for distorted propagation of this sort, by tracking (as a side exercise) some of the many known satellites that come whizzing by. If a familiar orbiting object is seen apparently to waver in its course when passing through a certain region of the sky, that waver can be recorded and applied to correct the apparent track of an unknown object just coming into view for the first time in roughly that same direction. There are obvious limitations to what can be done here, because the atmosphere is not stationary, but nevertheless quite useful first-order corrections can be made.

Although naval forces afloat may not be able to compensate their radars by use of itinerant objects in space, improved predictions of radar performance are being introduced to good effect. Because atmospheric conditions over water are far stabler than over land, unusual events such as inversions tend both to be larger in geographical extent and longer enduring. They are at the same time more readily predictable.

A carrier task force needs to know how far away its radars can likely be heard (so as to know the intercept range), how far away the radars can detect objects of interest, and whether there exist "holes" in the coverage patterns within which a target would escape detection. Better predictions and real-time measured data combined with new procedures such as essentially instantaneous ray tracing (made possible by low-cost computers) is bringing about solid improvement, and no end to this trend is yet in sight.

Measuring the Ocean's Moods Without Going to Sea

A line of investigation conceptually rather close to propagation research, but not quite the same thing, is study of the electromagnetic nature both of targets and the background from which target signatures must be extracted. Often propagation characteristics and the details of background clutter are interrelated, so that interpretation of one cannot be accomplished without consideration of the other. Studies of this kind often lead to unexpected and useful results. For example, radar reflections from the sea surface, using a variety of radio frequencies and platforms, have led to what has been called "radio oceanography" [9]. Information on sea state is transferred to radio waves after scattering or reflection and can be observed back at a distant source of illumination. Both high-frequency radar and sounders in satellites can monitor oceanic conditions at great distances

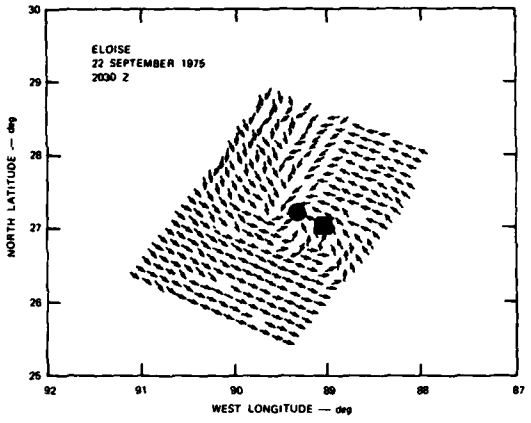


Figure 8–The black dot represents the position of Hurricane Eloise moving through the Gulf of Mexico at 2100 GMT on September 22, 1975, as determined by an experimental ground-based HF radar 3000 km distant in California. Arrows represent surface wind directions, derived by analysis of radar clutter from water waves. Although this was an ad hoc test, the radar-determined "eye" is only 35 km from the corresponding position (the square) deduced by the National Oceanic and Atmospheric Administration from satellite photographs and reconnaissance aircraft reports. Radar accuracy can undoubtedly be improved further. Storms can be tracked by this means for extended period of time at comparatively low cost.

in real or near-real time. The former, however, permits continuous "looks" at a given point on the sea surface and is less expensive to establish. A disadvantage is that it suffers outages from time to time. But it can also pinpoint and follow hurricanes. (See for example Figure 8.) HF radar can also indirectly measure surface currents, even localized currents generated by barometric forces and transient wind systems (as contrasted with major oceanic currents put in place by gross features of the global atmosphere circulation).

Real-time remote measurement of ocean currents and sea state is of clear economic importance in ship routing where the object is to minimize elapsed time and fuel consumption by readjusting a ship's course at frequent intervals to avoid regions where higher-than-average waves result in speed loss. Amphibious military operations also need wave and current information. Sea-state data are of further importance to the Navy because underwater sound generated by breaking waves represents a background noise that limits the detection range of sonar systems. Also, of course, high waves limit many kinds of Navy operations. It seems very likely that in the future, sea state will be reliably measured by shore-based means. In addition to improving forecasts, such data should also relieve ships' crews of the necessity to collect and send in oceanographic information as at present. This would be especially valuable in wartime or any other time when heightened tensions make a reduction in radio traffic desirable and mandatory.

Doing Something About the Radio Weather

Both ionospheric "weather" and its more familiar meteorological counterpart have been standard conversation starters over the years. Future generations, however, may find themselves deprived of that particular opening gambit as the ability to forecast and even modify our environment grows. At the present time the Navy-inspired Solrad series of satellites continuously checks the sun's output of radiant energy in those wavelength bands exerting the strongest influence on both short- and long-term ionospheric behavior. Since there is a time delay between causative radiation fluctuations and the resulting change in radio-reflecting power, such events can usually be anticipated in time to broadcast warnings to the fleet. Thus communicators can alter transmission frequencies, alter message routings, and take other steps to maintain an orderly flow of traffic. In the past, it not infrequently happened that the only warning of impending trouble was a major circuit failure. The new procedure should be of great benefit to all those systems that in any way depend on the ionosphere.

Since the atmospheric gas above, say, 100 km in height is highly tenuous (about equivalent, for example, to the vacuum of an inexpensive thermos bottle), the possibility of modifying its radio-wave reflecting characteristics to make them more useful is not as farfetched as it sounds. We know that the upper atmosphere is strongly affected by nuclear explosions. We know it is also measurably affected by thunderstorms, large blasts using conventional ammunition, and similar energy-releasing events (including tsunamis and landslides), on the earth's surface. We also know that when low-ionization-potential chemi-

cals such as barium or caesium are released at the right height, the local electron density can be materially increased for a matter of hours, assuming the region to be in sunlight. Conversely, the deliberate or accidental discharge of water into the ionosphere (from rockets or rocket exhausts) is very effective at causing free electrons to disappear; a localized decrease in their density results [10].

Water, of course, is very much a natural part of our atmosphere, so that a question of pollution does not arise. But a clearly nonpolluting technique for ionospheric modification is the radio-wave heating method. It can be expected that as knowledge of the details of radio-wave-induced effects improves, additional applications may well be found. Thus far, for example, there have been no studies (to the best of the author's knowledge) of the possibility of combined radio-wave and chemical modification. Extra electrons, released at the appropriate height, could raise the radio frequency at which heating is efficient, thereby significantly reducing the size and cost of the heating installation required.

Modifying the ionosphere to increase reflecting power is of obvious assistance in communication applications. (Such a reflector would have the unique advantage of nearly instantaneous controllability.) However, there are many other conceivable applications, many of them in the electronic warfare area, that have not yet been fully explored.

Another radio-wave modification of importance would be to *decrease* the amount of ionization in a given region, since ionospheric electrons represent a not-inconsiderable source of clutter for earthborne or spaceborne radars and communication systems which must transmit signals all the way through that region. This includes high-resolution side-looking radars, precision location and navigation systems, and the like.

Atmospheric nuclear weapons tests, and to some extent the natural aurora too, can create extra electrons in the ionosphere capable of making the targets of space-tracking radars appear to scintillate in position and grow either weaker or stronger. To dissipate these extra electrons by means of a powerful radio beam, which in effect "burns through" the affected region, has often been proposed. But the estimated energy requirements have thus far dampened prospects for this technique.

Can Induction be Substituted for Radiation?

Radio waves of really enormous lengths have never been of interest in commercial communication, because much more information can be transferred at lower cost in the MF and HF range. Geophysical prospecting does use this wavelength regime, but prospecting is normally concerned with analytical measurements at a particular location, rather than information transfer over long distances. Thus, communication technology in the 1-to-100-Hz range can hardly be said to be a mature art. Many of the concepts are relatively unfamiliar. For example, the normal variation in electron density resulting from changes in gas pressure with altitude takes place over a distance that is a tiny fraction of a wavelength in the case of superlong waves. The ionosphere, in effect, is a very thin shell. In addition, the effect of both electrons and ions needs to be taken into account, whereas at higher frequencies only electrons need be considered in calculating refractive index.

Although it is tempting to apply in this wavelength regime concepts and simplifications that have proven useful elsewhere in the frequency spectrum, such extrapolation is very risky. It may be preferable, for example, to abandon the concept of "radiation," implying as it does a decay of signal strength inversely as the distance, and to make use instead of a field component whose strength decreases with the square of distance. Optimum launching and retrieval of this field component might well lead to structures scarcely resembling conventional transmitting and receiving antennas at all. Whether these structures, when performing a given function, will be adequately low in cost remains to be established, but there is that hope.

Although the point is not immediately obvious, extremely long waves also have potential for detection and localization just as do their shorter counterparts. One can think, by way of illustration, the longest "wave" of all is a static or d.c. magnetic field. Let it be perturbed at a given point by (for example) a magnetic object. By measuring the detailed spatial distribution of the total field

over some aperture at another location, it is possible to deduce the position of the object, but only if the measurements can be made with sufficient precision. If this is feasible with a static field, it can also be done with a time-varying magnetic field, even when the time variation is comparatively slow. This procedure is greatly aided by digital recording techniques that make possible both easy storage and rapid processing.

CONCLUSION

Only a few of the more challenging electromagnetic propagation matters of potential interest to the Navy have been touched upon here. (For example, the many intriguing problems associated with laser communication and weaponry have been omitted.) Propagation is a research field that offers a delightful mix of physical effects, spanning as it does the frequency spectrum from 1 to more than 10^{10}Hz, and dealing as it does with transmission through materials as diverse as saltwater and the near vacuum of outer space. As electronic systems grow more complex and as the precision required of them grows ever greater, research must keep pace if the Navy is to retain its leadership in harnessing and exploiting the environment.

REFERENCES

1. T. J. Rosenberg, R. A. Helliwell, and J. Katsufrakis, "Electron Precipitation Associated with Discrete Very Low Frequency Emission," *J. Geophys. Res.* **76,** 8445 (1971).
2. W. F. Utlaut, "An Ionospheric Modification Experiment Using Very High Power, High Frequency Transmission," *J. Geophys. Res.* **73** (31), 6402–6405 (1970).
3. A. V. Gurevich, "Radio Wave Effect on the Ionosphere in the F-Layer Region," *Geomagn. Aeron.* **7,** 291 (1967).
4. Special Issue on Ionospheric Modification by High Power Transmitters, *Radioscience.* **9** (11), (Nov. 1974).
5. H. G. Booker, "The Role of the Magnetosphere in Satellite and Radio-Star Scintillation," *J. Atmos. Terr. Phys.* **37,** 1089–1098 (1974).
6. A. W. Wernik and C. H. Liu, "Ionospheric Irregularities Causing Scintillation of GHz Frequency Radio Signals," *J. Atmos. Terr. Phys.* **36,** 871–879 (1974).
7. P. L. Dyson, J. P. McClure, and W. B. Hanson, "In Situ Measurements of Amplitude and Scale Size Characteristics of Ionospheric Irregularities," *J. Geophys. Res.* **79,** 1497–1502 (1974).
8. S. M. Serebreny and R. H. Blackmer, Jr., "Satellite-Viewed Cloud Cover as a Descriptor of Tropospheric Radio-Radar Propagation Conditions," Final Report, SRI Project 7940, Stanford Research Institute, Menlo Park, Cal., Feb. (1974).
9. Special Issue on Radio Oceanography, *IEEE Trans. Antennas Propag.* **AP-25** (1), (Jan. 1977) (in preparation).
10. M. Mendillo, G. S. Hawkins, and J. A. Klobuchar, "A Sudden Vanishing of the Ionosphere due to the Launch of Skylab," *J. Geophys. Res.* **80,** 2217 (1975b).



Norbert Untersteiner has been Project Director of the Arctic Ice Dynamics Joint Experiment (AIDJEX) since 1971. In 1969 he and Dr. Kenneth L. Hunkins formed the initial scientific plan for the project. Dr. Untersteiner was Assistant Professor of Meteorology at the University of Vienna from 1951 to 1956. From 1957 to 1962 he was Resident Meteorologist at the Central Establishment for Meteorology and Geodynamics at Vienna. He became an Associate Professor of Glaciology at the University of Washington in 1963, and in 1967 was named a full Professor. He served as a consultant to the Rand Corporation from 1965 to 1972. He was Chairman of a committee of the National Academy of Sciences charged with developing a scientific program for a repetition of Fridtjof Nansen's historic drift across the Arctic Ocean. Dr. Untersteiner received a Ph.D. in Geophysics from the University of Innsbruck in 1950. He is a member of the International Commission of Polar Meteorology, the World Meteorologic Organization, the Committte on Polar Research of the National Academy of Sciences, the International Union of Geologists and Geophysicists, AAAS, and the American Geophysical Union. He is Vice President of the International Commission on Snow and Ice. In 1960 h received the Austrian Honorable Cross in Arts and Sciences.



Kenneth L. Hunkins is an Adjunct Professor and Senior Research Associate at Columbia University's Lamont-Doherty Geological Observatory, where he has been employed since 1960. He has participated in a number of Arctic Ocean research expeditions and in several oceanographic cruises in the North Atlantic Ocean. Dr. Hunkins received a B.Sc. in Physics from Yale University in 1950 and M.Sc. and Ph.D. degrees in Geophysics from Stanford University in 1960. He is a member of the Oceanographic Advisory Committee to the Secretary of the Navy, of the American Geophysical Union, and of Sigma Xi. He is a Fellow of the Arctic Institute of North America and of The Explorers Club.



Beaumont M. Buck founded the Polar Research Laboratory, Inc., in 1973 and now serves as its President. Mr. Buck served in the U.S. Navy from 1948 to 1961, including 3 years in the Electronics and Undersea Branch of ONR, and from 1961 to 1973 was Head of the Ocean Surveillance Section of the General Motors Defense Research Laboratory. He has led 23 field experiments in acoustics in the Arctic and Bering Seas. Mr. Buck received a B.S. in 1948 from the U.S. Naval Academy, a B.S. in Electronic Engineering in 1954 from the U.S. Naval Postgraduate School, and an M.S. in Applied Physics from the University of California, Los Angeles, in 1955. He is a member of the Acoustical Society of America and of the Technical Committee on Underwater Acoustics of that society.

ARCTIC SCIENCE: CURRENT KNOWLEDGE AND FUTURE THRUSTS

N. Untersteiner

University of Washington
Seattle, Wash.

K. L. Hunkins

Columbia University
New York, N.Y.

B. M. Buck

Polar Research Laboratories
Santa Barbara, Calif.

INTRODUCTION

The Arctic Ocean is a landlocked body of water covering the area around the North Pole. It is bordered by two continents and a subcontinent: Eurasia, North America, and Greenland (Figure 1). It is the fourth largest ocean, exceeded in size only by the Pacific, Atlantic, and Indian Oceans. The Mediterranean Sea is only one-fourth the size of the Arctic Ocean. It is a true ocean with depths in the deep basins averaging 3500 m and reaching as deep as 5000 m. Shallow shelf seas reaching widths of 700 or 800 km, the widest in the world, surround these basins. Another truly oceanic aspect of this north polar sea is the presence of the world-girdling midoceanic ridge system. The Mid-Atlantic Ridge system extends northward between Spitsbergen and Greenland and into the Arctic Ocean. This portion of the midoceanic ridge separates the Eurasia and American Plates.

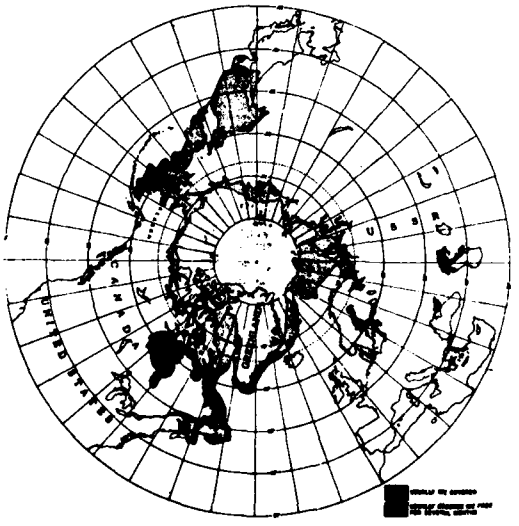


Figure 1—Ice over of the arctic regions [1]

One of the most distinctive features of the Arctic Ocean is its sea ice cover, a broken and ridged veneer of frozen seawater that covers the deep basins even during summer. The entire Arctic Ocean, and adjacent areas such as the Canadian Archipelago and Bering Sea, are ice-covered in winter.

Thus, the north polar sea is an important, diverse, and unique part of the global ocean. Despite this, the exploration of this area has lagged behind that of other oceans because the ice cover effectively prevents navigation by surface ships. Even now there is no icebreaker powerful enough to travel freely through the Arctic Ocean. Its exploration had to await the coming of airplanes to travel above the ice and submarines to travel beneath it. The most effective expedition prior to the invention of these vehicles was that of Fridtjof Nansen, who froze a specially designed ship into the ice in 1893. For 3 years the ship drifted while scientific observations were collected. This early

effort was not followed by any successors for many years. Two expeditions between the two World Wars gave indications of the direction that research platforms would take in the Arctic Ocean after World War I. One was the abortive 1931 submarine expedition under the ice led by Sir Hubert Wilkins. The other was the first scientific research camp, North Pole I, to be maintained directly on sea ice itself. It was established in the U.S.S.R. in 1937 at the North Pole; from there it drifted into the East Greenland Current and out of the Arctic Ocean. These precursors were to be followed after World War II by nuclear submarine cruises under the ice and by aircraft landings on the ice in all parts of the ocean.

A broad expansion of all types of research in the Arctic Ocean took place between 1950 and 1970. Many drifting ice research stations were established by the United States and the Soviet Union, each enduring for a year or two, on the average, as a base for studies of atmosphere, ice, ocean waters, and crust beneath. The Soviets also made hundreds of aircraft landings on ice for spot measurements. U.S. nuclear submarines first traversed the Arctic Ocean in 1957, making possible continuous profiles of many geophysical parameters. Research efforts by the United States recently culminated in the Arctic Ice Dynamics Joint Experiment (AIDJEX) [2].

Importance of the Arctic Ocean

The naval military importance of the Arctic has grown significantly over the past few years. One cause has been the chain of events leading to the national policy of energy independence and the consequent emphasis on accelerated exploitation of our North Slope and offshore Alaska oil reserves. As a result, strategic planners have considered the vital role of our naval forces in the protection of a new and important sea lane. A second factor has been extension to extremely long ranges of submarine-launched ballistic missiles, making the Arctic Ocean a possible patrol and launch area. A third factor involves freedom of the seas and geopolitical intentions. Some countries bordering the Arctic have indicated, so far in a mild way, that the Arctic Ocean should be changed in status from international to inland waters following the so-called "sector principle," or that it should be demilitarized as the Antarctic. Very recently other nations have threatened to extend their rights over contiguous waters to the 200-mi (320 km) limit, which could in some measure bottle up the narrow eastern entrance to the Arctic Basin. The nuclear attack submarine, with its unique mobility in ice-covered waters, has important potential roles in all of the above considerations.

Of course, those well-demonstrated abilities of nuclear submarines to operate in the Arctic would not be possible without sonar for detection, navigation, and communication. The Navy recognized this in the late 1950s and began a long-term research program in arctic underwater acoustics, as well as studies of many other arctic environmental factors that affect naval operations in this unique area.

SEA ICE

In the global thermodynamic cycle of atmosphere and ocean, the polar regions are the heat sinks. In the course of a year they lose more heat to space than they receive from the sun. As a result, they are cold and maintain a permanent ice cover of annually varying extent. To compensate for the loss of heat to space, the general circulation imports heat into the polar regions from lower latitudes. In the present climatic regime, the vertical extent of sea ice is extremely small compared with its horizontal extent (about $1:10^6$). Therefore, small perturbations in the heat balance of the sea surface may cause large changes in the sea ice cover, resulting in changes of terrestrial albedo, sea surface temperature, ocean mixing, evaporation, and so forth.

Like snow, sea ice is an extremely perishable constituent of the earth's surface. Unlike any other terrestrial solid, it is kept in continuous rapid motion by winds and currents. The following discussion is an attempt to summarize some of the findings and problems of modern sea ice research.

External Driving Forces

In most regions covered with sea ice, drift and deformation of the ice are primarily due to the

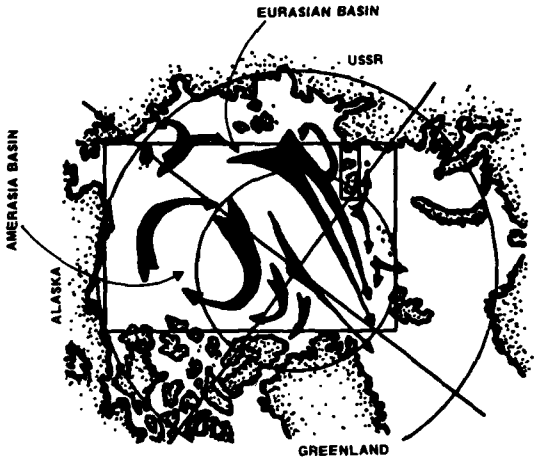


Figure 2—Drift patterns of arctic sea ice. Transit time from the Laptev Sea to the Greenland Sea is approximately 2-3 years. The Beaufort Sea Gyre requires about 10 years for one revolution.

tangential force exerted by the wind (exceptions are areas of swift and steady ocean currents—for example the Greenland-Spitsbergen Passage, shown in Figure 2, where the East Greenland Current exits from the Arctic Basin). Neglecting for the moment internal forces in the ice, which will be discussed later, the simple case of steady-state ice drift is that in which the velocity of the ice is such that the frictional forces between air and ice, and ice and water, are in balance. To analyze the actual balance of forces, one must add to this the Coriolis force due to the rotation of the earth, and a small component of gravity resulting from the slope of the ocean surface associated with currents.

In both fluid boundary layers, the frictional force, expressed as the vertical flux of momentum, depends on three main variables: the mean velocity, the intensity of turbulence, and the physical character (topography, roughness) of the solid surface (top and bottom of ice). To develop observational methods and theories relating, modeling, and predicting these variables has been one of the central subjects of geophysical fluid dynamics. Certain aspects of this problem peculiar to sea ice will be discussed below.

Stable stratification of the atmospheric boundary layer is the prevailing condition in the Arctic. It is caused primarily by radiational cooling of the ice surface and results in an "inverse" vertical temperature profile where, up to typically a few hundred meters above the ice, the temperature increases with height. In that case, turbulence is not "isotropic," meaning that a parcel of air displaced vertically by random motion is either heavier (going up) or lighter (going down) than the ambient air. Buoyant forces will tend to return that parcel of air to its original height. Stability of this kind consumes energy, taken from the work done by the overall field of atmospheric pressure (which drives the mean motion). The result is a reduction of the vertical flux of momentum, and a partial frictional decoupling between air and surface. In that case, basic precepts of isotropic turbulence, such as the linear increase of eddy viscosity with height and the linear increase of eddy viscosity with the mean wind, no longer apply.

Especially important in this context is the angle of turning between the (frictionless) geostrophic wind and the surface wind. According to classical Ekman theory, this angle is 45°. In a stratified boundary layer, the angle may vary from 10° to 40°, depending on the density gradient. This consideration applies to both the atmospheric and oceanic boundary layers.

The atmospheric boundary layer is most stable during the winter months, when net radiation is strongly negative. In the ocean, the boundary layer is most stable during the summer, when fresh meltwater from the ice surface is admixed to the uppermost layers of water and reduces its density.

Another boundary layer problem specific to sea ice is the great *local inhomogeneity* of the surface. Depending on locale and season, pack ice regions have a variety of surfaces: thick multiyear ice (2-5 *m*), *first-year ice* (0.2 *m*), pressure ridges, ploynyas, leads, meltwater ponds. During winter, the difference in surface temperature between thick ice and an open lead is 30°-40°C. The air overflowing a surface of such enormous heterogeneity is subjected to dramatic changes of its boundary layer over short distances. These changes are in themselves a subject of great scientific interest and can be studied most effectively in the Arctic. In addition, the extremely rapid heat loss from open leads is at times the controlling factor in the overall heat and ice balance.

Internal Ice Stress

The velocity of a given piece of pack ice is determined not only by the external forces acting at its location, but also by external forces acting elsewhere and being transmitted laterally through the ice. Natural pack ice is an assemblage of pieces, ranging from rubble to plates several kilometers in diameter. The description of the mechanical properties of such a complex material in mathematical terms is one of the core problems of sea ice research. Because of the great inhomogeneity of sea ice, the questions about its properties are intrinsically linked to the consideration of scales.

In the course of their shifting, rotating, rafting, and ridging motion, the individual ice "floes" most commonly break in tension (vertical loading). A large amount of data exist on the tensile strength of sea ice and its dependence on temperature and salinity. Because of experimental difficulties, for instance, the problem of accurately determining the porosity (air and brine) of a given sample, and because of differences in testing procedures (ring tests, beam tests, sample size, temperature control, etc.), the results scatter widely, especially for ice with a high brine volume. A definitive study on that subject, which has gained interest with the prospect of arctic transits by surface ships and the construction of offshore installations in regions of heavy pack ice, remains to be performed. Considering the mechanical properties of sea ice on a larger scale, where numerous flaws such as cracks, leads, and pressure ridges are contained in a single "sample," it is evident that a different physical reasoning must be applied, since an ensemble of ice floes separated by cracks is likely to have no tensile strength at all.

The search for a realistic constitutive law, relating stress and strain in sea ice on a scale of 100 km, has been the most important and productive pursuit in sea ice research during the last decade. In the face of having to model a material that is not a continuum and that has unknown mechanical properties, early investigators had to violate either physical intuition or facts, or both. On a basinwide scale, ice motions appeared smooth enough to suggest continuous behavior. Using classical concepts of fluid dynamics, models assuming a viscous ice cover driven by mean monthly or annual wind fields yielded acceptable mean velocity fields, but the same is true for a model that treats the ice as an incompressible material [3]. Both plastic and viscous constitutive laws can be formulated to contain nonlinearities and anisotropies to allow for certain known or assumed types of mechanical behavior of the ice (for instance, strain hardening). One of the most important advances achieved by the AIDJEX ice model (see below) has been that it connects the clearly discontinuous, plastic process of pressure-ridge formation to a large-scale, elastic-plastic constitutive law.

Since it is unlikely that a single description of the mechanics of sea ice can be totally satisfactory on all scales of space, future studies will doubtless adopt a pragmatic approach in which the form and content of constitutive laws will be selected according to the type and purpose of the model calculations. In the extreme case of predominantly thin ice, it may well be possible to neglect the internal ice stress altogether.

Arctic Ice Dynamics Joint Experiment (AIDJEX)

The twofold purpose of AIDJEX was

1. to acquire, during the period of a full year, an optimal set of data for studying basic processes, and for both "driving" and testing a large-scale dynamic model of sea ice,

2. to improve existing models, with special attention to finding a physically realistic representation of external and internal stress and suitable formulations of the conservation of mass and energy (which were neglected in earlier models).

Results from all phases of the project are described in the AIDJEX Bulletins (No. 1, Sept. 1970, to No. 32, June 1976). All data are stored in the AIDJEX Data Bank and are available to anyone.

An example of the data obtained is shown in Figure 3. In the course of one year (June 1975 to April 1976) the polygon delineated by automatic data buoys is both deformed and compressed. The general direction of ice drift was unusual (compare Figure 2) and caused the extreme difficulties encountered by the "bargelift" to locations on the Alaskan North Slope in the autumn of 1975.

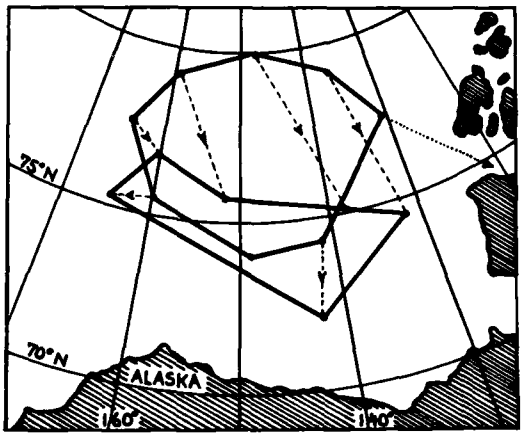
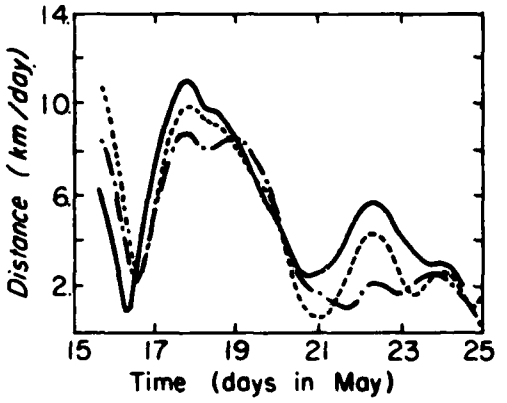


Figure 3—Displacements of the outermost ring of AIDJEX automatic data buoys from June 1975 to April 1976. Positions are determined by automatically received and telemetered NAVSAT signals and by RAMS (via Nimbus F). Buoy positions are observed several times per day, along with readings of barometric surface pressure and air temperature. One buoy drifted into McClure Strait (dotted line) and was lost there. In addition to the buoy array shown above, 23 buoys were deployed in late 1975 and early 1976, most of them within 200 km of the Beaufort Sea coast [13].



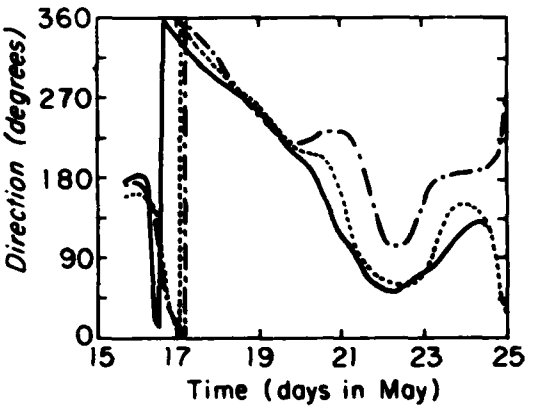


Figure 4—Comparison between 1975 AIDJEX field observations and model calculations. The solid lines represent speed and direction of the four manned camps, spatially averaged and filtered to remove frequencies greater than 1 cycle per day. The dot-dashed and dashed lines represent model calculations. The better fit of the dashed line was achieved by approximately doubling the drag coefficient in both the atmospheric and oceanic boundary layers [5].

Figure 4 shows an example of the numerical tests with the AIDJEX ice model [4, 5]. Additional, independent comparisons between computed and observed ice displacements can be made by means of successive LANDSAT pictures (Figure 5).

An important feature of the AIDJEX ice model is that it relates both the events of mechanical deformation and the heat balance (ablation and accretion) to one key parameter, the ice thickness distribution [6]. During the cold season, the heat loss from exposed sea surface is extremely rapid. Under these circumstances as little as 2% of open-water surface dominates the heat balance of an entire region. From the sparse information available, it appears that during winter the area of open water is smaller than 2% and that the thickness category 20-80 cm, which covers a greater area, determines whether the heat balance of a larger region is positive or negative [7].

It can be expected that the AIDJEX data and model calculation will be extremely valuable in selecting future arctic observing systems. Both the Pilot Study of 1972 and the Main Experiment of 1975/1976 proved that, in the spectrum of ice motion, high-frequency events (1 hr or less) are of only local importance. The significant features of the stress and strain field are covered by an observing system with grid spacing of 100-300 km in space and one-half day in time. Unfortunately, observations of the ice thickness distribution require a resolution of 10-100 m, which at present can be achieved only by airborne or submarine-borne sensors.

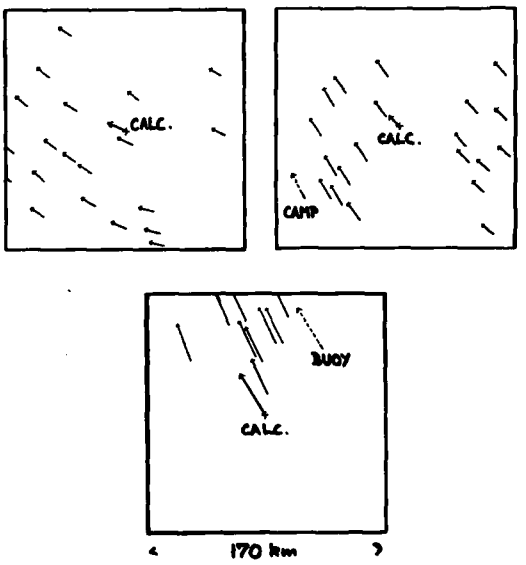


Figure 5—Comparison of 24-h ice displacement obtained from successive LANDSAT images (three pairs, May 17-18, 1975), from calculations with the AIDJEX model (vectors marked "X"), and from the displacement of one buoy and one camp that happened to lie inside the LANDSAT frames (dashed vectors) [4].

Ice Forecasting

Considering scientific and practical applications, it appears useful to distinguish between two kinds of ice forecasts:

1. In both polar regions the extent of sea ice undergoes a seasonal variation. (This is particularly large in the Southern Ocean.) If we assume that climatological data describing the mean annual cycle of the dynamic and thermodynamic forcing functions are available, then the computation of the seasonal variations of the sea ice cover becomes a kind of "ice forecast." If such computations could be performed, one might assume certain variations of the forcing functions (for instance, a different rate of solar energy output) to study the resulting changes in the sea ice cover with all its implications for global climate [8, 9].

At present, no realistic models exist that describe the annual variation of sea ice. The explanation (or "prediction") of an experiment that, with some variation, nature performs every year would be an important step toward understanding the role of sea ice in global climate. This problem has been assigned the highest priority in the U.S. contribution to the Polar Subprogram of the Global Atmospheric Research Program [10].

The ability to forecast regional ice conditions for a future month or season would obviously be of great operational and economic benefit. A number of schemes, based on extensive empirical studies, have been elaborated, primarily by Soviet authors. However, the skill of these forecasts is low, and the underlying physical mechanisms are not well understood.

Research on the physical basis of climate, methods of prediction, and limits of predictability have become an issue of worldwide concern [11].

2. The second type of forecast is one in which local ice velocity and concentration are predicted, generally for some operational purpose. Since the wind is the primary force driving the ice, the most important ingredient for that type of forecast is a prediction of atmospheric surface pressure and, hence, wind. It is a fortunate coincidence that, among all parameters making up a "weather" forecast, dynamic models of the atmosphere predict atmospheric pressure with the greatest precision. Numerous studies conducted in connection with planning the Global Atmospheric Research Project indicate that errors in determining the initial state of the atmosphere, and simplifications introduced by the models, limit the range of useful deterministic weather forecasting to about 2 weeks. Given the problems involved in deriving surface wind stress from a field of barometric surface pressure, one must expect that the limit of useful forecasts of ice motion may be considerably less than 2 weeks.

In addition to the dynamic influence of air and ocean currents, sea ice is affected by the local heat balance. The most important forecast to be made in that context would be the dates of the first and last presence of ice in a given location (freeze-up and break-up). These events depend on a combination of seasonal conditions (for instance, the amount of ice grown during one winter), and short-term events (for instance, a storm that coincides with high tide to shorefast ice). Shallow water and the proximity of a shoreline introduce a variety of complications. They are at present

under intensive study in connection with the development of natural resources on the arctic shelf.

Observing Systems

It was the intent and hope of the planners of AIDJEX that their project would introduce a pause in the need for maintaining multiple, long-term, manned ice stations, giving way to a different logistical approach (for instance, the icebreaker of the Nansen Drift Station [12]) and the use of unattended and remote-sensing devices.

The observational requirements of AIDJEX motivated a rapid development of sea ice data buoys, described in the final section of this review. As a result, AIDJEX and its corollary field programs have been the largest user of the Random Access Measuring System (RAMS) on Nimbus F in terms of buoy-years to date [13]. At the same time, a considerable effort by many agencies is underway to exploit satellite-borne remote-sensing methods for use in sea ice monitoring and research.

Sensors of electromagnetic radiation are available for a wide spectrum ranging from visible light (conventional photography and television) to waves many centimeters in length (both passive and active). The power of resolution of a sensor viewing the Earth from space generally decreases with increasing wavelength, while its power to look at the earth's surface through clouds and water vapor increases.

An example of the use of high-resolution images of sea ice in visible light was given earlier (Figure 5). Among the numerous other remote-sensing devices useful in sea ice research [14], the Electronically Scanned Microwave Radiometer (ESMR) and Scanning Multichannel Microwave Radiometer (SMMR) are of particular interest. They receive radiation emitted by the earth's surface at wavelengths of 1-6 cm, which passes the atmosphere almost unattenuated, making these systems independent of the weather. It was found that the emissivity and hence the apparent brightness temperature of sea ice depends more on its age than on its actual thermometric temperature. It was established that multiyear ice appears to be some 20K colder than first-year ice, while their actual surface temperatures may differ by only a few degrees. With the improving resolution of scanning microwave radiometers and the improving insight into the factors controlling the emissivity of sea ice, these radiometers should become increasingly useful in monitoring not only seasonal changes of the ice boundaries (open water appears extremely cold) but also the large-scale deformation and ice growth features in the interior of sea ice regions covered by sea ice.

Measurements indicate that brine volume (functionally related to temperature and salinity) is the most important factor determining microwave emissivity. Even though it has long been known that sea ice, in the course of its growth and aging, loses much of its initially high salt content, a thorough experimental study of the mechanisms of natural desalination [15] relating ice salinity to growth and temperature history remains to be performed. Such a study would be particularly useful in improving the interpretation of passive microwave images of sea ice.

The resolution of 100-300 km in space and one-half day in time mentioned earlier does not suffice to follow certain inertial and tidal effects. Although their "power" in the overall spectrum of motions is small, they may generate periodic phenomena (in space and time) whose significance can only be assessed when their physical nature is more clearly understood. As a result of the rising economic importance of the Arctic and of the recognition of cryospheric processes as a major component of the global climate system [16], the number of scientists in the United States engaged in sea ice research has been increasing during the past decade. If an adequate balance can be struck between "big science" programs, such as the proposed Nansen Drift Station Project and POLEX, and a number of specialized research activities by individual principal investigators, then there is little doubt that adequate progress in basic sea ice research can be achieved.

The need for environmental monitoring of the Arctic, and the high cost and operational hazards inherent in the customary sea ice camps, are making the use of automatic observing systems increasingly attractive, in terms of both efficiency and expense. An important task for the near future will be the selection and deployment of a long-term Arctic Ocean monitoring system that provides data for both operational use and scientific research.

OCEANOGRAPHY/GEOLOGY, GEOPHYSICS

Physical Oceanography

The Arctic Ocean is a mediterranean sea with straits connecting it to both the Atlantic and Pacific Oceans. The major influence on its water masses comes from the Atlantic via the Greenland and Norwegian Seas. The waters at depths greater than 200 m are of Atlantic origin. The Pacific influence is observed only in the layer between 50 and 200 m, which occurs in the Canadian and Alaskan side of this ocean. Besides the advetic influence through connection with other oceans, there is the influence of freshwater discharge from the many rivers that empty into the Arctic Ocean, lowering the salinity of the surface layer, which extends down to 50 m.

A unique feature of the Arctic Ocean is the presence of a frozen ice cover. This ice cover is not solid like that of a lake but fractured and ridged by constant movement under the influence of wind and current. Sea ice varies greatly in seasonal extent, covering the entire Arctic Ocean and adjacent ocean areas during winter but shrinking during summer to cover only 60% of the ocean. On the geological time scale even larger changes in area take place. It has been shown that polar waters (temperatures less than 0.5°C) invaded the North Atlantic during the Wisconsin ice advance. This implies a much greater ice cover on the oceans during that period than now.

There are strong differences of opinion on the role of the Arctic Ocean in global climate changes. According to some theories the ice-ocean-atmosphere system is inherently bistable and capable of switching from ice-covered to ice-free conditions with only a small triggering influence. Such a switch would undoubtedly produce profound changes in the climate of the Northern Hemisphere. Others have maintained that the present icepack is stable, so that even if it were removed by some means, natural or artificial, it would return to its original state. The uncertainties in our knowledge of fluid dynamics on a global scale do not yet allow a choice between these divergent theories [9].

The mean circulation of the ice has been charted from the drift of manned ice stations and unmanned buoys. A major transpolar drift stream crosses the North Pole and exits through the passage between Greenland and Spitsbergen. (See Figure 2.) There is a gyre in the Canadian-Alaskan side, which circulates in clockwise rotation and smoothly joins the transpolar drift stream [17].

The waters in the mixed layer, generally extending from the surface to a depth of 25 to 50 m, are frictionally coupled to the ice. The vertically integrated currents in this layer tend to move with the ice. The circulation of the surface water masses follows the ice motion and decreases with depth. The Pacific Water entering through the Bering Strait spreads northward into the Amerasian Basin. The spreading seems to be due to the eddies present in this 50- to 200-m layer rather than to a steady circulation. The Atlantic Water entering through the Greenland-Spitsbergen Passage follows the continental shelf along the Eurasian continental margin and spreads from there into the Canada Basin, where the exact circulation pattern is not so clear. Arctic Deep Water below 900 ms is of Atlantic origin and presumably is formed only occasionally, spilling over into the deep basins of the Arctic. There is clear indication of its progress from the Eurasian to Amerasian Basin, which it enters by flowing over the sill on the Lomonosov Ridge. These circulation patterns are deduced primarily from the distribution of temperature and salinity.

The unique conditions in the Arctic present an opportunity for fundamental oceanography and meteorological experiments. It was in the Arctic Ocean that Nansen, in his expedition on the *Fram* 1893-1896), first observed that ice drifts to the right of the wind direction. These observations stimulated Ekman's theory of boundary layers in which both friction and the earth's rotation are important. The theory is still one of the cornerstones of oceanography. Internal waves were also first observed by Nansen on the same expedition. More recently, detailed observations of turbulence, microstructure, and eddy motions have all been made possible by the ice platform from which instruments may be suspended without the interference of wave action. It seems reasonable to expect that future observations in the Arctic Ocean will provide further insight into basic oceanographic processes.

Measurements from ice platforms in recent years have confirmed the existence of a spiral current structure in the upper layers, predicted long ago by V. Ekman. The Arctic Ice Dynamics Joint Experiment (AIDJEX) has produced some especially good data on the spirals, as well as on other features of the oceanic boundary layer beneath drifting ice floes [18]. The stress below the ice has been measured by several techniques and is used to help establish the balance of forces acting on an ice floe drifting under the stress of wind.

Transient undercurrents, attaining speeds of 40 cm/s at a depth of 150 m, were noted on certain occasions. Although similar motions apparently have been observed a few times before in the Arctic Ocean, they were not noted in the 1970 or 1971 AIDJEX programs. The 1972 work clearly showed them to be subsurface eddies. Eddy diameters of 10 to 20 km were found in the depth range of 50 to 300 m [19, 20].

The arctic eddies contrast with those in other oceans, which generally have a larger diameter and a surface rather than subsurface maximum in horizontal velocity. The differing properties of the arctic eddies may be associated with the ice cover and with the steeper density gradient there. If so, the Arctic Ocean provides an opportunity on a geophysical scale to study eddies under altered conditions. The origin of these eddies and their part in the exchanges of momentum, heat, and salt are not known. It may be that they are formed in the oceanic front north of Alaska, which separates the more saline water entering from the Pacific via the Bering Strait from the less saline surface water of the Arctic Ocean. If this is the case, the eddies must play an important role in the transfer of properties between polar and temperate oceans in the Northern Hemisphere.

A knowledge of the exchange of heat, water, and salt among the Arctic Ocean, the atmosphere, and other oceans is a fundamental first step in understanding. However, the budget for these parameters in the arctic and subarctic seas is still not well known. Beyond the elementary need for budget information is the need for quantitative data on the processes involved. The exchange of energy and properties undoubtedly takes place by turbulent mechanism on many scales. Recent work in the Arctic Ocean has revealed the presence of such features as mesoscale eddies and step structure, which must play a role in horizontal and vertical mixing there, as they do in other oceans. There is opportunity in the Arctic Ocean to study these features in a parameter range different from other oceans and from a stable ice platform that permits detailed study of their structure.

One of the primary problems of physical oceanography in the Arctic Ocean is better knowledge of the processes that control sea ice extent. One of the processes is the heat balance, including both vertical flux and horizontal advection. Important factors influencing this are ocean current systems, mixing processes in the upper layers, and the fluxes of salt, which affect stratification.

Another important problem is the circulation of water and ice on the continental shelves. These areas are of importance to some aspects of deep sea circulation (for example, the role of submarine canyons in mixing). Increased exploitation of arctic resources such as oil and the attendant transportation and possible spillage questions make the study of these areas important.

The Earth Beneath the Arctic Ocean

Nansen first showed that the basin around the North Pole reached truly oceanic depths. It remained for expeditions of recent years to show that it is not a single basin but rather four basins separated by three nearly parallel ridge systems, which join the North American and Eurasian continents (Figure 6). How did this ocean originate and what produced its complex shape? An extensive bibliography on arctic geophysics may be found in the "Proposed Scientific Plan for the Nansen Drift Station Project" [12].

Geological science has been revitalized in the last decade by the concept that the Earth's crust is divided into relatively quiescent plates separated by narrow zones of concentrated seismic and volcanic activity. The insights of plate tectonics have provided a framework for reconciling many previously unrelated observations of crustal composition and structure.

One of the focal points of plate tectonic research is the delineation of plate boundaries on a

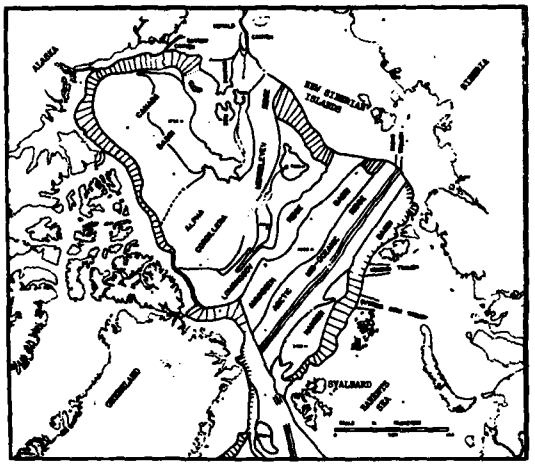


Figure 6—Physiographic provinces of the Arctic Ocean [21]

global basis. The plate boundary crossing this ocean, the Arctic Mid-Oceanic Ridge (sometimes also called the Nansen or Gakkel Ridge), is one of the most linear segments of this global ridge system. For almost 2000 km the line of seismic epicenters marking the center of the ridge forms an almost straight line. There are also other unusual characteristics of this ridge. It intersects the continental margin of the Laptev Sea, one of the shallow shelf seas north of Siberia, in one of the few cases of this type of behavior for midoceanic ridges. The pole of rotation about which this ridge opens is located relatively close by in the Eurasian continent. This leads to a low rate of spreading. The greatest depths found in the Arctic Ocean are located in the rift valley that marks the active center, where spreading takes place.

The Lomonosov is the central of the three ridges. Its characteristic smooth profile and its shape, which would seem to fit back into the Eurasian continental margin, suggest that this ridge is the former continental margin which was split away and carried to its present location by sea floor spreading. The symmetrical location of the Arctic Mid-Oceanic Ridge between the margin and the Lomonosov Ridge supports this idea.

The third and broadest of the three, the Alpha Ridge, is archlike in cross section and topographically rough. This feature is sometimes divided into two ridges, the Alpha Cordillera and the Mendeleyev Ridge. Like the Lomonosov, it is not seismically active. The origin of this ridge is least understood of all. One suggestion is that it is a former center of sea floor spreading that is no longer active. Alternatively, its genesis may be related to subduction or possibly to compression of an earlier ocean floor.

Most of the floor of the Arctic Ocean is covered with unconsolidated sediments, ranging in grain size from clay to pebbles and even boulders, which have been carried out from shore by ice. The greater part of the material is of glacial origin. There is a smaller organic fraction consisting of the skeletons of marine organisms. Between the ridges lie basins that have their deepest parts filled with sediments. The surface of these sediment deposits form the remarkably flat abyssal plains. Sediments have been carried into these basins by turbidity currents (submarine flows of sediment and water). Sediment depth reaches several kilometers in several of these basins. For example, stratified sediments reach a depth 3½ km below the Wrangel Abyssal Plain and 2 km beneath the Canadian Abyssal Plain. Turbidity currents occur infrequently, and the sediment fills are deposited irregularly in time.

On the ridges, however, sediments are laid down particle by particle in a rain of material from higher in the water column. In these places the sediments form a nearly continuous sequence in time, with individual layers varying in composition and thickness according to oceanic conditions at the time. Some of the conditions governing the layering are plant and animal life, and hence the state of the ice cover, which influences light penetration into the water. The presence of an ice cover in the past is of importance in consideration of ice ages. There are suggestions that the sea ice cover is a significant factor in triggering ice ages. Relations between sea ice and continental ice sheets are provided by the sedimentary record in the cores. So far, more than 500 sediment cores in the 2- to 5-m length range have been obtained along the route of ice island T-3 in the Canadian Basin and on the Alpha Ridge. Soviet workers have taken hundreds more in this and other regions of the Arctic Ocean.

Cores from the Canada Basin indicate that the present ice cover of pack ice has existed continuously through at least the latter part of the glacial

period [22]. The oldest dates of present cores differ but are at least 1 m.y.b.p. and possibly as early as 3.5 m.y.b.p. These are only minimum dates for the existence of arctic sea ice. So far the longest cores have been 4 to 5 m in length and have not reached deep enough to penetrate to layers formed before the glacial period. Longer cores, in the 10- to 15-m range, are needed; they will produce a complete climatic record of glacial age conditions in the Arctic Ocean. A large number of such long cores needs to be collected from various parts of the Arctic Basin. Any single core may have lost sections of its record by slumping or other local events. Only correlation between a group of cores can produce confidence that the complete climative sequence has been obtained.

In comparison with those of other oceans, the tectonic features of the Arctic Ocean are little known. The broad outlines of the major ridges have been charted, but the details of their rough surfaces, which contain clues to their origin, are not known. The greatest need at present is for more field data to help unravel the genesis and development of the topography and structure of this ocean. The Eurasian Basin as presently known seems to fit into the global scheme of plate tectonics with only minor discrepancies [23]. Some of the unusual features of this basin are the low amplitude of magnetic anomalies. This may be caused by the exceptionally deep sediment layer that has accumulated in a basin of such limited extent. Seismic studies are needed to decide this question. Another unique feature is the linearity of the Arctic Mid-Oceanic Ridge, which extends for 2000 km in a straight line. Soviet investigators claim to have found many small transform faults that break this straightness with short offsets. Detailed bathymetric and microseismic studies would be needed to confirm this. Also, although it is generally agreed that the Eurasian Basin opened in the period since 80 m.y.b.p., there are questions about the sequence of events during opening. Deep sea drilling would help answer questions about development here as well as in other parts of the Arctic Ocean. The U.S. Deep Sea Drilling Project has not ventured into this ocean so far.

The Alpha Ridge and Canada Basin have been explored geophysically on a reconnaissance scale from drifting ice stations, and nuclear submarines have obtained bathymetric profiles of the region [24]. The origin and development of this area is more obscure than that of the Eurasian Basin, since the pattern of bathymetry and structure here does not fit as neatly into the plate tectonic theory. Lack of detailed surveys has led to various speculations as to origins.

The earliest hypothesis was that subsidence in the Canada Basin and Alpha Ridge had resulted in a foundered continent. This idea requires the presence of some unknown process to convert the former continental crust into the oceanic crust that has been observed on the few available seismic refraction profiles. Certainly more seismic studies are called for to describe the crust in this region.

Another genetic hypothesis is that the Alpha Ridge is a former midoceanic ridge that was actively spreading up until 50 million years ago, when it became dormant. There is some indication of a rift valley along the crest of the Alpha Ridge, as well as magnetic spreading anomalies and transform faults. One variation of this hypothesis traces the former midoceanic ridge through the Labrador Sea and suggests that the Alpha Ridge became inert when the spreading axis suddenly switched from the west side of Greenland to its present position on the east side. There are also difficulties with the Alpha Ridge as a fossil midoceanic ridge. This ridge is much deeper than one would expect if the hypothesis is correct, and the amplitude of magnetic anomalies is greater than normally expected.

Still a third suggestion is that this ridge is related to subduction or at least to compression of an earlier ocean floor. Here again, geophysical studies and deep sea drilling would help decide between the proposed origins. Under the seafloor spreading hypothesis, the Canada Basin is an ancient section of ocean floor, older than 130 m.y.b.p. and perhaps older than 340 m.y.b.p., making it one of the world's oldest pieces of sea floor.

The major problem of Arctic Ocean tectonics centers around the history of plate motions in the Arctic Ocean and how they led to its present shape and structure. New data are needed here more than new theories, just as they are for the sedimentary history. These data will come from new geophysical surveys which include

bathymetry, gravity, magnetics, and seismic studies. The choice of vehicle is of prime importance to any survey in polar regions. The magnetic surveys and perhaps the gravity surveys can be carried out by airplane. Parts of the Arctic Mid-Oceanic Ridge and the Alpha Ridge have already been flown in U.S. aeromagnetic surveys. Bathymetry is undoubtedly best surveyed by nuclear submarine, although unmanned submersibles may be increasingly helpful here. Seismic reflection and refraction can probably be done best with helicopters or fixed-wing aircraft operating from temporary base stations on the ice. As much use as possible should be made of unmanned instrumental buoys.

Geophysical studies in the Eurasian Basin are of special interest, since no U.S. data are available from there. Only generalized results from Soviet sources describe the area. An ice station as a base of operations in this basin is not feasible, because the long airplane distances and change of breakup reduce safety too much. An icebreaker frozen into the ice makes a base that is safe from breakup. This concept is now under active consideration as the Fridtjof Nansen Drift Station.

ARCTIC UNDERWATER ACOUSTICS

The earth's environment is observable and measurable only by reception of radiant and reflected energy. In the water medium that makes up most of the Earth's surface, sound is by far the most useful, since it is the only form of energy that propagates efficiently. To understand, and through this understanding to utilize, this energy is the raison d'être of the science of underwater acoustics and its technological adjunct, sonar engineering. Underice acoustics is a branch of underwater acoustics.

The Uniqueness

The ocean environment affects the behavior of underwater sound energy and both limits and enhances the usefulness of sonars in many ways. The unique environmental feature of the Arctic—the one that effectively precludes extrapolation of generalized acoustics theory, models, and data from the more thoroughly researched open ocean areas—is the ice canopy. Its presence grossly affects the two parameters, sound propagation and noise, that are of prime importance to the ultimate users of acoustics knowledge, the sonar designers and operators.

Consider some of the effects of ice and the resulting uniqueness of the Arctic Ocean. The high albedo of the ice cover prevents warming of the upper layers, causing a stable, nearly isothermal vertical temperature structure. This results in a positive gradient of sound velocity and upward refraction of propagating sound energy, forming a natural waveguide bounded by the surface. Sound rays reflect from the surface, refract and reflect from the surface again and again as they propagate. Figure 7 shows a typical arctic vertical sound velocity structure and samples of rays. The surface-bounded waveguide may be thought of as a variation of the more familiar "Deep Sound Channel" of the open oceans, but of course in the Arctic the sound channel axis is not "deep"; it is at the surface. Just as in the open ocean, a sound source near the axis transmits energy with great efficiency in the Arctic. This description is oversimplified and warrants a few qualifications. For efficient horizontal propagation, neither the source nor receiving hydrophone can be close to the water-ice-air interface, which forms a "pressure-release" surface. On the other hand they cannot be too deep either, or they will not "couple" to the wave guide. This dependence on nearness to the acoustic pressure-release surface is a function of wavelength, and therefore frequency, of the sound. The ice introduces another frequency-dependent consideration; as the rays strike the rough ice bottom the energy is only partially reflected, the rest being back-scattered and absorbed in the ice. To the very low frequencies, the ice is not "rough," but rather a near-specular reflector causing very little bounce loss. Progressively higher losses are suffered as the frequency increases. Therefore, the bottomside ice topography is of considerable importance to sound propagation.

Ice is by far the most important source of background sonic noise in the Arctic. The mobility of the central arctic ice pack, driven primarily by surface winds but also by water currents, causes relative motions between ice masses,

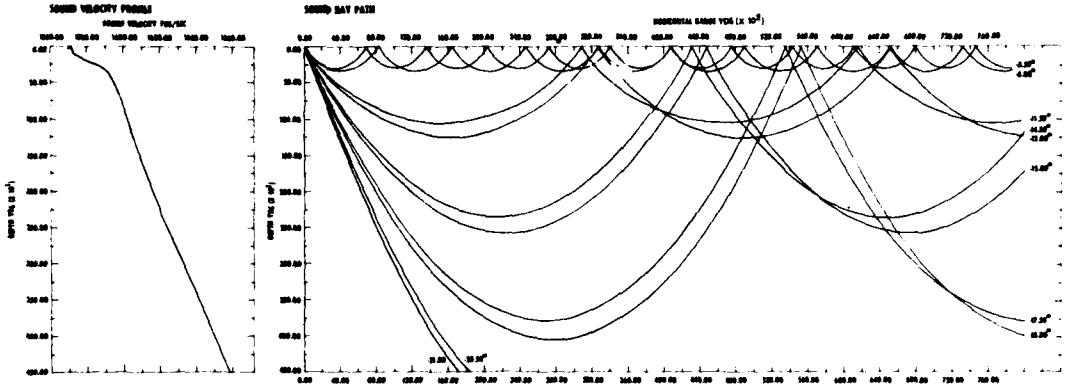


Figure 7—Typical arctic deep water sound-velocity profile and resulting acoustic rays.

which in turn produce stresses at the boundaries and consequent pressure ridging. A considerable amount of transient acoustic noise results. In areas where, because of adjacent land masses, the ice is locked in and immobile, another ice phenomenon is important in the production of noise. Air temperature changes cause the top of the ice to expand or contract while the ice bottom, being in water of constant temperature, does not. The result is thermal cracking, which is the main source of noise. Also, windblown snow and ice crystals on the ice surface cause noise in those areas.

Even biological noise in the Arctic is related to the presence of the ice. Whales, seals, and walruses can produce a cacophony of noise in certain frequency bands. The ice allows these animal populations, at least the latter two, to be widely dispersed and present during all seasons. Without the ice they would concentrate only in very shallow areas and therefore be of no particular bother to sonars. Although some sounds are heard from marine mammals almost all of the time, the calls, whistles, clicks, and moans are loudest in the spring, like the songs of birds in the forest—and probably for the same reason.

Reverberation is an acoustic phenomenon important to active sonar operation, and here again the ice cover exerts a strong influence. In open oceans, gravity waves churn up a considerable amount of trapped air, thereby increasing volume reverberation in the upper layers, but there are no waves in ice-covered waters. Plankton layers of various types are also important to reverberation, and their distribution and vertical migration habits are influenced by the low levels of light penetration through ice cover. Of course, surface reflection reverberation is totally dependent upon ice bottom topography and structure and quite different from that produced by gravity waves.

Of practical importance to sonars is the problem of identifying targets as belonging to a particular class of vessels. The open oceans of the world contain a multitude of surface ships, all producing noise that obfuscates the sonar picture. No so in the Arctic, which is primarily a mediterranean sea. Futhermore, the ice cover precludes free and easy navigation, limiting surface vessels primarily to the shallow water rim and then only in summer. These features cause a drastically different background sound spectrum at the very low frequencies.

Except for the relatively simple single-hydrophone telemetry systems ("sonobuoys"), all sonars use multisensor arrays to gain directional information and discriminate against noise, thereby improving the signal-to-noise ratio. To obtain bearing resolution, an array must have an aperture that is reasonably large in terms of signal wavelengths, and its geometry must be known precisely. For efficiency it must be located at an optimum depth. Very large array dimensions are required at very low sonic frequencies; these are practically attainable in the Arctic because of the

stable ice cover and the proximity of the sound channel axis to the surface. Thus, advanced systems can be explored and used in acoustics research in the Arctic at but a small fraction of the cost and effort that would be required in the deep open ocean.

The influence of ice on acoustics and oceanography in those peripheral arctic areas where the pack meets open water (the Marginal Sea Ice Zones) warrants special mention. Such areas are of importance to modern submarine operations. These zones have been described by submarine captains who have cruised in them as having the "worst sonar conditions in the world," "worse than the edge of the Gulf Stream." Such conditions are caused by anomalous oceanographic conditions of great spatial variability, which might be expected in a zone interfacing the disparate water masses of the open and ice-covered areas. By definition, the ice in these zones varies from zero to continuous cover with highly mobile and variable concentrations between those extremes.

It is not meant to imply that ice is all-important to arctic acoustics; it is not, of course. The same factors and phenomena that complicate the acoustics picture in open-ocean research are present in the Arctic. Internal waves, oceanographic fronts and eddies, bottom and subbottom reflective qualities, etc., all exert their influence on sound propagation, albeit to different degrees and extent, just as they do in different open-ocean areas. It is the ice canopy that makes the Arctic unique and demanding of separate scientific exploration and study. Because the ice is at the interface of the atmosphere and the water and reacts to both to influence underwater acoustics in the manner described, meteorology and ice dynamics are of direct interest to acoustics researchers, as much so as oceanography per se.

The Past

The International Geophysical Year saw the birth of the Navy's research effort in arctic marine science. Impetus was provided by the first submarine transit of that ocean by U.S.S. *Nautilus* in 1958. Despite the success and operational vistas opened by that operation, marine science was not the primary thrust of the Navy's arctic research program in those early years. The emphasis was on "Man in the Environment," and therefore most of ONR's arctic effort was centered on biological sciences. However, significant though preliminary work carried out in basic oceanography, geophysics, and meteorology had application to underwater acoustics, and some preliminary work was done in sampling acoustics propagation and noise in the late 1950's and early 1960's.

Although the effort was limited, primarily by the lack of good support facilities for on-ice work and by the attendant high costs, the inherent character and dissimilitude of arctic acoustics were discerned and needs for further research were indicated. Those needs were primarily for basic acoustics survey data, for it was apparent even from these early investigations that the arctic acoustic environment varied greatly both spatially and temporarily. Central arctic deep water, arctic shallow water, the locked-in ice of the Canadian Archipelago, and the Marginal Ice Zones all differed in character and magnitude in important ways. These basic survey data were necessary before meaningful predictive models could be derived, although early models based on ray acoustics and wave acoustics were of considerable help in understanding the nature of propagation phenomena. Therefore, in the 1960's emphasis was placed on gathering these needed data. Unfortunately the difficulties and high cost of placing and maintaining men on the ice to do this job extremely limited this effort. It was not until 1975 that new technological advances in remote instrumentation using radio telemetry opened new vistas in cost-effective acoustic data collection in the Arctic.

Most of the acoustics work during the 1960's was centered at the only available U.S. ice stations, and these were ice islands. Ice islands are floating fragments of thick and massive glacial ice and provide station longevity, and important consideration to the logistics budget. However, they are in some ways undesirable as acoustics platforms, primarily because they are rare and do not have the same character as the prevalent pack ice. Moreover, the ice islands housed diverse scientific experiments, many of which produced noise interferences to acoustics projects. However, ice islands were the only available capability and had to be used. In the spring of 1970 a fortunate oppor-

tunity to use two dedicated, quiet floe stations, ARLIS 5 and ARLIS 6, became available for concentrated acoustics experimentation. A considerable amount of basic propagation and ambient noise data resulted from this effort.

In 1971 a special effort, under the aegis of the Arctic Submarine Laboratory of the Naval Undersea Center and with logistics support of ONR, started in earnest to study the acoustics and oceanography of the Marginal Sea Ice Zones on both the Pacific and Atlantic sides of the Arctic. This is a long-term effort involving support submarines, icebreakers, fixed-wing aircraft, rotorcraft, and short-term manned ice camps.

The primary investigators in Arctic underwater acoustics from 1958 to the present, and their most important publications in that field, are given in the short Arctic Acoustics Bibliography at the end of this paper.

The Future

The arctic acoustics program has progressed at a slow pace, with periods of activity centered on infrequent and short in duration submarine cruises and on the availability of equipped manned ice stations or icebreakers. Because of costs and the lack of suitable advance land bases and aircraft, basic propagation and noise data collection efforts are still in a rudimentary stage and concentrated in the south Beaufort Sea—the only region within reach of the single U.S. arctic logistics support base at Barrow. This is all about to change, and the prospects are exciting to arctic scientific investigators who have labored so long under extremely adverse field conditions to scratch out the needed data base. The change is being brought about by the development of new technologies for remote sensing from aircraft and satellites and remote instrumentation telemetering data through satellites and direct ice-to-land sites by high-frequency radio.

Soviet scientists have used remote unmanned telemetry stations extensively since the early 1950's. Their platform is the DARMS (Drifting Automatic Radio Meteorological Station) which uses the medium-frequency radio band for telemetering data of all types to shore stations. DARMS are used as operational weather stations in direct support of the Northern Sea Route and also as scientific data collection stations over the entire Arctic Ocean. It was not until 1975 that the United States used "data buoys" to any extent for either operational or scientific uses in the Arctic, and that was for the Arctic Ice Dynamics Joint Experiment (AIDJEX). Several types of remote stations were developed and used successfully at that time, primarily for the collection of barometric-pressure, air-temperature, ice-strain, and acoustics ambient-noise data and, less extensively, for water-current and temperature data.

The AIDJEX data buoys included 10 Arctic Environmental Buoys (AEB) (Figure 8), which were large, sophisticated platforms using high-frequency radio telemetry to a central control station on the ice. Each buoy included two precision barometers, two air-temperature sensors, and a NAVSAT (the Navy's Transit Satellite) receiver as its primary sensor suite. It also had relocation aides, engineering data sensors, and two digital memories to hold data during radio blackouts. Also used were 10 SYNRAMS (Synoptic Random Access Measurement System) buoys that relayed data through the NIMBUS 6 satellite (Figure 9). Those buoys measured location, barometric pressure, air temperature, and ambient noise levels in four one-third-octave bands. These AEB and SYNRAMS buoys made all of their measurements at the "synoptic weather" observation times every 3 h for the yearlong experiment. Four of a third type of data buoy equipped with a barometer, water-current meters and water-temperature sensors also used NIMBUS 6 for location and data telemetry. In the middle of the AIDJEX field experiment yet another NIMBUS 6 data buoy was developed and readied by December 1975 for ice strain measurement use. This was the ADRAMS (Air Droppable Random Access Measurement System). Where the other data buoys required aircraft landings on the ice for installation, ADRAMS was designed to be air-dropped in any weather, from any altitude and during day or night. Sixteen of these 8-month-life buoys were dropped and used successfully to track ice movement. Two were equipped with precision barometers for automatic atmospheric-pressure measurements.

While the on-ice buoys provided a giant step ahead in arctic data collection, especially from the

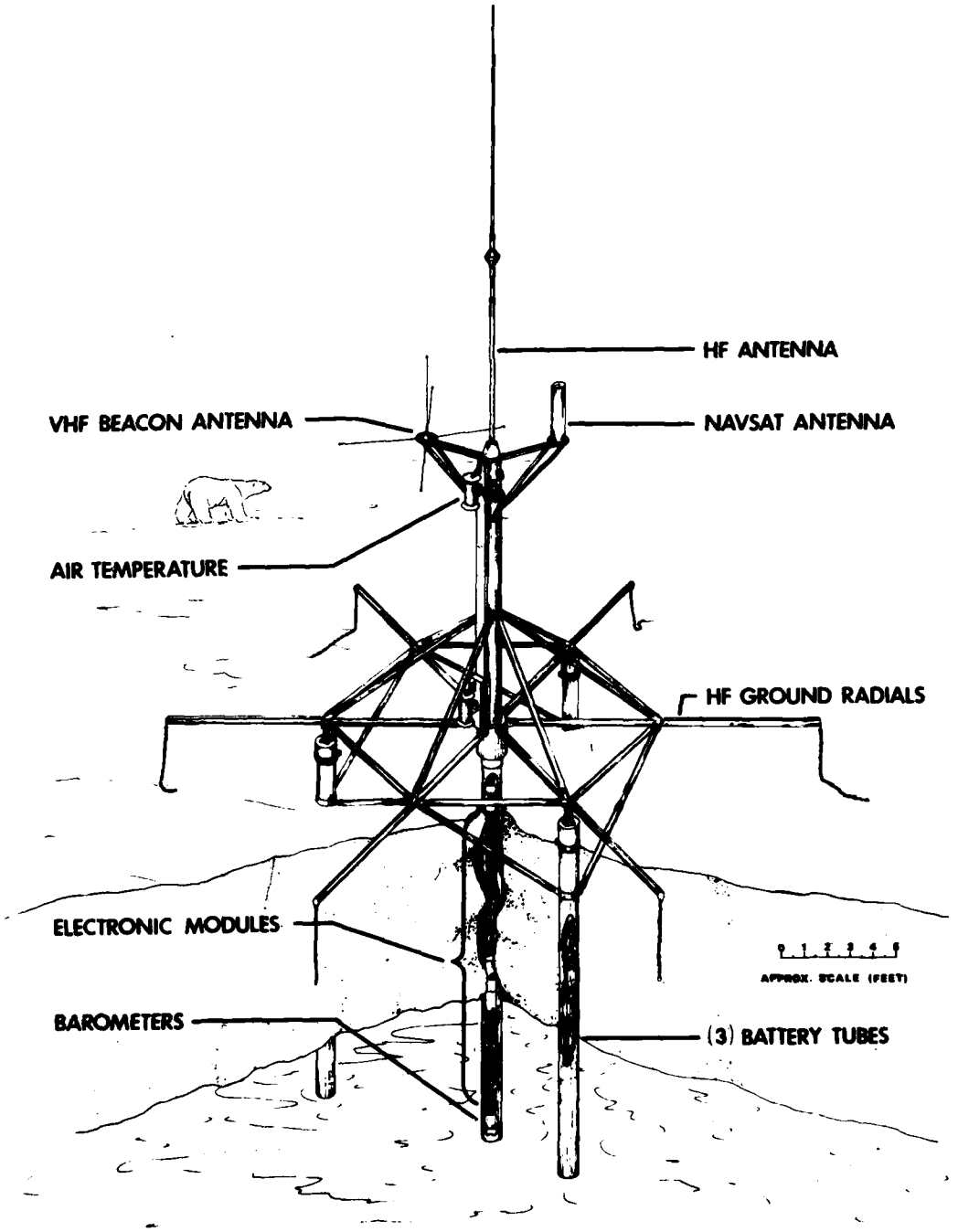


Figure 8—Arctic Environmental Buoy (AEB) with HF radio for data transmission, as used in AIDJEX 1975-1976.

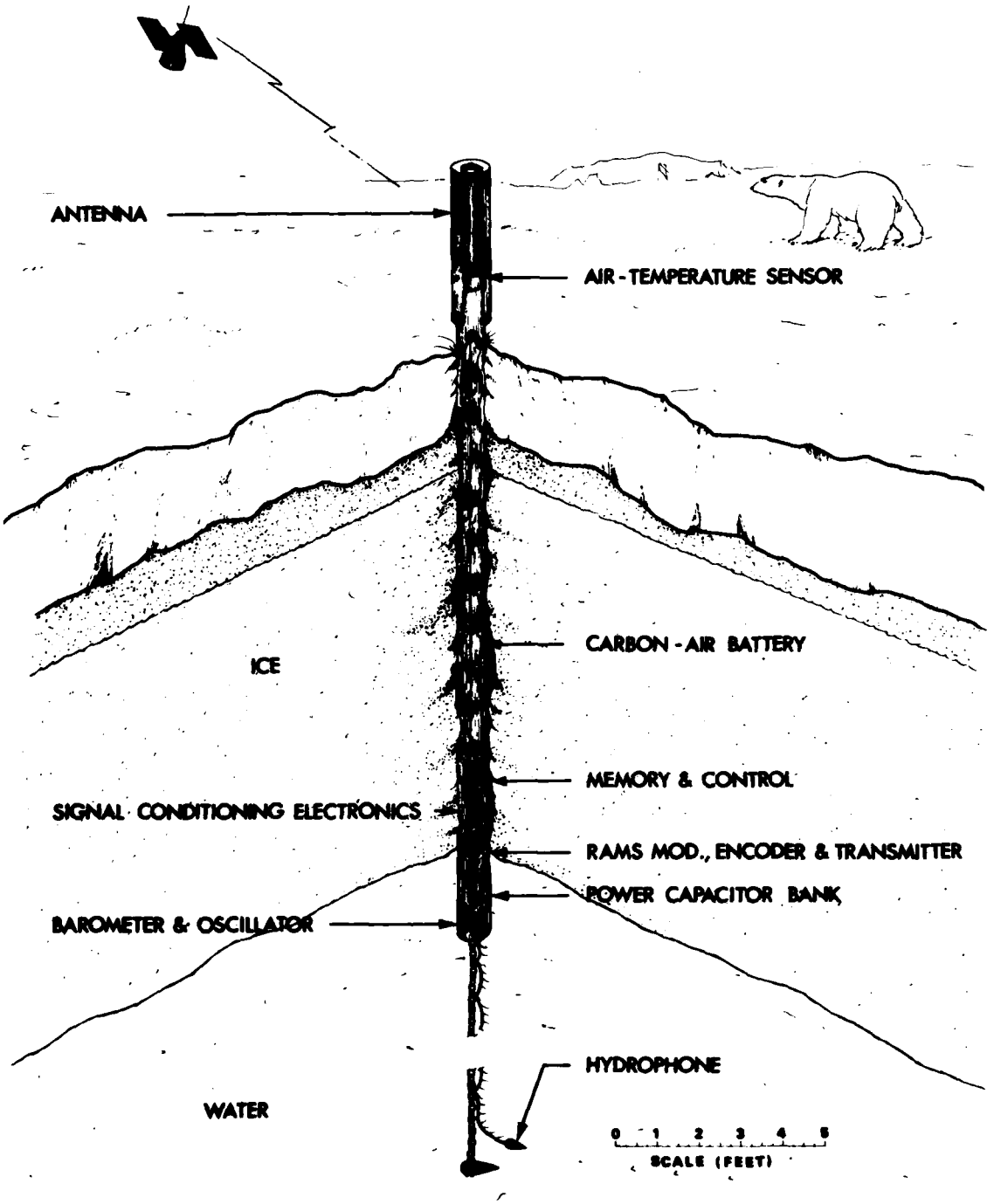


Figure 9—Synoptic Random Access Measuring System (SYNRAMS), with data transmission via Nimbus 6, as used in AIDJEX 1975-1976.

standpoint of cost reductions relative to manned ice camps, the air-dropped version went considerably beyond even that in that it further reduced installation costs.

The AIDJEX data buoys and newer, more sophisticated developments have a tremendous potential for future arctic scientific and operational use, not only in the underwater acoustic program but in most arctic science disciplines. Through their use, the next few years could see a quantum jump in our understanding of the Arctic Ocean.

For example, long-life data buoys that draw their operating power from the environment (e.g., wind, solar, water currents below the ice) can be installed where the known pack movement will carry them into still unresearched areas, or they can be air-dropped directly in those areas for research purposes. The buoys can be used to provide real-time weather data inputs for improving weather forecasting and, using the new AIDJEX models for ice dynamics, for forecasting ice movement. They can also be used for ground-truth measurements in conjunction with remote-sensing techniques from aircraft and satellites.

One of the more attractive applications of remote automatic data buoys in the Arctic is in conjunction with a mannable ice camp. Scientific and operational data collection falls into three categories: regular, long-term statistical sampling (e.g., of acoustic ambient noise or barometric pressure); continuous recording (e.g., arrays of current meters and thermistor strings for the study of oceanographic fronts and eddies); and concentrated diverse experimentation requiring investigators on the ice. Therefore, a self-navigating station that can be relocated and occupied for manned experiments or visited to collect tape recordings and that can operate all of the time as an automatic data collection station would be highly cost-effective. Such stations could also provide the United States with a "presence" over the total Arctic Ocean. A prototype station to determine concept feasibility will be installed a few hundred miles northeast of Barrow in the fall of 1976. The concept is called MUMMERS (Manned-UnManned Multipurpose Environmental Research Station), and the first station will be equipped to

- Self-navigate to an accuracy of 100 m
- Provide quarters for three men
- Relay, on command by HF radio, signals from underwater explosives for propagation studies
- Measure every 3 h, store, and relay data on barometric pressure, air temperature, water-current speed and direction, windspeed and direction, earth's magnetic field x and y vectors, solar radiation, water depth, and acoustic ambient noise in one-third-octave bands.

The automatic data collection tasks that can be performed with future MUMMERS installations are limited only by our imagination and technological competence. While the day of men on the ice is not past, the future will see more and more manual tasks performed by automatic systems. This will mean greatly improved quality, temporal and spatial extensions, and reduced costs.

REFERENCES

1. J. Sater, ed., *The Arctic Basin*, Arctic Institute of North America, Washington, D.C., 1976, 319 p.
2. N. Untersteiner, "Arctic Ice Dynamic Joint Experiment," *Arctic Bull.* 1 (4), 145–159 (1974).
3. D. A. Rothrock, "The Mechanical Behavior of Pack Ice," *Annu. Rev. Earth Planetary Sci.* 3, 317–342 (1975).
4. M. D. Coon et al., "Calculations To Test a Pack Ice Model," *AIDJEX Bull.* 31, 170–187 (1976).
5. R. S. Pritchard, M. D. Coon, and M. G. McPhee, "Simulation of Sea Ice Dynamics During AIDJEX," *Proceedings of American Society of Mechanical Engineers; International Joint Petroleum Mechanical Engineering and Pressure Vessels and Piping Conference*, Mexico City, 1976 (in press).
6. A. S. Thorndike et al., "The Thickness Distribution of Sea Ice," *J. Geophys. Res.* 80, 4501–4513 (1973).
7. G. A. Maykut, "Energy Exhange over Young Sea Ice in the Central Arctic," *AIDJEX Bull.* 31, 45–74 (1976).

8. M. I. Budyko, "The Future Climate," *EOS* **53**, 868–874 (1972).
9. W. W. Kellogg, "Climate Feedback Mechanisms Involving the Polar Regions," in *Climate of the Arctic*, G. Weller and S. A. Bowling, eds., Twenty-fourth Alaska Science Conference, University of Alaska, Fairbanks, Alaska, Geophysical Institute, 1973, pp. 111–116.
10. National Academy of Sciences, *U.S. Contribution to the Polar Experiment (POLEX)*, Part 1, POLEX-GARP (North), National Academy of Sciences, Washington, D.C., 1974, 119 p.
11. Global Atmospheric Research Programme (GARP), "The Physical Basis of Climate and Climate Modelling," GARP Publ. Series No. 16, 1975, 265 p.
12. National Academy of Sciences, *Proposed Scientific Plan for the Nansen Drift Station*, National Academy of Sciences, Washington, D.C., 1976, 260 p.
13. P. C. Martin and C. R. Gillespie, "Five Years of Data Buoys in AIDJEX," *Proceedings of the Symposium on Meteorological Observations from Space: Their Contributions to the First GARP Global Experiment*, Philadelphia, June 1976 (in press).
14. W. J. Campbell et al., "An Integrated Approach to the Remote Sensing of Floating Ice," *Proceedings of the Third Canadian Symposium on Remote Sensing*, 1975, pp. 39–72.
15. N. Untersteiner, "Natural Desalination and Equilibrium Salinity Profile in Perennial Sea Ice," *J. Geophys. Res.* **73**, 1251–1257 (1968).
16. N. Untersteiner, *Dynamics of Sea Ice and Glacier and Their Role in Climatic Modelling*, GARP Publ. Series No. 16, ICSU-WMO, 1975, pp. 206–224.
17. L. K. Coachman and K. Aagard, "Physical Oceanography of Arctic and Subarctic Seas," in *Marine Geology and Oceanography of the Arctic Seas*, Springer, New York, 1974, pp. 1–72.
18. K. Hunkins, "The Oceanic Boundary Layer and Stress Beneath a Drifting Floe," *J. Geophys. Res.* **80**, 3425–3433 (Aug. 1975).
19. K. Hunkins, "Subsurface Eddies in the Arctic Ocean," *Deep Sea Res.* **21**, 1017–1030 (1974).
20. J. L. Newton, K. Aagard, and L. K. Coachman, "Baroclinic Eddies in the Arctic Ocean," *Deep Sea Res.* **21**, 707–710 (1974).
21. R. M. Demenitskaia and K. L. Hunkins, "Physiographic Provinces of the Arctic Ocean," in *The Sea*, vol. 4, part 11, John Wiley & Sons, New York, 1971, pp. 223–249.
22. D. L. Clark, "Arctic Ocean Ice Cover, Late Cenozoic History," *Geol. Soc. Amer. Bull.* **82** (12), 3313–3323 (1971).
23. P. R. Vogt and O. E. Avery, "Tectonic History of the Arctic Ocean: Techniques and Interpretations and Unsolved Mysteries," in *Marine Geology and Oceanography of the Arctic Sea*, Springer, New York, 1974, pp. 83–117.
24. N. A. Ostenso and R. J. Wold, "Aeromagnetic Survey of the Artic Ocean: Techniques and Interpretations," *Mar. Geophys. Res.* **1**, 178–219 (1971).

ARCTIC ACOUSTICS BIBLIOGRAPHY

Anderson, J. O., et al., "Results of Acoustic Measurements, Marginal Sea Ice Zone, Winter Bering Sea 1973—Part A," Polar Research Lab TR003, Nov. 1973.

Anderson, J. O., B. M. Buck, and R. G. Paquette, "Marginal Sea Ice Zone—Pacific Study 1971 Experiment (Part A)," Delco Electronics Rep. TR71–62, Dec. 1971.

Buck, B. M., "Arctic Acoustic Transmission Loss and Ambient Noise," presented at ONR Arctic Drifting Stations Symposium, Warrenton, Va., Apr. 13–15, 1966.

Buck, B. M., "Low Frequency Underwater Acoustic Measurements in the Arctic Ocean 1965–1968," ACDRL Rep. TR67-10, Feb. 1969.

Buck, B. M., and C. R. Greene, "Arctic Deep Water Propagation Measurements," *J. Acoust. Soc. Amer.* **36**(6), 1526 (June 1964).

Buck, B. M., and C. R. Greene, "Arctic DIMUS Sonar Performance," presented at the 24th Navy Symposium on Underwater Acoustics, U.S. Navy Air Development Center, Johnsville, Warminister, Pa., Nov. 29–Dec. 1, 1966.

Buck, B. M., M. McLennan, and M. Springer, "Underwater Acoustic Measurements in the Arctic Ocean Using a Tropospheric Scatter Radio Telemet-

ering System," General Motors-DRL Report TR63-201, Jan. 1963.

Diachok, O. I., "Effects of Sea Ice Ridges on Sound Propagation in the Arctic Ocean," *J. Acoust. Soc. Amer.* **59** 1110–1120 (May 1976).

Ganton, J. H., and A. R. Milne, "Temperature and Wind-Dependent Ambient Noise Under Midwinter Pack Ice," *J. Acoust. Soc. Amer.* **38,** 406–411 (Sept. 1965).

Garrison, G. R., and E. A. Pence, "Studies in the Marginal Ice Zone of the Chukchi and Beaufort Sea, A Report on Project MIZPAC-71B," University of Washington, Applied Physics Laboratory, Rep. No. APL-UW 7223, Jan. 31, 1973.

Greene, C. R., "Under Ice Acoustics at High Frequencies," AC-DRL Rep. TR 6576, Oct. 1965.

Greene, C. R., and B. M. Buck, "Arctic Ocean Ambient Noise," *J. Acoust. Soc. Amer.* **36** (6), 1218 (June 1964).

Greene, C. R., and B. M. Buck, "Directional, Spectral and Statistical Properties of the Underice Noise in the Arctic," presented at 21st Navy Symposium on Underwater Acoustics, U.S. Naval Research Laboratory, Washington, D.C., Dec. 2–4, 1963; DRL Rep. TR 65-22, Mar. 1965.

Hunkins, K. L., "The Seasonal Variation in the Sound Scattering Layer Observed at Fletcher's Ice Island (T3) with a 12 Kc/s Echo Sounder," *Deep Sea Res.* **12,** 879–881 (1965).

Hunkins, K. L., and H. W. Kutschale, "Shallow-Water Propagation in the Arctic Ocean," *J. Acoust. Soc. Amer.* **35,** 542–551 (1963).

Hunkins, K. L., H. W. Kutschale, and J. K. Hall, "Studies in Marine Geophysics and Underwater Sound From Drifting Ice Stations," Lamont-Dohert, Final Report NONR 266 (82), Sept. 1969.

Kutschale, H. W., "Long-Range Sound Transmission in the Arctic Ocean," *J. Geophys. Res.* **66,** 2189–2198 (1961).

Kutschale, H. W., "The Period Equation by Ray Theory for Propagation in the Arctic Sofar Channel," *J. Underwater Acoust.* **21,** 37 (Jan. 1971).

Lyon, W. K., "Ocean and Sea Ice Research in the Arctic Ocean via Subm., *Trans. N. Y. Acad. Sci.* **23,** 662–674 (1961).

Marsh, W. H., and R. H. Mellen, "Underwater Sound in the Arctic Ocean," *J. Acoust. Soc. Amer.* **35,** 552–563 (1963).

Mellen, W. H., and W. H. Marsh, "Underwater Sound Reverberation in the Arctic Ocean," *J. Acoust. Soc. Amer.* **35,** 1645–1648 (1963).

Milne, A. R., "Sound Propagation and Ambient Noise Under Sea Ice," in *Underwater Acoustics,* Chap. 7, Vol. 2, Plenum Press, New York, N.Y., 1967.

Milne, A. R., "Statistical Description of Noise Under Shore-Fast Sea Ice," *J. Acoust. Soc. Amer.* **39,** 1174–1182 (1967).

Milne, A. R., "Underwater Backscatter Strengths of Arctic Pack Ice," *J. Acoust. Soc. Amer.* **36,** 1551–1556 (1964).

Milne, A. R., and J. H. Canton, "Diurnal Variations in Underwater Noise Beneath Springtime Sea Ice," *Nature* **221,** 851–852 (Mar. 1969).

Paquette, R. G., and R. H. Bourke, "Oceanographic Investigation of the Marginal Sea Ice Zone of the Chukchi Sea—MIZPAC 1974," Naval Postgraduate School Rep. No. NPS-58PA76051, May 1976.



Oscar Karl Huh is an Associate Professor at the Coastal Studies Institute of Louisiana State University, Baton Rouge. He is a geologist and oceanographer specializing in remote sensing of coastal processes. Dr. Huh was employed by the U.S. Naval Oceanographic Office, Research and Development Department, Ocean Science Center, from 1967 to 1976. His geological studies of stratigraphy, sedimentation, and structural geology in east-central Idaho and southwestern Montana resulted in new correlations and the now standard formational subdivision of the Mississippian System in this region. He has conducted research on bottom currents off Southern California; investigated currents and water masses in South Korean coastal regions; made coastal oceanographic studies in the Korea Strait and the Sea of Japan (by remote sensing as well as by conventional methods); and developed and tested methods for locating and measuring sea surface temperature gradients with the Defense Meteorological Satellite system. Dr. Huh was born in Hackensack, N.J. He received a B.S. in Geology from Rutgers, The State University of New Jersey, in 1957 and M.S. and Ph.D. degrees in Geology from Pennsylvania State University in 1963 and 1968. He is a member of the Society of Economic Paleontologists and Mineralogists, the American Geophysical Union, the Oceanic Society, and the American Association for the Advancement of Science.



Vincent E. Noble is Special Assistant for Navy Environmental Remote Sensing at the Naval Research Laboratory, where he has been employed since 1972. Dr. Noble's responsibilities include the development of remote-sensing techniques and data handling and analysis methods for Navy requirements. From 1960 to 1968 he was an Associate Research Physicist at the Great Lakes Research Division of the Institute of Science and Technology, the University of Michigan, and from 1968 to 1972 he was employed by the U.S. Naval Oceanographic Office as Research Physicist in the Airborne Remote Sensing Oceanography Project and the Polar Oceanography Division. He has worked in the fields of atmospheric turbulence, the air-sea energy exchange, physical limnology, and remote sensing of the environment. Dr. Noble received B.A., M.S., and Ph.D. degrees from Wayne State University, Detroit, Mich.

REMOTE SENSING OF ENVIRONMENT: ACHIEVEMENTS AND PROGNOSIS

Oscar Karl Huh

Coastal Studies Institute
Louisiana State University
Baton Rouge, La.

Vincent E. Noble

Naval Research Laboratory
Washington, D.C.

With the nation's energetic quests to maintain military superiority, exploit aerospace technology, and cope with burgeoning environmental problems, an explosive growth in Earth observations has taken place in the last 30 years. Methods of remote sensing of environment have played a major role. Remote sensing of environment is the detection of conditions of terrain, waters, and atmosphere of the Earth by remotely positioned sensors that detect the properties of reflected, scattered, and emitted electromagnetic energy. Information on these conditions is obtained by interpretation of the acquired data arrays, using models, equations, simultaneous direct measurements, or a prior knowledge of the environment. Remotely positioned sensors, as discussed in this article, are predominantly cameras, photometers, radiometers, and radar receivers mounted on aerospace platforms. (Underwater remote sensing by acoustical means is specifically excluded from this discussion.)

The remote-sensing approach is in most cases a logical supplement to the existing capabilities in the environmental disciplines for extending present measurements and observations. In many cases, however, it has provided the first opportunities to discover and deal with a whole new set, or a previously intractable set, of problems. With today's sensors, it is possible to map the earth in great detail and in any portion of the globe to make soundings of temperature and humidity; trace gas profiles of the atmosphere; sound depths and infer particulates in shallow seas; and measure the spectrum of land and ocean roughness. The scientific core of these measurement capabilities is the applied physics of electromagnetic radiation—its propagation, detection, and, most of all, its interactions with the solids, liquids, and gases of the earth. This branch of physics, combined with advances in sensor engineering, electronic data processing, communications technology, and aerospace technology, makes up the technological field of remote sensing. Remote sensing of environment is the application of this technology to environmental research or problems. However, the data must be converted into information, and key to the utility and relevance of the technology lies in the abilities of the environmental sciences to provide concepts and models for correct data interpretation. A powerfully synergistic interaction between the technology and scientific disciplines has taken place. Remote-sensing technology spawns new kinds of scientific achievements, and the sciences in turn spawn new concepts in remote sensing. This technological and scientific field has changed drastically in the last 30 years. Originating as subjective analysis of occasional daytime aerial photographs, it has advanced to the automated analysis of data on the Earth's surface, waters, and atmosphere, acquired several times

daily by solar-powered manned and unmanned satellites.

Four major factors have contributed to this growth of remote sensing: defense requirements for early warning reconnaissance and surveillance, rapid expansion of aerospace technological capabilities, rapid development of large-capacity computers and numerical methods, and widespread political awareness of high-priority environmental problems. With the advent of nuclear stalemate and high-speed weapons systems such as the ICBM or fractional orbital missile, priority was placed on surveillance of potential enemies. As a result, remote-sensing systems were developed to monitor potentially hostile activities without violation of treaties or the sovereignty of nations. Motivated by the challenges of superiority in space, space exploration, and placing man on the moon, the technological capabilities of this industrial society made immense strides in aerospace technology. The National Aeronautics and Space Administration, founded in 1958, became the focus of the civilian effort in U.S. Aerospace programs. The first environmental products of this effort were the instrumented aircraft and the experimental and operational satellite systems. The Television Infrared Observational Satellite (TIROS I, 1960) was the first environmental satellite; it inaugurated the photography of cloud cover from unmanned spacecraft. The experimental TIROS, NIMBUS (1964), and the Advanced Technology Satellites (ATS-1, 1966) led to the presently operational NOAA LANDSAT and SMS/GOES series of environmental satellites. These systems have provided large quantities of remotely sensed data for experimentation, and various Federal agencies funded investigators who had imaginative or utilitarian experimental concepts. Data became available, and at relatively low cost. Parallel with the aerospace developments was the development of large-capacity computers and numerical methods.

A major statistical and mathematical awakening has occurred in the environmental sciences in the last 30 years, particularly in those involving regional studies and geographic variability. The large volumes of data from operational satellites rapidly overwhelmed all previous concepts of data processing. The operational requirements in meteorology forced development of rapid-turnaround capability, from data acquisition to analysis and dissemination. These new meteorological data decreased in value rapidly with age. Timeliness became vital, and so rapid ingestion, processing, and output became as important as the basic acquisition of the data. As observations and measurements with LANDSAT focused on the subtleties of the earth scene, quantitative analyses of data on computer-compatible tapes rapidly superseded qualitative studies of photographic image reproductions. Mathematical modeling programs of time-dependent natural processes have created a strong "appetite" for the time-lapse, geographically extensive numerical data fields available from remote-sensing systems.

Thus came the means, the functional establishment, and the myriad of individuals capable of action. The final ingredient, political pressure on behalf of the environment, grew in parallel. Environmental concerns have expanded rapidly in the last 30 years, particularly in the advanced industrial nations. Excessive pollution—degradation of waters, lands, and atmosphere—began to severely affect the quality of life in the rapidly expanding urban and suburban regions. In extreme cases, pollution produced serious health problems for large population centers. Pollution in one form or another has affected all citizens, and ecology has become an appropriate populist concern. Shortages of low-cost resources required by industrial economies have stimulated exploration and survey of large remote regions. New development in overcrowded regions has required plans for development as many major interest groups within society have competed, presenting to political leaders conflicting demands for space and environmental quality. Prediction of the consequences or environmental impact of major development plans have become a primary consideration. Increased sophistication of weapons systems and required precision of military operations have made the systems more environment "sensitive" than previously. If military missions are carried out with inadequate information on adverse environmental conditions, the result may be operational failure and loss of lives, capital assets, and military objectives. Thus the objectives of relevant military research and development, even more stringent

than in the case of civilian requirements, must be constantly weighed against a "zero-failure" criterion.

Faced with these problems, politicians and program managers turned to environmental scientists and engineers for new knowledge, and they in turn required the capabilities of advanced technology for dealing with problems of unprecedented size and complexity. Thus, remote sensing of environment and a sister technology, the satellite relay of data telemetered from remotely dispersed in-situ sensors, have become vital tools. The basic tools for the assault on the environmental problems of the late 1970s, 1980s, and beyond are now costly conventional surveys, remote sensing of environment, telemetry from automatic in-situ sensors, and numerical modeling.

The birth and growth of modern remote sensing of environment has received major impetus from the Office of Naval Research (ONR). The leadership role has far exceeded the proportion of funding provided through this program. The very term "remote sensing" originated with the Geography Programs of this agency in about 1961. It was created in the redesignation of a project entitled "Interpretation of Aerial Photographs" to "Remote Sensing of Environment." It was a natural and appropriate change, in view of the development of sensors to make observations in regions of the electromagnetic spectrum beyond the ranges of human vision and photographic sensitivity. It was recognized that a new term was needed to encompass the total of observational processes from remote platforms.

In February 1962, ONR sponsored the first symposium on remote sensing of environment at the University of Michigan's Willow Run Laboratories. By the seventh symposium, in 1971, these meetings had become international, and attendance had expanded from 70 U.S. scientists in 1962 to more than 800 from 27 countries [1]. In October 1975 more than 165 papers were presented at the Tenth International Symposium on Remote Sensing of Environment, at which there were 646 attendees. Sponsors included 13 U.S. Government agencies, an agency of the Republic of China, a Japanese corporation, and a Spanish university. A professional journal (*Remote Sensing of Environment*, established in 1972) and trade journals, along with numerous special education programs, have come into existence. Existing scientific societies regularly hold remote-sensing sessions at conferences. ONR provided an important start here and continues to fund selected areas of research with potential for Navy missions.

ACHIEVEMENTS AND STATUS

The achievements and status of the field of remote sensing of environment are most vividly illustrated by today's operational systems and the concepts of those under development for use in the late 1970s and early 1980s. Even a brief review of the achievements in remote sensing of environment is a large task. Here it will be abbreviated to an outline of past and near-future milestones (Table 1, at the end of this paper) and a discussion of a few selected concepts, including the Defense Meteorological Satellite Program, sea-surface temperature measurements, vertical temperature and humidity profiling of the atmosphere, uses of reflected visual and near-infrared radiation, remote sensing by active and passive microwave sensors, tracking of balloons, buoys, floats, and satellite data collection platforms, laser sounding of the ocean, and detecting and measuring properties of soil and rock. The existing systems include the Defense Meteorological Satellite Program (DMSP Block 5-C Satellites), the NOAA Improved TIROS Operational Satellites (ITOS), the NOAA Stationary Meteorological Satellites (SMS/GOES), the LANDSAT series (ERTS A and B), and the NASA NIMBUS series and Skylab experimental satellites. Table 1 lists the systems chronologically, with the major capabilities achieved. The systems under development include the DMSP Block 5-D satellites, the TIROS-N, the NIMBUS-G, SEASAT-A, LANDSAT-C, STORMSAT-A, Applications Explorer Mission (AEM-A), Remote Ocean Measurement System (ROMS), and Synchronous Earth Observatory Satellite (SEOS). Each of these latter systems is built upon the concepts and technological advancements conceived or developed in the earlier aircraft or satellite experiments. They will expand our environmental remote-sensing capabilities in both a

qualitative and a quantitative sense. SEASAT-A represents a new thrust, as the first dedicated oceanographic satellite and the first incorporating passive and active microwave primary mission sensors. The TIROS-N satellite marks a new age of cooperation in space, among the Department of Defense, the Department of Commerce, NASA, and several NATO allies. This system will be based on the spacecraft of the DMSP, with sensors developed by NASA, the United Kingdom, and France. The Department of Defense will launch these satellites (two operational) into space and NOAA will conduct the remote-sensing operations and distribute data [2]. In the following paragraphs a review of selected concepts is presented in a framework of results from operational systems.

Defense Meteorological Satellite Program

The Defense Meteorological Satellite Program is a premium system for military use, having been originated and developed by the Air Force in response to real-life problems of Southeast Asia operations [3]. It is now a joint services program managed by the Air Force and a paradigm for environmental information support systems. The system includes 2 polar orbiting satellites, 2 primary receiving sites (Maine and Washington), the Air Force Global Weather Central in Nebraska, some 20 transportable terminals deployable within hours, and 2 Navy shipboard units on carriers. The satellites are in 830-km-high, sun-synchronous, polar orbits, one with the orbital plane at solar meridian and one in near terminator meridian. The sensor package includes the scanning visible/infrared radiometers, the scanning infrared radiometer (a vertical atmospheric temperature profiler), and the ambient electron monitor.

The scanning visible/infrared radiometers use two spectral bands, the visible near infrared (0.4–1.1 μm) and the thermal infrared (8–13 μm). The swath of data beneath each satellite is approximately 3000 km. There are four channels of data: the 3.7-km-resolution visible (HR) and infrared (HRIR) and the 0.6km-resolution visible (VHR) and infrared (WHR). The HR visible channel is a unique achievement. It has an automatic gain control, which uses data from an incident solar radiation sensor to control the channel gain. It thus acquires radiance values that represent scene albedo. The sensor has sun shades and glare-suppression devices that, combined with the automatic gain control, allow this channel to provide usable visual data through the day-night terminator and on the dark side of the earth. It has also been able to record such astonishingly low light levels as nighttime city lights, lightning flashes, aurora borealis, and "moon glint" on the sea surface [3]. This sensor is presently the sole available source of the satellite imagery that illustrates North America and Western Europe by nighttime patterns of city lights.

The capability of transmitting direct readout HR, HRIR, and either VHR (daytime) or WHR (nighttime) in digital form to any tactical DMSP receiving site has paid major practical dividends. The DMSP provides near real-time, readily assimilated imagery on weather conditions to the analyst, yielding a quantum jump in the quality of environmental support. Input of satellite data into weather forecasts by analysts and digital processing for use in numerical prediction models have also made important improvements in these services.

The Defense Meteorological Satellite Program will soon launch a new advanced series of satellites, designated the Block 5-D. The Block 5-D satellite is a unique integrated spacecraft into which functions of the uppermost stage of the launch vehicle are incorporated. It guides itself into orbit from liftoff. This system is designed for longer spacecraft lifetime and to incorporate major improvements in data, as noted in Table 1.

ONR assumed a leadership role here in the early 1970s by sponsoring a study of the coastal oceanographic processes with the imagery at the U.S. Naval Oceanographic Office and a post-operations analysis of the tactical data obtained from the carrier U.S.S. *Constellation*. The former study, a combined ONR-7th Fleet/NAVOCEANO operation, crowded system capabilities far beyond normal meteorological requirements and achieved some surprising and useful results. Data from the VHR (0.6 km) visual sensors were used to detect sea ice and turbid river discharge plumes in coastal waters. The HRIR thermal infrared data were used to detect

sea-surface temperature gradient features and measure their movements and changes with time. Despite the coarse 1.6°C temperature quantization interval of the data, the HRIR sensor acquired an excellent series of images in the Sea of Japan, the Yellow Sea, and East China Sea in the fall and winter of 1972 [4]. This unique electro-optically contoured imagery provided near real-time information on the position and structure of sea-surface temperature gradients, as well as a means of quantitatively estimating temperature differences between water masses [5]. Examples of sea-surface data, including surface temperature gradients, albedo of turbid coastal waters, and sea ice, are shown in Figures 1–3.

Sea-Surface Temperature Measurements from Space

The surface of the sea radiates thermal infrared energy with an intensity proportional to its temperature. The ability of the satellites to detect and measure this radiation provides a potentially vital environmental data bonus for receiver-equipped aircraft carriers or task forces. Many regions of the globe have horizontal sea-surface temperature gradients that delineate tactically significant changes in the conditions for sound propagation. A wide range of very common water column features, both transient and permanent, vary significantly from climatological mean locations or conditions of temperature and salinity (i.e., sound velocity) for any given area. Some examples are ocean eddies (5–500 km scales) that thicken or thin the warmer surface layer of the sea by hundreds of meters, convergence zones between major currents, water mass boundaries on local or regional scales, and transient upwelling of deep waters.

Advances in studies of sea-surface temperature have been accomplished through research with the NOAA ITOS series satellites (Table 1). This system, in which scanning radiometers use the narrow 10.5–12.5 μm band in the 8–14 μm thermal infrared window region, is subject to less moisture and CO_2 absorption and avoids the 9.0-μm radiation-absorption peak of zone. Temperature corrections have been smaller than those required by the DMSP, and the Very High Resolution

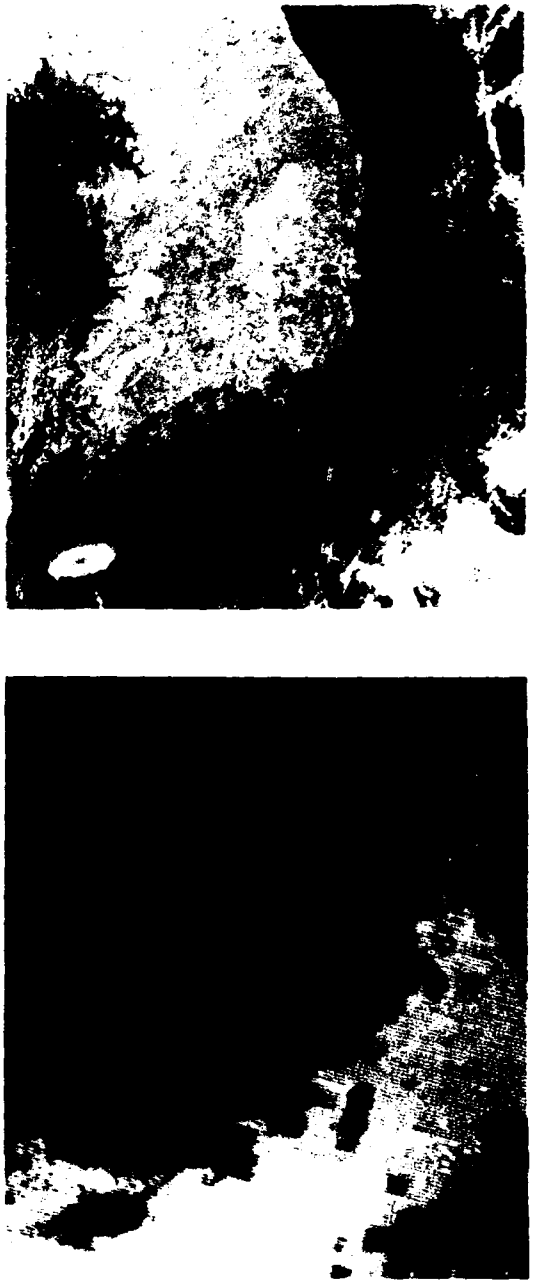


Figure 1—Direct-readout visual and infrared imagery from the Defense Meteorological Satellites. (A) VHR visual (0.6-km spatial resolution) image, a small portion of the full swath showing the Korean Peninsula and surrounding oceanic areas. (B) HRIR infrared (3.7-km spatial resolution) image, a small portion of the swath showing the Korean Peninsula and electro-optically contoured surface temperatures in surrounding seas (special enhancement not obtained simultaneously with A).



Figure 2—DMSP VHR albedo patterns along the Louisiana-Mississippi-Alabama coast. This image shows the turbid plumes of suspended sediment-laden coastal waters at the mouth of the Mississippi River, Atchafalaya Basin, Mississippi Sound, Mobile Bay, and west Louisiana continental shelf area.

Radiometer (1-km spatial resolution) has provided spectacular images of sea-surface temperature gradient featurs around the globe. Figures 4 and 5 show a portion of the Gulf Stream off the southeastern United States and the Kuroshio-Oyashio convergence off Honshu and Hokkaido, Japan, respectively. The direct-readout Automatic Picture Transmission (APT) capability of the NOAA satellites has been a most successful program. More than 500 low-cost APT sites are in operation around the world for direct access to the imagery for meteorological analysis twice daily. Portable APT has been successfully used to detect sea-surface temperature gradients and sea ice, as well as to direct field experiments [6]. This is the coarse 7.5-km spatial resolution infrared data that is available twice daily. The low cost and portability greatly facilitate deployment for experimental and practical use of the data. NOAA now operationally incorporates satellite data into the Monthly Gulf Stream Summary on the East Coast and provides location of oceanic fronts associated with upwelling along the West Coast.

Major problems beset this near real-time sea-surface temperature mapping capability. The outstanding results obtained by the infrared sensors of the DMSP and NOAA satellites are very contingent on atmospheric conditions. Field experiments have demonstrated how dependent the infrared remote-sensing capabilities are on outbreaks of dry continental air. These experiments have taken place in a variety of regions, including the U.S. East and West Coasts, the western Mediterranean, the Mexican west coast, New Zealand, Korea, Japan, the east coast of Africa, and the Hawaiian Islands. For example, with the



Figure 3—Direct-readout VHR visual data, 600-m spatial resolution, showing the sea ice canopy over the northwestern two-thirds of the Sea of Okhotsk. Note the shore leads around Sakhalin, the northern coast, leads in the icepack outlining the granular mesoscale structure of the ice, the frozen Tatar Strait, volcanic mountains on Kamchatka Peninsula, and the cumulus cloud streets where the cold, dry winds blow off the ice canopy over the open water.

advent of a cold front, cloud masses are swept seaward with the cloud shield of the front, and cool, dry, polar continental air replaces the warm, moist marine airmass (Figure 6). This provides optimum satellite imaging conditions for oceanic and terrestrial regions with both IR and visual scanning radiometers. The near uniformity, clarity, and low humidity of these airmasses provide reduced atmospheric attenuation of the reflected and radiated energy from the earth's surface. The warm, moisture-laden marine airmasses of the world severely attenuate the infrared radiation, not only changing the surface temperature by a variable amount but actually reducing the apparent strength of the sea-surface temperature gradients (Fig. 4). The processes of attenuation are



Figure 4—NOAA satellite Very High Resolution Radiometer, infrared image, April 4, 1975. Note the suppression of thermal gradients along an east-west line crossing Florida at about the latitude of Lake Okechobee. The Gulf Stream seems to disappear, and the Florida Peninsula nearly fades from view through the humid marine air mass. This is an atmospheric humidity front separating polar continental air to the north and warm, moist, but clear marine air to the south.



Figure 5—NOAA satellite Very High Resolution Radiometer, infrared image, April 4, 1975. Image shows portions of Honshu and Hokkaido, Japan, and the Kuroshio-Oyashio currents in the northwest Pacific. Strong outbreak of polar continental air from Siberia reveals spectacular sea-surface temperature patterns outlining the convergence of these major currents and the array of eddies at their juncture. Note the buildup of cumulus cloud streets offshore, obscuring the seaward extension of the gradient features.

absorption and scattering by water droplets and particulates, plus absorption and reemission of infrared radiation by the triatomic gas molecules in the atmosphere (water vapor, carbon dioxide, ozone, and nitrous oxide). Chief and most variable of these is water vapor. To develop a reliable operational capability, research and development must provide means of penetrating cloud cover, avoiding atmospheric attenuation, and exploiting the tactical advantages such data provide.

Vertical Temperature and Humidity Profiling of the Atmosphere from Space

In addition to mapping earth scenes with scanning radiation sensors, recent developments are providing data in the third dimension and remote sounding of temperature and humidity profiles of the atmosphere. The Vertical Temperature Profiling Radiometers (VTPR) are multiple-wavelength, infrared (and now microwave) radiometers that provide temperature and humidity soundings. This technique was suggested by Kaplan in 1969, was proven in NIMBUS III and IV experiments, and became operational in 1972 aboard NOAA satellites. It uses measurements of radiation emerging from the top of the atmosphere in a series of spectral intervals ranging from the centers to the wings of the strong constituent gas absorption bands (carbon dioxide, nitrous oxide, and oxygen absorption peaks or maxima). Thermal radiation from a spectral interval near the opaque band center arises from higher altitudes because of atmospheric absorption of energy emitted from lower levels. Radiation measured from the wing of the absorption band comes from lower altitudes because of high transparency of the upper atmosphere. Since the distribution of these gases is known well enough, the measured variation of outgoing radiation in the various spectral intervals can be interpreted in terms of vertical atmospheric temperature profiles [7]. With the temperature profile and one or several water vapor channels, the VTPR provides temperature and humidity profiles of the atmosphere needed for numerical atmospheric prediction models [8]. Until the satellite VTPR development, profiles of the atmosphere were available

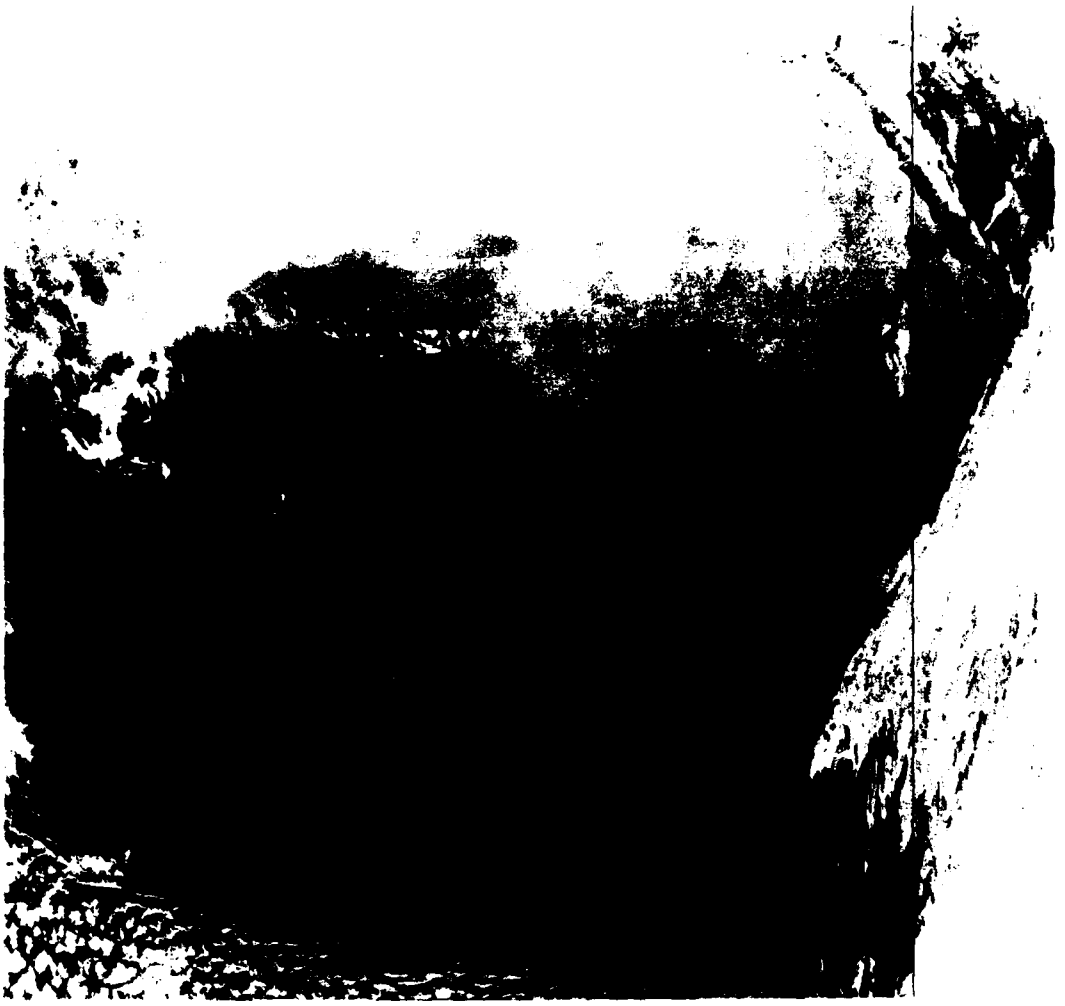


Figure 6—NOAA satellite Very High Resolution Radiometer, infrared image, Dec. 14, 1974, central Pacific region. A large outbreak of continental air moved west from Southern California/Baja California (to the far east, i.e., right side, in the image) to provide good imaging conditions for detection of mesoscale temperature gradients featured on the sea surface, the subtropical convergence of the North Pacific.

only from balloon and rocket sounding devices, which were poorly distributed in space and time [9]. The vertical temperature and humidity profiling capacity is still undergoing advanced development. This capability will eventually help correct sea-surface radiation temperatures for atmospheric effects. Quantitative measurements of temperature and humidity at various pressure levels in the atmosphere are among the most fundamental observations required for weather forecasting. These atmospheric sounders thus contribute to atmospheric forecasts, sea-surface temperature measurement, and weather forecast capabilities of the Navy.

Reflected Visual and Near-Infrared Radiation

The spectral composition of visual and near-infrared reflected skyward from the earth's surface has been altered by differential absorption, scattering, and attenuation by terrain features,

ocean, and atmosphere. The visual and near-infrared scanners of the operational and planned satellite systems are superseding the more costly aerial photography for larger scale environmental monitoring, resource exploitation, crop inventory, soil mapping, and hydrological-limnological-oceanological applications. The operational visual sensors in the LANDSAT series satellites provide opportunities for global coverage with a radiometric equivalent of color photography. The multispectral scanner has capabilities for detecting a wider dynamic range of signal intensities and electronically calibrating the measured intensities of reflected light. It provides an electronic output on computer-compatible tapes that are amenable to computer processing of the imagery. In contrast, color photography is extremely difficult to calibrate for color fidelity and most difficult to use quantitatively. A distinct advantage of the satellite imagery is that the long focal length of the satellite imaging devices provides a simple geometry across the field of view of the image. Because the scanner data are electronically calibrated, the effects of variations in solar illumination and atmospheric attenuation can be computed to provide quantitatively the reflective spectral signatures of the image features. The multispectral images from LANDSAT A and B are equivalent to four photographs, one each sensitive to the green, yellow, red, and near-infrared portions of the spectrum, with an 80-m spatial resolution.

Experiments of the NASA Earth Resources Program have demonstrated how isolated ground-truth points in an image may be used as calibration points to determine the spectral, textural, and structural characteristics of specific features in the scene. Computer classification programs can then be used to map or contour similar areas throughout the scene. For example, in agricultural scenes, individual farm fields can be identified as to crop type, bare soil, soil types, irrigated nonirrigated fields, grasses, forests, roads, or bodies of water. In coastal regions, computer programs may be used to discriminate sand beaches, marsh grass, palmetto, and scrub pine in the near-shore terrain. Bottom reflections from shallow depths provide a brighter image than reflectance from similar bottom types in deep waters. The spectral nature of the bottom reflectance (i.e., color) can be used to discriminate among sands, silt deposits, and vegetation such as eel grass and kelp. The most stringent requirements arise from military needs to describe the offshore shoals, surf zone width, beach slope, tidal ranges, and trafficability from the standpoint of soil bearing strength, roads, creeks, swamps, and other advantages or impediments to maneuvers.

The spectral composition of light reflected by the oceanic regions of the earth is altered by differential absorption and scattering by the sea surface, water, dissolved substances, particulate matter (living and nonliving), and bottom. The aggregate of this light may be classified as bottom reflection (bottom color), diffuse radiance (water color), or specular radiance (wave-facet reflectance). Incident radiation from the Sun is peaked at 0.475 μm, radiation emitted by the surface of the sun at a temperature of 6000°K. A naturally occurring spectral interval of high transparency in clear water at this same wavelength allows some passive measurement of oceanographic properties as a function of depth. Oceanographic uses have included bathymetric mapping by photogrammetric methods (clear, shallow-water regions), study of variations of suspended sediment [10], measurement of chlorophyll content (biological productivity) of coastal waters [11], and detection of ocean currents [12].

Measurement of bathymetry with passive sensors such as LANDSAT is severely limited by lack of stereo capability, low sensor gain (LANDSAT), lack of homogeneous sea floor, variable sea state, sun angles, and interference by near-bottom or upper-water-column suspended sediment loads. Couple these difficulties with the low repetition rate of LANDSAT (18-day revisit period), and the result is low reliability. Remote measurement of suspended sediment using variations in diffuse radiance is equally encumbered by

1. Variations in depth of the turbid layer that will alter the spectrum as well as the intensity of the diffuse radiance.
2. Gelbstoff, the yellow human substance in coastal waters from terrestrial runoff, which can severely bias any measurements using, for example, LANDSAT band 4 (green band [12].

3. Differences in sea state, which will alter the total radiance with specular reflection from waves in the sun glint portion of images.

The appropriate direction here is for development of active sensors, initially for low-flying aircraft or drones. In this mode, stereo capability, laser sounding, measurement of sea state, and detection of suspended sediment are possible.

Remote measurement of chlorophyll in surface waters has promise. Larger amounts of chlorophyll are associated with relative decrease in the blue portion of the spectrum and an increase in the green [11]. Useful remote measurement must detect a change of 0.3 mg/m^3 of chlorophyll, representing a change of 10% of the normal range of oceanic levels [13]. Remote measurement of color is complicated by loss of scene contrast owing to air light at high altitudes, but not sufficiently to prevent most required measurements. High chlorophyll concentrations in surface waters indicate high levels of biological productivity. The potential here is for detection of areas with probability of high concentrations of soniferous marine life (biological noise sources).

The use of color to detect the position and movement of current boundaries across broad regions of the sea has a large number of practical applications. This capability has been demonstrated repeatedly by aircraft and satellites using infrared sensors over the edge of the Gulf Stream and other strongly baroclinic features around the world. Based on the sea-surface temperature discontinuity, this capability is lost during summer months, when isolation makes all surface waters isothermal. Color measurements by satellite can detect the current edges in two ways: by color change (difference in optical property of seawater) and by change of specular radiance (sea state/surface albedo) [12]. Differences in albedo of the sea surface have also shown sets of linear surface features identified as surface expression of internal waves [14]. This capability is not feasible for any operational capability inasmuch as it is very contingent on sea state, sun angle, and weather and satellite subpoint track. All-weather, day-night radar systems, however, do have potential for operational detection of internal waves for sonar applications.

There are three major difficulties in applying data from LANDSAT sensors to oceanographic uses: the spectral bands available are incorrect, the gain settings of the sensors (sensitivities) are too low, and the data are susceptible to contamination by specular reflectance from the wind-roughened sea surface. There are no spectral bands of the multispectral scanner in the blue portion of the spectrum (0.40–0.50μm). This is a severe limitation for oceanography because most of the spectral information from the sea is contained in this interval. The sensor gain settings are too low for the illumination levels of seawater regions. The radiance ranges of LANDSAT I sensors, for example, as estimated by Maul and Gordon [12], range from a minimum in band 7 of 0.05–0.40 mW cm^{-2} ster^{-1} to a minimum of 0.15–75 mW cm^{-2} ster^{-1} in band 4. A multispectral scanner optimized for coastal oceanography will be flown on NIMBUS-G. This Coastal Zone Color Scanner (CZCS) will have needed gain capabilities ranging from 1.34 to 11.46 mW cm^{-2} ster^{-1}. Specular reflection from the sun glint will contaminate radiance values obtained by LANDSAT. Sun glint avoidance is planned for the Coastal Zone Color Scanner by tilting the scanner away from the sun, up to 10 degrees ahead of or behind the spacecraft in 2-degree steps (NIMBUS-G; see Table 1).

For the unique advantages available from the visual portion of the spectrum it will be necessary to shift to sensor systems with illumination sources, the active sensors. The approach that appears most promising is that of driving a series of high-intensity pencil beams of near-monochromatic visible light down onto the earth scene. These illumination sources, dynamically coupled with high spatial and spectral resolution detectors, will much more reliably measure the desired properties of the environment. This approach will avoid the many vagaries of natural illumination and detect the many oceanic and terrain features with unique color and spectral signatures. Such sensors will require large power supplies and will probably be limited to reconnaissance aircraft or drone aircraft systems for many years.

The visual-range sensors of the meteorological satellites, NOAA, DMSP, and SMS/GOES have important applications (beyond meteorological), even with their low spatial and spectral resolu-

tions. Most important is detection of icefields and snowfields on the basis of the extreme reflectivity differences against water and terrain. Detection of sea ice, the marginal ice zone, and leads of open water in sea ice are of direct operational importance. Studies of ice dynamics benefit directly from the high repetition rates of the polar orbiting satellites, which provide imagery every couple of hours. Poor solar illumination levels and cloud cover are the present encumberances to this capability. NOAA has successfully monitored the snow fields of the United States, measuring changes in snowfield area and using these to predict meltwater production and thus river stages downstream.

Remote Sensing of Environment by Active and Passive Microwave Sensors

Microwave remote measurement systems promise to eliminate one of the major disadvantages or limitations of present systems, weather dependence. Detection of environmental conditions during periods of cloud cover, extreme atmospheric humidity (tropic and temperate summer seasons), as well as nighttime overcast, is a major requirement, particularly for the military, whose operations often must continue during periods of inclement or severe weather, closely following the limiting conditions of feasibility for mission accomplishment. Microwavelengths are actually long in comparison to visible and infrared wavelength, ranging between 1mm and 1m. Polarization is often used as a parameter of feature discrimination in the microwave region of the spectrum, inasmuch as the transmitting and receiving antennas are readily built with single polarization directions. In comparison to visual and infrared systems, the geometry of radar range measurement facilitates the use of observational angles well away from the vertical. By comparison, the visual and infrared sensors are most effectively used in the near vertical mode. Active microwave sensors provide their own illumination, whereas passive systems measure radiation originating elsewhere. With active systems the shape of the return pulse, the polarization, and the intensity of the backscattered microwave energy are all modulated by

1. Terrain roughness, plant canopy, snow cover, ice, soil type, and soil moisture.
2. Ocean surface winds, waves, temperature, salinity, nutrient and pollutant content, current and upwelling motions, falling rain, surface pressure, and the molecular species distribution and density in the atmosphere.

Similarly, thermal microwave energy emitted from the surface is modified by a series of microprocesses that vary with the wavelengths used. The various microwave bands vary in sensitivity to different scale features and have different transmissivities within the atmosphere and upper ocean or terrain surface. These differences across the spectrum of microwavelengths allow separation and quantification of various environmental effects using sensors mounted on the ground, aircraft, drone aircraft, or satellites.

The use of side-looking imaging radar in the synthetic-aperture mode is a technique, long familiar in aircraft operations, to obtain high-resolution, all-weather, rectilinear map images. The imaging radar detects changes in ocean-surface backscatter and yields imagery of a wide range of surface roughness or smoothness features, including deepwater gravity waves, smooth oil slick areas, internal waves, coastal waves, island shadows, and current and water mass boundaries. Figure 7 is an image, from a side-looking synthetic-aperture radar, of the ocean surface. It was obtained from the Naval Research Laboratory four-frequency radar. The approximate spatial resolution of this aircraft image is 25m. It illustrates the capability for measuring the ocean surface wave structure. The spatial variations in ocean wave structure outline important oceanic features. Often the radar reflection anomalies occur at the location of major and operationally important surface temperature gradients such as the boundary of the Gulf Stream. Under certain conditions the location of internal waves below the ocean surface may be manifest in the synthetic-aperture radar images because of changes in ocean-surface reflection coefficient resulting from concentration of slick-forming contaminants over the internal wave troughs.

Active microwave systems are in various stages of advanced development for measurement of the ocean surface topography, surface wave structure, and surface windspeeds and directions.

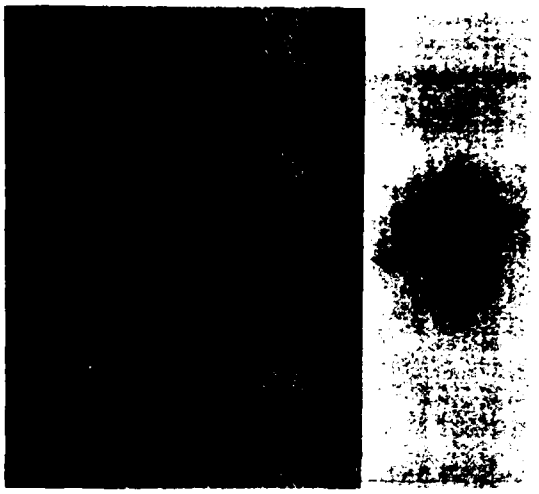


Figure 7—L-Band synthetic aperture radar image of the ocean surface at the Gulf Stream Boundary obtained with the NRL four-frequency radar system. (Moskowtiz [15]).

The parameters measured by these sensors are the range (distance) from the instrument to the reflecting surface, the shape of return pulse waveform, and the intensity of the reflected energy as a function of the angle of incidence and polarization of the radar beam. The shape of the radar return pulse is a convolution between the transmitted pulse shape and the roughness characteristics of the ocean surface in the illuminated spot. Analysis of Skylab, GEOS-3, and aircraft flight measurement data has demonstrated that the slope of the return pulse waveform may be used to estimate the significant wave height to an accuracy of 0.5 m within the radar altimeter footprint (a 7-km spot for SEASAT). Determination of the effects of sea state upon the return pulse waveform is necessary to design the altimeter range tracker so that the electromagnetic range may be related to the true mean sea level in order that measurements of the ocean surface topography will not be biased by the effects of local sea states. The Synthetic-Aperture Radar (SAR) will yield high-resolution images of coastal features and sea ice structures under all weather conditions; thus the SAR is a valuable complement, to, and in many cases a replacement for, multispectral visible region images such as those obtained from LANDSAT.

Analysis of airborne radar altimeter measurements made in a wind-wave radar configuration under a limited fetch condition has demonstrated that the slopes of the leading and trailing edges of the return radar waveform provide independent measurements of the significant wave height and surface windspeed. Analysis of return radar energy levels received from high incidence angles (30–50°) has demonstrated that the radar backscatter coefficient may be used to estimate windspeeds over large areas of the ocean surface. Two measurements made from orthogonal directions, with respect to the same point on the surface, make it possible to correct determinations of the radar cross section as a function of wind direction and wind-wave interaction. This offers the potential of determining the directional wind vector at the ocean surface. Scatterometer measurements from the SEASAT experiment will permit evaluation and calibration of the precision of these measurements. It is essential to note that analyses based on determinations of the radar backscatter coefficients from measurements of return radar signal strength must account for atmospheric propagation losses in order to determine accurate values for the ocean-surface characteristics that are being inferred. Data from passive microwave and infrared radiometers and sounders may be used to determine the atmospheric propagation corrections.

Altimeter measurements from Skylab and GOES-3 have demonstrated the feasibility of determining the shape of the marine geoid from the relatively stable satellite orbit. This measurement technique may be used to determine dynamic oceanographic processes by measurement of departures of the local ocean surface from the geoid. For example, dynamic topography associated with major current boundaries (such as the Gulf Stream) may be as large as 1 m, open-ocean tide ranges are on the order of 0.5 m, and hydrostatic pressure differences associated with atmospheric pressure changes (from 950 mbar low to 1050 mbar high) can cause changes of ocean-surface elevations up to 1 m. Assuming the capability for precise satellite orbit determinations, and assuming local knowledge of the marine geoid, high-precision radar altimetry (10 cm planned for SEASAT-A) may be used to provide data for evaluation and improvement of prediction models for ocean dynamics.

Although the effects of clouds and atmospheric constituents render the atmosphere opaque to infrared and visible portions of the spectrum, they appear translucent to selected frequencies in the microwave portion of the spectrum. Radiometric measurement techniques in the microwave portion of the spectrum can therefore be used to obtain near all-weather ocean measurement capability. Passive microwave radiometers measure the black-body radiation emitted by, and incident radiation reflected from, the ocean surface at microwave frequencies on the order of 0.3, in contrast to the infrared emissivity, which is nearly 1.0. Further, owing to the complex nature of the dielectric constant at microwave frequencies, the emissivity is a function of the frequency and polarization of the measurement. The measurement technique is to measure the microwave energy received from the sensor target and express the measured energy in terms of the equivalent temperature of a true black body that would have emitted the amount of energy that is received. Passive microwave radiometer measurements are therefore expressed in units of degrees of brightness temperature, which is the black-body temperature of the target being measured. Conceptually, then, the brightness temperature is a measure of the emissivity (or complex dielectric constant) of the target. For example, a microwave brightness temperature measurement of the ocean surface is a function of the frequency, polarization, and angle of incidence of the measurement; the temperature, salinity, and roughness of the ocean surface; and the propagation characteristics of the atmosphere, which are largely dominated by liquid water and water vapor. At frequencies considered for oceanographic applications, the roughness effect on the apparent brightness temperature is dominated by the capillary wave statistics generated by the local wind field. Therefore, the apparent dependence of brightness temperature upon surface roughness can be expressed as a dependence upon surface windspeed. At higher windspeeds, the ocean surface begins to show patches of foam coverage, which also tends to increase the apparent brightness temperature of the measurements. Thus the wind effects upon brightness temperature extend to windspeeds beyond the saturation range of the capillary wave structure. Figure 8 shows the frequency dependence of sensitivity of brightness temperature measurements as a function of salinity, sea-surface temperature, surface windspeed, and atmospheric liquid water and water vapor. As shown in the same figure, the lower microwave frequencies (2-6 GHz) are most sensitive to sea-surface temperature, but single-frequency measurements of surface temperature are not possible because of the high dependence upon surface windspeed. In practice, then, passive microwave sensor systems for measurement of ocean-surface parameters must utilize several microwave frequencies to permit the solution of multiple equations for simultaneous determination of windspeed, sea-surface temperature, salinity, and atmospheric propagation corrections. The atmospheric parameters determined from the passive microwave measurements may be used, too, for propagation correction for active microwave sensors. Examples of this class of sensor are the Scanning Multifrequency Microwave Radiometer (SMMR) being developed for the

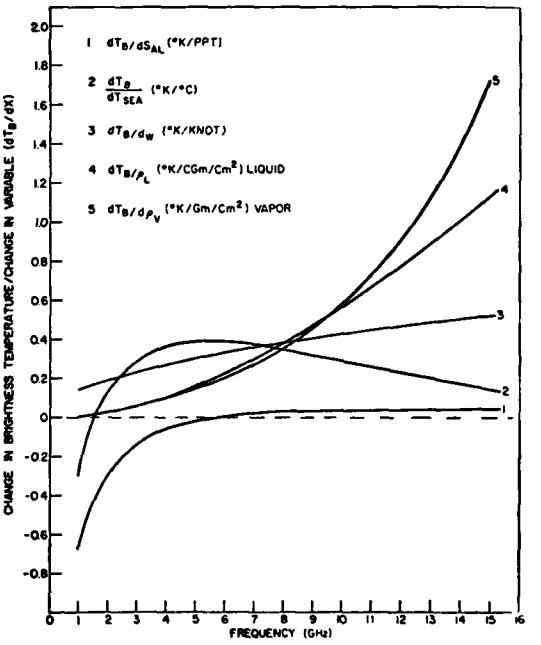


Figure 8—Frequency dependence of the sensitivity of brightness temperature measurements as a function of salinity, sea surface temperature, surface wind speed, atmospheric liquid water and water vapor. (Hollinger et al. [16]).

NIMBUS-G and SEASAT experiments and the follow-on Passive Microwave Sensor (PMS) component of the Remote Ocean Surface Measuring System (ROMS) under development at the Naval Research Laboratory.

The microwave emissivity of sea ice is a function of the dielectric constant of the medium. Sea ice crystals are surrounded by interstitial pockets of brine and contain air bubbles within the matrix. As the ice anneals with age and accretes fresh ice from consolidated precipitation on the surface, the emissivity also changes. Therefore, open water, first-year ice, and multiyear ice exhibit different brightness temperatures, which are also frequency dependent. Aircraft measurements have demonstrated that dual-frequency passive microwave measurements over sea ice can be used to minimize atmospheric propagation effects and yield the combined percentages of open water and first-year and multi-year ice within the instantaneous field of view, or "footprint," of the microwave sensor [3, 17].

Remote sensing of environment with active and passive microwave sensors is clearly an important area requiring much further research and development. The potential for measuring sea state, geographic patterns of sea-surface roughness, windspeeds and directions, surface temperatures, surface salinities, ocean dynamic topography, soil moisture, terrain roughness, foliage density, and sea ice properties in all-weather daytime or nighttime conditions should be aggressively pursued. Development of data interpretation and use capabilities is severely hampered by difficulty of access to sensor systems and data. More widespread availability of aircraft missions in support of oceanographic, coastal, geologic, geographic, and meteorological research programs is essential.

Tracking of Balloons, Buoys, Floats, and Satellite Data Collection Platforms

A series of advances in Lagrangian object displacement measurements of winds and ocean currents has taken place. These advances have been made possible by new telemetry and data-link technology that has been combined with remote-sensing platforms. The first spacecraft experiment was the Interrogation Recording and Location System (IRLS) operated from NIMBUS 3 and 4 satellites [18]. It tracked about 30 stratospheric balloons [19]. The next advance was the French EOLE satellite, launched on August 16, 1971, by a Scout Rocket from the NASA Wallops Island facility. This operation culminated in the successful tracking of a network of 280 stratospheric balloons simultaneously [20]. A spar buoy with subsurface drogue was tracked by the EOLE satellite for 89 days in the Agulhas Current system, off South Africa. The buoy described a cyclonic eddy, a 140-degree eastward turn of the current, and an anticyclonic eddy [21]. Other experiments have been conducted using buoys, icebergs, animals, stations, and ships. The EOLE experiment provided for the first time a completely homogeneous set of highly accurate in-situ measurements of horizontal wind at nominal density level (200 mbar ± 1 percent) on a planetary scale. Of greatest importance from the EOLE experiment was the demonstration of a powerful new tool for measuring and understanding general circulation kinematics of both the atmosphere and the upper water column of the oceans. A new experiment of this kind entitled Tropical Wind Energy Conversion and Reference Level Experiment (TWERLE) is underway with the NIMBUS project [22], as noted in Table 1.

For smaller scale studies, ONR-sponsored research produced the over-the-horizon radio direction-finding system for tracking coastal and shelf currents [23]. This system, using simple, inexpensive, lightweight, and portable equipment, requires only one operator at each of two shorebased tracking stations. With it one can track the movements of as many as 15 drogues within 450 km of the shore stations. These are similar achievements to the development of the neutral-buoyance acoustically tracked subsurface oceanic floats such as the Swallow Floats [24] and the SOFAR Floats [25]. Review of these programs demonstrates capabilities of making Lagrangian measurements of atmospheric, sea-surface, and subsurface ocean circulation from planetary scales to microscales. These experiments rather impressively demonstrate the new tools available for environmental data support for military operations (a) from a real-time readout of measurements critical to operations or (b)

from numerical modeling of circulation with increased data base and new understanding of the kinematics of flow.

Satellite Data Collection Platforms (DCP) provide data-link capabilities for transmission from remotely located in-situ sensors. Originally LANDSAT, and now the Stationary Meteorological Satellites (SMS/GOES-geostationary satellites), provide this capability for all situations where the spacecraft is simultaneously in view of the remote data collection platform and the ground receiving site [26]. The satellite serves as a remotely positioned relay station. It is a direct throughput only, with no onboard storage capabilities. The data collection platforms are low in cost and can transmit data from a number of sensors. They can transmit at preselected intervals or through interrogation commands relayed through the Stationary Meteorological Satellite from the operations control center. The data collection platforms are completely self-contained and may be used on land or on oceanic platforms (moored or drifing buoys). Greatly expanded data message capabilities will be available from the commercial communications satellites. This capability, like the more expensive satellite tracking of balloons or buoys, provides access to measurements of the environment with minimal commitment of resources and with greatly extended duration.

Laser Sounding of the Ocean

Vertical laser sounding for temperature structure, thermocline depths, depth to bottom, or particulate concentration using pulsed lasers may be a most powerful capability for future development. Operating in the blue-green portion of the spectrum, lasers can provide maximum capability for penetration of the water column. Airborne utilization of lasers in this portion of the visible spectrum provides the capability for simultaneous measurement of the sea-surface wave profile and of bathymetry in the coastal regions. Further, analysis of the laser signal return as a function of depth within the water column will provide a measure of the turbidity within the column. This turbidity may be related to optical visibility and to littoral sediment transport. Further, coupling of the laser beam with the molecular resonance of the water molecules can stimulate Raman frequency shifts and polarization changes in the received energy which can provide temperature and salinity profiles within the water column.

Detecting and Measuring Properties of Soil and Rock

Studies are underway to develop a technique for rapid assessment of surface soil types and conditions (including moisture content and temperature) over large areas of the globe. Satellite-based remote-sensing techniques have been shown to satisfy the requirement for rapid global monitoring, but the particular approach that will yield the best soil data has yet to be determined. Microwave radiometers are sensitive to soil water content over depths on the order of a few centimeters, but to date their applicability has been severely limited by the coarse spatial resolution [27]. By using visible and infrared sensors, resolution can be improved, though depth of detection is reduced (to several millimeters) and weather conditions are limiting. Idso et al. in 1975 demonstrated that albedo is useful for characterizing surface soil water content of bare soils [28]. Measurements after heavy rain or irrigation show that soil water content from surface to a depth of 10 cm is well correlated with surface albedo. Two problems occur with this approach: correlation breaks down with light rainfall or irrigation, and albedos of various soils differ so greatly that soil type must be well known in advance. Idso et al. found that, if soil type is known, the thermal inertia, or amplitude of the diurnal surface soil temperature wave, can be used to yield good estimates of water content in the surface layer, as can the maximum value of the temperature differential between surface soil and air. Remote sensing of the diurnal temperature wave of the Earth's surface has also been used to detect bedrock types. Materials with low thermal inertia are relatively insensitive to temperature perturbations at the surface. At the Jet Propulsion Laboratory, a thermal model of the response of the earth's surface layer to diurnal heating has been developed. When the model is applied to aircraft or satellite measurements of surface albedo and midday and predawn temperatures, the thermal inertia of the surface can be inferred, because, thermal inertia is a body rather

than a surface property, the effect is measurement of bulk property of a surface soil layer or bedrock [29]. NASA's Heat Capacity Mapping Mission (HCMM), scheduled for launch in 1978, will measure albedo in the 0.5–1.1 μm range and surface temperatures in the 10–12 μm range. Spatial resolution will be 500 m and repetition rate will be approximately 8 days for postnoon, postmidnight revisit within one 24-h period. In addition to surface soil moisture content, measurement of the freezing isotherm on the surface of the earth is of importance to agriculture as well as to military operations. Measurement of cold surface temperature or migration of surface isotherms has been monitored using the Stationary Meteorological Satellite (Stephen Baig, personal communication, 1975). Even allowing the fact that IR emissivities are not 1.0 for the surface mosaic of soil types, plant canopy, and bedrock, it was possible to usefully predict the migration of the freezing isotherm for the benefit of Florida citrus growers.

In retrospect, essential to recall is that all the systems provide only numbers, i.e., variations in voltage. There is no interpretation of these spatial or time variations in sensor readings without concepts of the earth scene and its changes, concepts of electromagnetic radiation interacting with solid liquid and gaseous components of the scene, and concepts of how the sensor data distort the scene appearance. Over the past 20–30 years data interpretation concepts used have changed dramatically from the subjective classification and identification of features in aerial photographs to the use of precisely formulated quantitative concepts, numerical models, and statistical classification techniques for correcting, classifying, analyzing, and interpreting the data. Experience shows one fact to be clear, however: in research, the full range of approaches will coexist. Contemplative study of mapped data by specialists is the concept development stage and in many cases is the necessary forerunner of automated processing.

PROGNOSIS FOR THE FUTURE

The power of advanced remote-sensing technology for dealing with environmental problems is manifest. The monitoring and forecasting of the intervening natural environment (military environment reconnaissance) is a logical extension of and adjunct to military target reconnaissance, the monitoring of position, movements, and strengths of hostile forces. The fertile combination of the advanced technology with the geophysical environmental sciences has produced an impressive array of advances in an increasingly critical problem area for military and civilian sectors of society. The questions remaining here are, what scientific and technological directions to be pursued, and what can be expected from this field in the future?

Technology must improve platforms, sensors, and data links. Needed are greater spatial resolution, increased repetition rates, narrower spectral bands, (and more of them), changeable sensor gain, and all-weather day/night capabilities. In meteorology and oceanogaphy a major thrust is coming in microwave sensors and in all-weather and day/night sensing. *Increased control* of sensors is needed—to point, to change resolution ("zoom"), to change look angle, to vary the mix of spectral bands according to special user needs, and to escape the vagaries of natural illumination. Active sensors will play a major role in the future, but power limitations and antenna sizes demand new technological developments. Control of remote-sensing systems is presently by telemetry of commands and by man-in-space Skylab-type projects. With the space shuttle and orbiting space stations proliferating in the coming several decades, teams of scientists and engineers will jointly perform "field" work, operating sensors and conducting work from space. It is interesting to note that they can also make simultaneous ground-truth measurements, interrogating in-situ sensors that record, store, and telemeter direct measurements for analysis and comparison in space or at faraway ground stations. The results of discoveries will be nearly immediate inasmuch as near real-time data acquisition and processing allows the experiments and thought processes to continue without interruption. Tests of concepts in some cases will be continued in a series of different earth locations with remotely sensed data acquired farther along orbit or on a later orbit.

Further control of remote sensing is available through the use of remotely piloted vehicles. Here the sensor package can be moved closer or farther

from the target phenomenon. This allows maneuvering with respect to claud decks and use of active sensors of lower power. Drone aircraft can loiter for days at high altitudes, awaiting conditions or commands to acquire the desired data. Artificial illumination of the earth scene with varying portions of the electromagnetic spectrum multiplies enormously the measurements possible.

Of major importance is the development of automated in-situ measurement technology, coupled with telemetry. The concept here is that automated sensors and data links create an artificial "nervous system," extending man's perception capabilities to the remotest, most inclement, and most inhospitable regions of the earth. New measurements and information on extreme conditions and on the earth will be increasingly possible. Improved sampling and on-scene processing and telemetry can increase capabilities by transmitting required information rather than raw data.

An expansion of application of remote-sensing technology awaits the low-cost distribution of high-quality data to operational decision makers. The DMSP remote vans and NOAA APT capabilities have demonstrated the appetite that field groups have for high-quality data. Necessary are improved capacity of portable remote receivers, antennas, minicomputers, and display devices. Required is low-cost access to a larger variety of data for small military field units, university field camps, commercial operators, etc.—all who make important practical decisions on the basis of environmental prognoses and conditions.

The environmental disciplines are evolving rapidly, experiencing expansion of problems to solve, information available, and technological power. Far more knowledge is needed on the reflective and emissive characteristics of the ocean atmosphere and terrain, through an increasingly wider range of the electromagnetic spectrum. A shift from measurement of parameters to measurement of fluxes is underway. It will include measurement of fluxes of energy and matter through complex ecosystems with interactions at many levels of scale. Research with the increasing array of tools will uncover new information on the temporal and spatial structures of processes, revealing regularities of pattern extending over large geographic domains. Already studies of the relationships of ocean and atmospheric events over long distances, called "teleconnections" have been investigated by several workers. Examples include research on the influence of northern hemisphere circulation on droughts in Brazil [30], the influence of strong flow in the equatorial countercurrent on the occurrence of El Nino [31], and the teleconnections among the Aleutian Low, the westerlies, the trade winds, and convective activity near the equator. These examples are environmental events structured in time and space—expansive cause-and-effect chains. The study and modeling of these system interactions will add new strength to environmental prediction. Other teleconnections can be envisioned, and remote sensing is a key to detection and forecasting. For example, early snowmelt seen by satellite can be used to initiate a model run duplicating the rate of movement of spring runoff and its impact on the entire hydrologic basin, river mouth, and coastal oceanic region.

Much inference is necessary to deduce environmental conditions. Researchers must work intensely to improve interpretation capabilities. One optimum use of remotely sensed data is for tuning, correction, and monitoring the divergence of numerical models from reality. Models are mandatory to provide quantitative interpretation of conditions detected, and are needed for quantitative extrapolation and interpolation between observation periods and spaces. Improved models of four classes are needed:

1. Atmospheric transmission and attenuation models for the increasingly greater range of spectral regions of interest
2. Models of variations in reflected and emitted brightness temperatures and surface scattering processes
3. Numerical fluid dynamical models of atmospheric and hydrospheric processes
4. Models of morphodynamic processes that shape the sediment accumulations and bedrock surfaces of the solid earth.

Vital to the success of the modeling in remote sensing is keeping the efforts very closely linked to the field investigations, from inception to testing. Increased or optimized realism requires that these efforts occur within environmental research programs rather than in programs strictly oriented toward numerical methods.

Future research directions call for combined multidisciplinary studies using multitechnological approaches—for instance, meteorologicaloceanographic studies, meteorologicalhydrogrological studies, studies of surface reflectivity-emissivity with soils and geology, studies of heat capacity-thermal inertia with geology and soils. Imaginative and unprecedented aggregates of technology will appear spontaneously, such as electro-optical remote sensing combined with acoustic surveillance, or in-situ sensor telemetry with seismic monitoring of environment. These are, of course, high-cost efforts, and increased competition and more stringent cost-benefits analyses will be necessary.

In the future, military analysts will receive much more accurate, extensive, and responsive environmental intelligence than ever in the past. Near real-time inputs on conditions of the sea, atmosphere, and terrain will rapidly increase user confidence in the information, and it will play an increasingly important role in the strategic and tactical decisionmaking process. For example, task forces will receive detailed information on the surface temperature structure, sea-surface roughness, water levels, currents, boundaries, sea ice, and atmospheric conditions of a target objective region. From these, models will generate specific sonar conditions, optimum ship transit speeds, and radar detection conditions. Amphibious and special forces will receive near real-time instead of historical information on near-shore conditions such as surf zone width, wave height, locations of bars and shoals, beach trafficability, and conditions on the route to the amphibious objective. Detailed environmental data of adequate resolution will come in large quantities and will need automatic reduction and, most critically, some measure of reliability. Military commanders will not weigh environmental data of spotty reliability in the face of hard military intelligence in the making of important decisions.

For more than a third of a century, the Navy has been a driving force and technological leader in oceanographic research in the United States. In the last 15 years, the Office of Naval Research has consistently supported basic research stressing use of remote-sensing systems. Navy requirements for the measurement, analysis, and prediction capabilities needed to support military missions demand that high-resolution, high-accuracy, and high-reliability environmental prediction products be available to field units operating anywhere on the globe. Knowledge of the physics of oceanographic, atmospheric, and geomorphic processes is necessary to the construction of reliable operational prediction models. High-accuracy measurement technology is necessary to support research efforts that define the basic environmental physics and that sustain subsequent prediction models. For the Navy, remote sensing of environment will play an increasingly vital supporting role. Of all the services, it is the Navy that launches operations in all environments of the globe—in the air, on the sea, under the sea, on coastal regions, and across the ice and snow of the poles.

Table 1

Satellite Systems, Achievements and Plans

Satellite Launch Date	*Achievements*	*Satellite*
Apr. 1, 1960	Daytime cloud cover photography from space	TIROS I
Dec. 21, 1963	Direct readout of cloud pictures to local ground stations	TIROS VIII
Aug. 28, 1964	Nighttime cloud cover imagery	NIMBUS I
Jan. 21, 1965	Global daytime cloud cover photography in Sun-synchroous orbit	TIROS IX
July 1, 1965	First operational satellite	TIROS X

Table 1 (continued)

Satellite Launch Date	*Achievements*	*Satellite*
Feb. 28, 1966	Inauguration of world's first operational satellite system	ESSA 1 and 2
Dec. 7, 1966	Continuous black-and-white cloud cover pictures from geosynchronous orbit	ATS 1
Nov. 5, 1967	Continuous color cloud cover pictures from geosynchronous orbit	ATS 3
Apr. 14, 1969	Vertical Atmospheric Temperature Sounder	NIMBUS 3
Jan. 17, 1970	Operational satellite with scanning radiometer (daytime and nighttime coverage)	ITOS 1
Aug. 16, 1971	Tracking and data collection from a large fleet of balloons or buoys (France)	EOLE
July 23, 1972	Operational multispectral scanner with 80-m resolution, 185-m-square field of view, repetition rate every 18 days. Sensor response in four channels: green-yellow, orange-red, red-near infrared, and infrared bands. Satellite data collection system for relaying data transmitted from in-situ data collection platforms dispersed around the United States.	ERTS-A
Oct. 15, 1972	Operational satellite with very high resolution radiometer and vertical temperature profile radiometer.	NOAA 2
Dec. 12, 1972	Microwave spectrometer and electrically scanning microwave radiometer for vertical temperature profiles and sea ice boundaries through clouds.	NIMBUS 5
Mar. 7, 1973	DMSP data and capabilities made public by the U.S. Air Force with original system designation *Data Acquisition and Processing Program* (DAPP). First operational system with two polar orbiting satellites and a nighttime high-gain, visual range earth-imaging capability for city lights and the aurora borealis.	Block 5-C
May 1973	Manned orbital mission with *Earth Resources Experiment Package* (REP) included a 6-band multispectral earth terrain camera, infrared spectrometer, 13-band multispectral scanner, microwave radiometer/scatterometer and altimeter, and L-band radiometer.	Skylab
May 17, 1974	First geosynchronous operational environmental satellite with visual and infrared spin-scan radiometer (VISSR).	SMS-1
Mar. 10, 1975	Inauguration of two-satellite system for near-continuous viewing of United States and adjacent waters.	SMS-2
June 12, 1975	Continuation of previous NIMBUS experiments including the *Temperature Humidity Infrared Radiometer*	NIMBUS 6

Table 1 (continued)

Satellite Launch Date	*Achievements*	*Satellite*
	(THIR), a two-channel scanning infrared radiometer with an 11.5μm channel for images of cloud cover, temperatures of cloud tops, land and sea surfaces (8.2-km spatial resolution) and a 6.7-μm channel for upper troposphere and stratosphere moisture and location of jet streams/frontal systems (22-km spatial resolution). The *Electrically Scanning Microwave Radiometer* (ESMR) experiment, a single-channel (250-MHz band centered at 37 GHz) electrically scanning radiometer that measures thermal microwave radiation upwelling from the Earth's surface and atmosphere. It is used for mapping liquid water content of clouds, distribution and variation of sea ice and snow cover on ice, and characteristics of land surfaces (spatial resolution 25 × 25 km at nadir to 160 × 45 km at extremity of scan). Complete global coverage 12 h. Other new experiments, including: The *Earth Radiation Budget* (ERB) experiment, which involves a 22-channel radiometer viewing Earth and Sun, to provide highly accurate (to 1% or less) radiation measurements of sun and earth for computation of radiation budget at synoptic and planetary sclaes. The *High Resolution Infrared Radiation Sounder* (HIRS) experiment, a third-generation sounding experiment using a 17-channel radiometer for obtaining surface temperature, vertical atmospheric temperature profile, vertical humidity profile, integrated water content of clouds, surface albedo, average total albedo, total outgoing longwave flux, and pressure altitude and amount of clouds. Maximum resolution 25 km. The *Scanning Microwave Spectrometer* (SCAMS) is a five-channel radiometer for producing global maps of troposphere temperature profiles, liquid water and water vapor in the atmosphere, snow cover, ice type, soil moisture, and ocean roughness. Spatial resolution ranges from 145 km at nadir to 330 km at scan margin. The *Limb Radiance Inversion Radiometer* (LRIR), a four-channel multispectral scanning radiometer to measure vertical distribution of temperature, ozone, and water vapor by inverting the limb radiance profiles obtained from scanning the earth's horizon. The *Pressure Modulated Radiometer* (PMR) experiment includes a two-channel radiometer with pressure modulated transmission of radiance through gas-filled cells to sensor. It measures atmospheric temperature distribution in the upper stratosphere and mesosphere (between 40 and 85 km altitude) by selected radiation emitted by CO_2 emission. The frequency component of atmospheric radiation in phase	

Table 1 (continued)

Satellite Launch Date	*Achievements*	*Satellite*
	with cell gas modulation is measured by the detector. Vertical resolution 10 km at nadir, horizontal resolution 500 km. *The Tropical Wind Energy Conversion and Reference Level Experiment* (TWERLE) is a meteorological observation system using lightweight, low-cost balloons to record temperature, pressure, geometric altitude, and location, transmit to NIMBUS for relay to ground for processing. Over 500 platforms operating with location accuracy to reference sites of 1.5 km of true position.	
	Plans	
1976	DMSP Block 5-D satellites, first to achieve constant cross-track spatial resolution of scanner data for automated data processing and accurate Earth location of data. Incorporates an advanced atmospheric sounder for temperature and humidity profiles and total ozone. A highly accurate attitude determination and control system is used for precise pointing of imaging sensor. Twin (redundant) digital computers are on board, programable by message to control satellite functions.	Block 5-D
Sep. 1977	Inauguration of new thermal channel in the *LANDSAT Multispectral Scanner*, providing 240-m spatial resolution infrared imagery in 10.4-12.6 μm spectral interval. Improvement of Return Beam Vidicon system for high resolution panchromatic Earth images. Conversion to all-digital processing for increased production. Initiation of Cubic Convolution method for geometric correlation of LANDSAT video data, a very high quality data interpolation technique.	LANDSAT-C
Late 1977 or Early 1978	The *Heat Capacity Mapping Mission* for study of the thermal inertia of Earth materials to differentiate surface materials and identify conditions such as soil moisture content. Uses a single sensor, the Heat Capacity Mapping Radiometer, with two channels, 0.5–1.1 μm, and 10.5–12.5 μm, 500-m spatial resolution, and a repeat time of 1–3 days. A small dedicated satellite, the Applications Explorer Mission-A.	AEM-A
Early 1978	Start of new series of NOAA series polar-orbiting satellites, includes the *Advanced Very High Resolution Radiometer* (AVHRR), the *TIROS Operational Vertical Sounder* (TOVS), and the new *Data Collection System* (DCS). The AVHRR has 1.1-km spatial resolution and four channels, 0.55–0.90 μm, 0.725–1.0 μm, 3.55–3.93 μm, 10.5–11.5μm, with digital data downlink. The	TIROS-N

Table 1 (continued)

Satellite Launch Date	*Plans*	*Satellite*
	TOVS will provide vertical atmospheric temperature profiles, water vapor amounts at three levels, and total ozone content of the atmosphere. The new data collection system will monitor nearly 2000 data collection platforms around the globe.	
May 1968	Launch of first dedicated oceanographic satellite. This experiment includes five sensors: The *Radar Altimeter* for measurement of detailed shape of the marine geoid as influenced by ocean currents, storm surges, and tides. *Wind Field Scatterometer* to measure surface windspeed and direction on a global scale for evaluation of potential impact on numerical wave forecasting models. *Synthetic-Aperture Radar* to obtain ocean-surface imagery for directional wave spectra, monitoring of coastal processes, charting of icebergs, icefields, and leads. *Visual and Infrared Imaging Radiometer* will provide feature recognition and cloud position data, clear air sea-surface temperatures, and cloud top brightness temperatures to supplement microwave experiments.	SEASAT-A
Late 1978	NIMBUS G experiments include: the *Coastal Zone Color Scanner* (CZCS), a six-channel scanning radiometer to detect water color for chlorophyll, sediment and gelbstoffe (yellow humic compounds) content, and surface temperature. The *Scanning Multichannel Microwave Radiometer* (SMMR), a five-channel microwave radiometer for mapping sea ice, continental ice sheets, heavy weather patterns, atmospheric water (liquid and vapor), sea-surfce winds, sea-surface temperature, and soil moisture on a nearly all-weather basis, 33–245 km spatial resolution. The *Solar and Backscattered Ultraviolet* (SBUV) and *Total Ozone Mapping System* (TOMS), measuring the time variability of solar spectral irradiance and atmospheric backscatter and the total ozone field. 50-km spatial resolution with vertical distribution to 55-km altitude along nadir track. *Limb Infrared Monitor of the Stratosphere* (LIMS), a six-channel infrared radiometer to map vertical profiles of temperature and concentrates of O_3, H_2O, NO_2, and HNO_3. The *Stratospheric Aerosol Measurement* (SAM II), a single-channel solar photometer that measures the extent of solar radiation at spacecraft sunrise and sunset. It will map concentrations of submicron stratospheric aerosols as a function of altitude with supplementary ground-truth LIDAR and in-situ balloon-borne aerosol measurements. *Measurement of Air Pollution from*	NIMBUS G

Table 1 (continued)

Satellite Launch Date	*Plans*	*Satellite*
	Satellites (MAPS), a three-channel nadir-looking radiometer to map global distribution of the total integrated CO CH_4, and NH_3 levels in the trophosphere. NIMBUS G will also continue previously run experiments, including the *Temperature Humidity Infrared Radiometer* (THIR) and the *Earth Radiation Budget* (ERB).	
	A planned geosynchronous satellite with greatly improved spatial and temporal resolution and radiometric sensitivities over present SIMS series. It will scan small areas of known severe storm activity and mesoscale phenomena of interest. It will include advanced atmospheric temperature and moisture profilers.	STORMSAT-A
1981	*Remote Ocean Measurement System* (ROMS), a DOD active and passive microwave sensor suite for ocean-surface measurements.	DMSP
1985	*Synchronous Earth Observatory Satellite* (SEOS), an advanced geostationary satellite with mission assignments in mesoscale atmospheric phenomena and Earth Resources Observations. Plans include an advanced multispectral imagery (including microwave) and an advanced IR and microwave atmospheric sounder.	SEOS A

REFERENCES

1. G. J. Zissis, "The Development of Remote Sensing of Earth Resources," Proc., Comm. on Science and Astronautics, House of Representatives, 92d Congress, Rep. 13, 1972.
2. United States Department of Commerce, *The Federal Plan for Meteorological Services and Supporting Research, Fiscal Year 1976*, National Oceanic and Atmospheric Administration, Washington, D.C., 1975, p. 79.
3. R. Reeves, A. Anson, and D Landen, *Manual of Remote Sensing:* vol. I, "Theory, Instruments and Techniques"; vol. II, "Interpretation and Applications," American Society of Photogrammetry, Falls Church, Va. 1975, 2144 pp.
4. O. K. Huh, "Coastal Oceanographic Use of the Defense Meteorological Satellite Program (DMSP)," U.S. Naval Oceanographic Office, Tech. Rep. 241, 52 pp, 1973.
5. O. K. Huh, "Detection of Oceanic Thermal Fronts off Korea with the Defense Meteorological Satellites," *Remote Sensing of Environment*, 1976 (in preparation).
6. P. E. Laviolette, L. Stuart, Jr., and C. Vermillion, "Use of APT Satellite Infrared Data in Oceanographic Survey Operating," *Trans. Am. Geophys. Union* **46**(5), 276–282 (1975).
7. W. L. Smith, "Satellite Techniques for Observing the Temperature Structure of the Atmosphere," *Bull. Am. Meteorol. Soc* **53**(11), 1074–1082 (1972).
8. L. M. McMillin et al., "Satellite Infrared Soundings from NOAA Spacecraft," NOAA Tech. Rep. NESS 65, 112 pp., 1973.

9. S. Fritz et al., "Temperature Sounding from Satellites," NOAA Tech. Rep. NESS 59, 49 pp., 1972.
10. V. Klemas, J.F. Borchardt, and W. M. Treasure, "Suspended Sediment Observations from ERTS-1," *Remote Sensing of Environment* **2**, 205–221 (1973).
11. G. L. Clarke, G. C. Ewing, and C. J. Lorenzen, "Spectra of Backscattered Light from the Sea Obtained from Aircraft as a Measure of Chlorophyll Concentration," *Science* **167**, 1119–1121 (1970).
12. G. Maul and H. Gordon, "On the Use of the Earth Resources Technology Satellite (LANDSAT-1) in Optical Oceanography," *Remote Sensing of Environment* **4**, 95–128 (1975).
13. W. Hovis, M. Forman, and L. Blaine, *Detection of Ocean Color Changes from High Altitudes*, Goddard Space Flight Center, Greenbelt, Md., 1973, pp. 1–23.
14. J. R. Apel et al., "Observations of Oceanic Internal and Surface Waves from the Earth Resources Technology Satellite," *J. Geophys. Res.* **80**(6), 865–881 (1975).
15. Moskowitz, L. I., "The Feasibility of Ocean Current Mapping via Synthetic Aperture Radar Methods," *Proc. Am. Soc. Photogrammetry*, Fall Convention, Walt Disney World, Lake Buena Vista, Fla. Oct. 2–5, 1973.
16. Hollinger, J. P., Robert M. Lerner, and MacMillan M. Wisler, "An Investigation of the Remote Determination of Sea Surface Temperature Using Microwave Radiometry," NRL Memorandum Report 3159, Nov. 1975.
17. W. J. Campbell et al., "Beaufort Sea Ice Zones as Delineated by Microwave Imagery," *J. Geophys, Res.* **81**(6), 1103–1110 (1976).
18. C. E. Cote, "The Interrogation, Recording and Location System," *IEEE Trans. Geosci. Electron.* **8**, 243–245 (1970).
19. J. K. Angell, "Air Motions in the Tropical Stratosphere Deduced from Satellite Tracking of Horizontally Floating Balloons," *J. Atmos. Sci.* **29**, 570–582 (1972).
20. P. Morel and W. Bandeen, "The EOLE Experiment: Early Results and Curent Objectives," *Bull. Amer., Meteorol. Soc.* **54**, 298–306 (1973).
21. C. C. Stavropoulos and C. P. Duncan, "A Satellite Tracked Buoy in the Agulhas Current," *J. Geophys. Res.* **79**(18), 2744–2746 (1974).
22. J. E. Masterson, "A Random Doppler Measurement Technique for the Global Atmospheric Research Program," *Bull. Amer. Meteorol. Soc.* **51**, 222–226 (1970).
23. S. P. Murray et al., "An Over-the-Horizon Radio Direction-Finding System for Tracking Coastal and Shelf Currents," *Geophys. Res. Lett.* **2**(6), 211–214 (1975).
24. J. C. Swallow, "A Neutral-Buoyancy Float for Measuring Deep Range Currents," *Deep Sea Res.* **3**, 74–81 (1955).
25. T. Rossby, A. D. Voorhis, and D. Webb, "A Quasi-Lagrangian Study of Mid-Ocean Variability Using Long Range SOFAR Floods," *J. Marine Res.* **33**(3) 1975).
26. National Atmospheric and Space Administration, *Data Users' Handbook: Earth Resources Technology Satellite*, Goddard Space Flight Center, Greenbelt, Md., 1971, 200 pp.
27. T. Schmugge et al., "Remote Sensing of Soil Moisture with Microwave Radiometers," *J. Geophys. Res.* **79**(2), 317–323 (1974).
28. S. B. Idso et al., "The Utility of Surface Temperature Measurements for the Remote Sensing of Surface Soil Water Status," *J. Geophys. Res.* **80**(21), 3044–3049 (1975).
29. A. B. Kahle, et al., "Thermal Inertia Mapping," *10th Internat. Symp. on Remote Sensing of Environment: Summaries*, Ann Arbor, Mich., p. 142 (1975).
30. J. Namias, "Large-scale and Long-term Fluctuations in Some Atmospheric and Oceanic Variables," Scripps Institution of Oceanography, La Jolla, Cal., 1972.
31. K. Wyrtki, "Teleconnections in the Equatorial Pacific Ocean," *Science* **180**, 66–68 (1973).



John G. Heacock is the director of the Earth Physics Program of the Office of Naval Research. He worked as a geophysicist for the Shell Oil Company from 1953 until he joined ONR in 1962. His work at ONR has included studies of Earth and ocean tides and their interaction, of the rotational parameters of the earth's axis, of geodetic positioning aimed at measuring continental drift, of the physical properties of the Earth, including field and laboratory studies of the physical properties of the Earth's crust, of the detection of hostile weapons by seismic means, and of the use of geothermal energy to power remote naval bases. He is editor of the AGU Geophysical Monograph 14 on *The Structure and Physical Properties of the Earth's Crust*. His current efforts involve the application of broad geological and geophysical techniques to solve problems of direct interest to the Navy. He received his undergraduate training in physics at Franklin and Marshall College and his graduate training in physics and geophysics at Columbia University.



Jack E. Oliver is Chairman of the department of Geological Sciences at Cornell University. From 1953 until he joined the faculty at Cornell, Dr. Oliver held various positions with Columbia University, including that of Chairman, Section of Seismology of Lamont Geological Observatory, Professor of Geology (1961-1971), and Chairman, Department of Geology (1969-1971). His research work has included exploration of the upper atmosphere by acoustical methods; participation in the first U.S. aircraft landings for scientific purposes on the Arctic ice pack; marine seismic refraction measurements; analysis of long-period seismic data from Columbia University's worldwide seismograph network; study of Rayleigh wave phase velocities; and studies of strainmeter data, crustal movements from leveling, source mechanisms of seismic waves, the new global tectonics, and broad aspects of seismology. Dr. Oliver earned a B.A. at Columbia College and an M.A. in Physics and a Ph.D. in Geophysics at Columbia University. He served in the U.S. Naval Reserve from 1943 to 1946. He is a Fellow of the Geological Society of America and of the American Geophysical Union. He is past president of the Seismological Society of America; was Councilor of the Geological Society of America and President of the Section on Seismology of the American Geophysical Union; and has filled posts on numerous advisory committees of national and international stature. He is Chairman (1976-1979) of the Office of Earth Science of the National Research Council.



George V. Keller has been with the Department of Geophysics at the Colorado School of Mines since 1964; he is currently Professor and Department Head. He served as a geophysicist with the U.S. Geological Survey from 1952 to 1964. Dr. Keller has conducted research on exploration for geothermal energy, the development of electrical prospecting methods, and data analysis associated with electrical probing of the Earth's crust. He earned a B.S., M.S., and Ph.D. in Geophysics at Pennsylvania State University.



Gene Simmons is Professor of Geophysics at the Massachusetts Institute of Technology. He served for two years as Chief Scientist of NASA's Manned Spacecraft Center in Houston during the Apollo Program. His research contributions have been in geophysics and have included data on the physical properties of rocks and minerals, measurement and interpretation of terrestrial heat flow beneath continents and the oceans, measurement and interpretation of the Earth's gravitational field in the Adirondack region, and a surface experiment done on the Moon during Apollo 17. His current research emphasizes studies of the controls exerted by microcracks on the physical properties of rocks as detailed in the laboratory and applied to the analysis of field data. Dr. Simmons received undergraduate training in electrical engineering at Texas A&M and graduate training in geology (M.S.) at Southern Methodist University and in geophysics (Ph.D.) at Harvard University.

SOLID EARTH PROPERTIES AND THEIR IMPORTANCE TO THE NAVY: CURRENT KNOWLEDGE AND FUTURE PROSPECTS

John G. Heacock

Office of Naval Research
Arlington, Va.

Jack E. Oliver

Cornell University
Ithaca, N.Y.

George V. Keller

Colorado School of Mines
Golden, Colo.

Gene Simmons

Massachusetts Institute of Technology
Cambridge, Mass.

ABSTRACT: This paper discusses recent progress in understanding the relations among various physical properties of the Earth's crust (e.g., seismic vs strength properties); advances in measuring techniques for evaluating the electrical properties of the crust, supporting laboratory studies of the influence of microfractures on the seismic, electrical and other physical properties of rocks, and advances in seismology. Such research is leading to new insight into the physical properties of deep Earth materials and a new capability for inferring quantitively properties that are otherwise unmeasurable, such as crustal temperature, lithology, strength, porosity, electromagnetic propagation characteristics, and state of stress at various depths. These studies are important to the Navy as they relate to the development of geothermal energy, the engineering strength of underground structures, communication through the Earth, and the evaluation of earthquake risk, to name some of the more obvious possibilities. As a result of the anticipated advances, it appears likely that within the next 10 to 20 years (or sooner) certain naval bases will be operating on local geothermal energy sources; that the risk of earthquake damage will be assessable in order that a rational judgement can be made on the extent of protective measures which should be taken to protect a particular naval base and the associated civilian community; that the resistance of subterranean caverns to surface overpressures will be computable; and that the utility of the earth's solid body as a communication medium will have been evaluated. While the above research is related primarily to the physical properties of the Earth's crust, additional research related to whole earth properties are also important to the Navy, as for example, the influence of the elastic behavior of the solid earth as it strongly affects the phase and amplitude of ocean tides. Such information is essential for predicting tides and tidally induced currents both in the deep sea and in coastal areas where no tide gauges are available and where it is presently not possible to predict tides.

The solid body of the Earth affects naval operations in many important ways. To understand this, let us consider the schematic figure of the Earth shown in Figure 1.

Earthquakes and Geothermal Energy

We observe that the Earth's interior is very hot (8600F° at the core). The Earth's internal heat produces convection currents amd stresses that ultimately cause earthquakes. Furthermore, the internal heat of the Earth is responsible for both volcanic and geothermal activity, the latter in the form of hot springs, geysers, or simply areas of high thermal gradients with a potential for producing usable energy in the form of heat.

Just these thermal aspects of the Earth have their own importance for the Navy. In the first

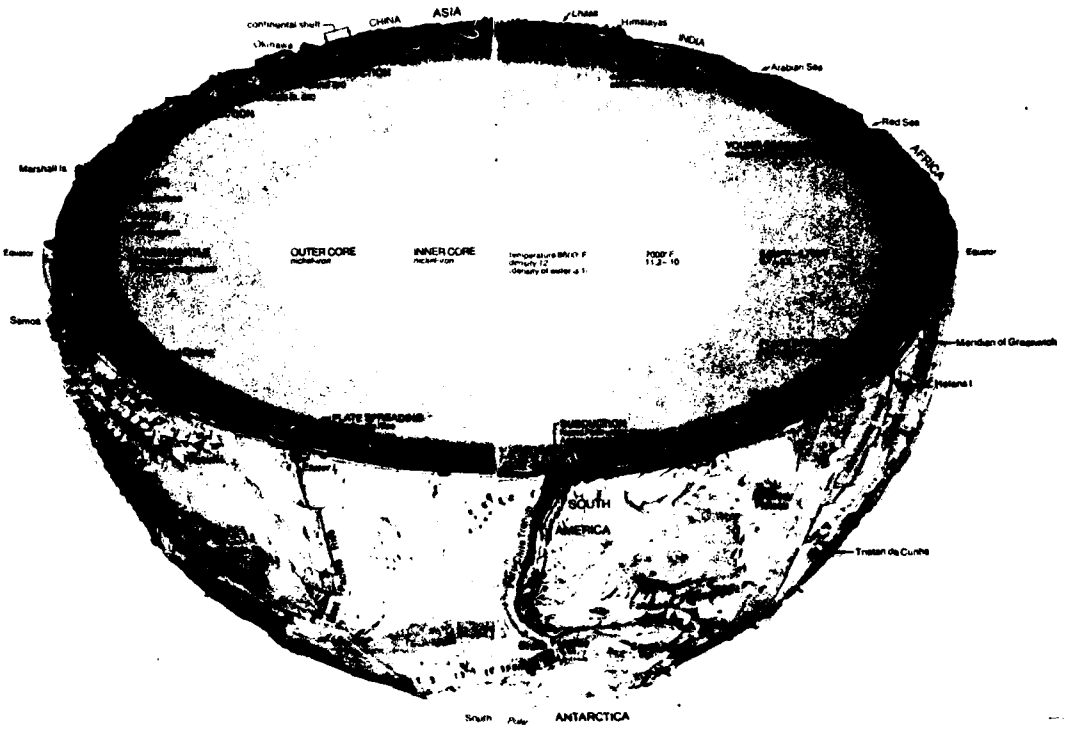


Figure 1—It should be noted that the thickness of the Earth's crust is exaggerated in this illustration to show its structural features. © Smithsonian Institution 1974; from Smithsonian *magazine, January 1975. Illustrator, Richard Edes Harrison. Used by permission.*

place, earthquakes are a threat to naval bases in seismically active zones; the bases are especially vulnerable because they are of necessity built on marine soils or even on filled land. Such areas are notoriously unstable and are subject to liquefaction (causing buildings and foundations to fail) when shaken by earthquakes. Thus, an ability to reduce earthquake risk is highly important for naval bases that are threatened. In the second place, geothermal resources close to naval bases, especially those in remote areas, offer a potential source of energy to operate those bases. The use of geothermal energy to operate remote naval bases offers an increasingly attractive alternative to the use of fossil fuels, given the continuing need to import increasing amounts of foreign fuels at elevated prices. By making its bases self-supporting in energy, the Navy can not only reduce its operating expenses and save precious hydrocarbon fuels for operating ships and aircraft, but in wartime it can relieve the need to provide escort vessels and the necessary manpower normally required to insure that fuels intended for base operations reach their destination. One day it may be possible to operate naval vessels and aircraft with synthetic fuels produced from geothermally generated electricity, although no practical technique exists for the production of such fuels today.

Earth Rotation

We note that the Earth rotates about an axis whose orientation defines the north-south direction. However, what is not so obvious is that although the rotational pole remains reasonably fixed in space, the surface of the Earth moves about the polar direction. This so-called Chandler Wobble has a 13-month periodicity and an elliptical pattern with an amplitude of roughly 20 by 30 meters in the minor and major axis directions, respectively. Thus, the true north-south direction

changes to this extent. Also, when measured astronomically, changes occur in the rate of Earth rotation. Both types of changes in Earth rotation are important, especially for reasons of ensuring missile accuracy.

Tides and Earth Elasticity

Let us recognize that the elastic behavior of the Earth greatly affects the Navy. The solid body of the Earth flexes tidally due to forces exerted by the Moon and the Sun. In addition, the Earth bends under the weight of shifting masses of tidally displaced water. The latter motion is known as the ocean loading effect. The elastic response of the Earth to these two effects has a first-order influence in controlling both the phase and amplitude of ocean tides [1].

Communication Potential

Note that the solid Earth is potentially a medium for either electromagnetic or seismic communication. Radio communication today is possible because of the earth-ionosphere waveguide formed by the electrically conducting ionosphere consisting of ionized gases surrounding the Earth, by the Earth's surface which is also conducting, and by the nonconducting (highly resistive) atmosphere between. Electromagnetic energy is reflected back and forth between the ionosphere and the Earth's surface and thus is confined to the so-called Earth-ionosphere waveguide, where it propagates with low attenuation because of the nonconductive nature of the atmosphere. A similar possibility has been suggested for the propagation of electromagnetic energy through the outer regions of the Earth, where the Earth's surface can act as the upper reflecting region and the Earth's hot, conductive interior can behave as the lower reflecting boundary. Between these two reflecting regions may be a zone of nonconducting (resistive) rock. Such a relationship, by analogy to the earth-ionosphere waveguide, may form what has been called the lithospheric waveguide. For the lithospheric waveguide to exist, the intermediate zone of resistive rock must be both continuous and sufficiently resistive that electromagnetic energy can propogate through it with low dissipation. As yet, we have no final answer to this intriguing possibility. Similarly, seismic energy propagates through the body of the Earth by various paths and different modes (e.g. body waves or surface waves), each with its own characteristics of frequency type of particle motion and amplitude distribution with depth. Seismic communication, while less attractive because of its lower data rate and limited range, might prove useful for some naval uses.

THE EARTH'S CRUST

Since the crust of the Earth is closest to us, it is clearly a region of great potential interaction with naval systems. Referring again to Figure 1, note the arrows in the vicinity of the East Pacific Rise, indicating the spreading sea floor in that region. Note also the subduction zone where the oceanic lithosphere plunges beneath the continents, shown in Figure 1 on either side of the Pacific Ocean (South America to the east and Okinawa, Japan, etc. to the west).

The subduction zone is a zone of stress concentration. Both heat and earthquake activity are characteristic of this region, in addition to the oceanic trench created by the tectonic forces at work. In the vicinity of the subduction zone stresses develop that can cause disastrous earthquakes. It is here that we can extract useful geothermal energy (as in a dozen countries around the world—including the U.S., which currently produces 500 MW from the Geysers area, some 90 mi (145 km) north of San Francisco). It is on the crust that we depend for our natural resources, and because of impending shortages both in fuels and critical minerals the Navy is strongly interested in understanding how the crust was formed in order to determine the most likely locations for resources in the future when today's more readily accessible fuels and minerals have been exhausted. This is a matter of critical importance for the continuation of modern civilization as we know it today, and as such is a matter of direct importance to the Navy. For these reasons, and to maintain a reasonable length for this paper, we shall limit its scope to a discussion primarily of the crust of the Earth and to a description of three of

the geophysical techniques that are important to the Navy for crustal studies.

Deep Crustal Studies—The New Frontier

The crust of the continents is perhaps the major frontier of the solid Earth sciences today. Within the next two or three decades we expect a rapid gain in our knowledge of this important part of the Earth and, as a consequence, major steplike advances in our understanding of the Earth and its history.

To understand why the above statements are likely to prove true, let us look at Earth sciences in the broad perspective of history as it relates to major advances in geology.

Geology (which in its broadest sense includes all studies of the solid Earth) advances at an irregular pace, as do all sciences. Commonly, periods of slow advance and gradual accumulation of observation are interrupted by intervals of discovery, synthesis, and rapid advance in our understanding of Earth phenomena. Sometimes the entire field is caught up in such cycles, sometimes only subdisciplines. But it is nearly always new observations that are the basis for such rapid developments in modern Earth science. Consider some major developments of the past.

Hutton, the founder of modern geology, showed that "the present is the key to the past" through extensive observation of sedimentary rocks, modern depositional, igneous, and other processes. The famous Neptunist-Plutonist debate of the late 1770s (on whether basalts and granites are of aqueous or igneous origin) was ultimately resolved by the "go-and-see" attitude of Demarest. This experimental attitude led to an appreciation of the immense importance of igneous activity in the development of the Earth. Likewise, understanding the pronounced landshaping effects of the Pleistocene icecaps and the importance of glaciation in Earth history resulted from observations of modern glaciers. Most recently the concept of plate tectonics or sea-floor spreading [2-5] has produced a revolution in the Earth sciences. This revolution depended vitally on the exploration of deep ocean floors that followed World War II, in which the Office of Naval Research played a major role [6-8].

Given these examples of steplike advances in our knowledge of the Earth, where can we look for the next series of major advances in the solid Earth sciences? The answer seems clear; it is in the deep basement rocks of the continents—and for several reasons.

First, such rocks are widespread and cover a large fraction of the Earth; to understand them is important for this reason alone. Second, deep crustal rocks are intimately related to surface rocks, on which man depends for his livelihood. Hence, it is critically important to understand this interaction more effectively in order to solve problems related to the distribution of minerals and other natural resources critical to the welfare of civilization. Third, the deep rocks are largely unknown in a detailed sense, and it is important to understand this region of the Earth thoroughly, not only to benefit from the potential resources hidden there but also to benefit from a knowledge of the interaction of this region on surface rocks in terms of Earth stress, tectonic activity, volcanic eruptions, etc. Fourth, modern technology has just now reached a stage where new tools (especially new seismic, electrical, and laboratory techniques) are available for exploration of the deep crust. This modern situation is analogous to that which occurred just after World War II for the exploration of ocean basins. Many new instruments (hydrophones, seismic recording equipment, precision echo sounders, gravimeters, magnetometers, etc.) and techniques were developed during wartime and were ready for adaption to the new scientific study of the ocean basins. These four reasons, plus the substantial recent stepwise advances in our knowledge of basic Earth structure and processes, stimulated by the concept of plate tectonics, clearly designate the deep crust as a major frontier of the Earth sciences at the present time.

The deep crust has already been partially explored in some places. This work has been helpful, but the information is limited because of the sparse and erratic application and the inherently low resolving power of the methods used. Compare, for example, the fine detail given by a geologic map of the surface rocks of an area with the crude crustal models generated to satisfy gravity data, magnetic data, or various forms of seismic data. Clearly, low resolution of standard

geophysical methods is a major obstacle to an improved understanding of the Earth. We must try to enhance geophysical methods for observing the deep crust, and there is good potential for substantial improvement in the near future.

SEISMIC CRUSTAL PROBING

Let us look at various seismic methods in some detail, first considering results of past studies, then possible future capabilities. A variety of seismic methods are applied to study the crust [9]. They can be broken into the following categories: methods based on seismicity and earthquake mechanisms, earthquake body waves, earthquake surface waves, controlled-source refraction methods, and controlled-source reflection methods.

Seismicity and Earthquake Mechanism: Stress Patterns and Earthquake Risk Reduction

A great deal has been learned about the Earth solely through study of spatial patterns of earthquake occurrence. On a gross scale, the worldwide pattern of hypocenters was an important observation in the development and testing of the concept of plate tectonics, which describes the process whereby the sea floor moves outward from spreading ridges and plunges downward generally beneath continental masses. On any scale the history of past earthquake activity is the most important source of information on the earthquake hazard of the future. Unfortunately, the record is frequently too short to provide reliable predictions of future activity, so that it is necessary to obtain additional information from other sources, such as geologic records of fault movements. In the case of major earthquakes in normally aseismic areas (Charleston, S.C., for example), the historical record commonly includes only one such shock, so that an attempt, using all available information, to understand the cause of that earthquake is vital. Until this is understood, the only safe course is to assume that such earthquakes may occur anywhere in the same province and to prepare, at great expense, for such an event. This example has clear implications for the threat not only to the naval base at Charleston, but also to other coastal installations in the eastern United States.

On a smaller scale, hypocenters precisely located in depth can define faults in the Earth and mark the extent of the rupture zone of a major earthquake. Some earthquakes are associated with surface phenomena such as volcanoes, potential geothermal areas, and, perhaps, near-surface magma bodies. These earthquakes yield information about the state of stress in the earthquake, its structure and, indirectly, the availability of geothermal energy in the area, all of which have potential interest for the Navy.

Earthquake Prediction

Temporal variations in seismic activity are not yet well understood, but the subject offers some fascinating possibilities, including the potential for earthquake prediction. The "seismic gap" method, for example, depends on the condition that earthquake activity in an active seismic belt will, over a sufficiently long interval of time, tend to be distributed more or less uniformly over the belt. Segments lacking recent major activity are those most likely to experience major shocks in the future. Several successful predictions of locations, but not precise times, of large earthquakes have been made in this way. In the temporal patterns there is also sometimes a suggestion of propagation of epicenters along an active belt. A particularly good example of this effect occurred along the Anatolian Fault in Turkey over an interval beginning in 1939, as epicenters moved from east to west. In some cases, periods of quiescence appear to precede buildups in activity preceding major quakes. Tilts and other deformation, water level changes, velocity variations, and other effects may precede earthquakes and with further study serve as predictors. This subject will bear much further investigation.

Over the last 15 years, stimulated by the interest in seismic sources of all types in the context of the nuclear test ban treaty, seismologists have made considerable advances in understanding the earthquake source mechanism. The far-field radiation pattern of initial motions for most shocks fits the simple "double-couple" model that results

from shear failure (Figure 2). (The ambiguity in actual fault-plane direction must be resolved from a knowledge of regional stresses and fault patterns.)

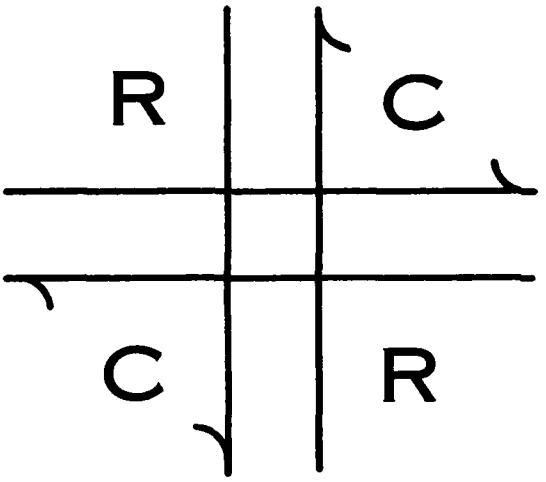


Figure 2—Vectors represent relative motions in the vicinity of the ebicenter, such that strain energy released by the earthquake produces an initial compression in those quadrants labeled C and an initial rarefaction in the quadrants labled R.

From observations of the radiation pattern, the orientation in space of the focal plane and the direction of slip along it can be determined, and following that, the orientation of the principal stresses. More refined studies of the character of the radiated seismic waves, particularly the frequency spectrum, produce additional information on the magnitude of the stress drop (typically some tens or hundreds of bars) and information on the size of the rupture (fault length). Future studies will further develop these methods for the study of stress patterns.

Earth Structure from Large Seismic Sources (Earthquake Body Waves)

Early studies of the Earth's interior were based on measurement of the travel times required for seismic (body) waves to penetrate the body of the earth. Partly because surface waves were not so well understood initially, and partly because of the great penetration and resolution of body waves, much of our knowledge of the earth's interior derives from the latter source. The body wave method continues to be of great importance in the study of the deep interior. Recently, there has been a trend toward the use of nuclear or other controlled sources (see next section), since crustal studies from earthquake sources are less reliable because of the uncertainty in the origin time of the waves. A second problem with earthquake sources is the expense of operating many stations for long intervals while waiting for the appropriate earthquake to occur. In the case of the nuclear explosions, the problem of poorly known origin time and hypocenter are overcome, and very precise and useful surveys can be made. They are, however, limited to sources at the nuclear test sites and hence to certain paths only.

In recent years, to overcome these problems, we have placed a growing emphasis on body wave studies in which the *relative* arrival times are measured at a network of stations near the zone of interest. In this way, travel time anomalies, and hence structures, can be determined on a scale that is comparable to the spacing of the stations of the network and the depth explored. Great redundancy, in the form of large quantities of data, is an advantage in delineating the structure, and recent instrument development, such as good, cheap crystal clocks and improved telemetry, have helped. These methods suffer from the problem of sorting out near-station effects from effects elsewhere along the path, but clearly have potential. Some such studies use the spectra as well as arrival times but as yet these methods have not proved definitive or productive of information on a fine scale.

Surface Waves from Earthquake and Nuclear Explosions

Although the term "surface waves" implies a phenomenon confined to the outer layers of the Earth, some such earthquake-generated waves are so long, and have a skin depth so great, that they penetrate and provide information on much of the interior; such waves are components of some of the free oscillations of the Earth. At shorter wavelengths, surface waves are useful in distinguishing continental structure from oceanic

structure, in measuring details of each, and in determining the approximate thickness and properties of the lithosphere (which is the outer, rigid part of the Earth to depths of 80-100 km). Still shorter waves, some propagating as the fundamental modes and some as the higher modes of guided waves of the Rayleigh*, Love* and shear** types, permit resolution of variations in the crust or crust-mantle system. For a particular source-receiver combination and a particular path, effects are averaged over the path. By combining data for many criss-crossing paths, a somewhat finer lateral resolution of structure can be obtained. The problem of lateral refraction, or so-called multipathing, of surface waves is a major one that has not been resolved and that limits the effectiveness of the method. An additional limitation of the method is the great length of the waves, which leaves them insensitive to features of smaller dimension. Even though, in principle, waves of very short wavelength are trapped in crystal waveguides and could be used in exploration of the crust, difficulties arise in practice because of the rapid attenuation in the Earth of such higher frequency waves and because of lateral heterogeneities that make the approximation of the crust by a simple layered waveguide invalid. Surface wave methods will continue to yield new information on the Earth, but their averaging property is both an asset and a liability, and they will always be limited in resolution.

Controlled-Source Refraction Methods

In order to overcome the difficulties of source location and timing inherent in studies using earthquake sources, controlled explosions are commonly used as sources in studies based on the refraction (and wide-angle reflection) techniques [10]. Except in the case of nuclear explosions, this method is commonly limited to profiles a few hundred kilometers or less in length and a few tens of kilometers or less in depth. A trend in recent years, however, has been to use large explosions in water to extend these ranges somewhat. The method has the advantage that it provides information on both velocity and structure. However, the structures are generally based on relatively crude layered models and interpretation is also limited by spacing of detectors and sources. In the Soviet Union, and increasingly elsewhere, very detailed refraction and wide-angle reflection studies of the crust are carried out, and the studies are beginning to provide information of sufficient detail to be correlated with some large-scale geological features, e.g., crustal uplifts, downwarps, etc. In the United States, crustal refraction studies have been carried out primarily by a few universities, private institutions, and the U.S. Geological Survey, but the effort has never attained the massive scale and focus of the Soviet program, and as a consequence the crust, with its resources, in the United States is still largely unexplored.

Controlled-Source Reflection Studies

The seismic reflection profiling method is a sophisticated, highly developed technique that finds its chief application at present in exploration for petroleum in sedimentary basins. Developed by the petroleum industry at great expense, it provides by far the highest resolution of structural features, as well as good information on both vertical and lateral variations of velocity. The information is very detailed and much more "geological" in character than that provided by any other geophysical method. Elaborate arrays of many seismic sources and receivers are used in a way analogous to those of radar. Explosive and other impulsive sources are used on land and sea, but on land the VIBROSEIS technique, developed and registered by the Continental Oil Company, has gained prominence in recent years. This technique uses as sources giant truck-mounted vibrators that shake the ground with a radar-like

* Rayleigh and Love waves are surface waves, decreasing rapidly in amplitude with depth. Rayleigh waves are characterized by a retrograde elliptic particle notion in a vertical plane parallel to the direction of propagation, while Love waves require a surface layer and are characterized by a horizontal shear motion.

** The shear waves referred to here are "normal mode" waves trapped in a surface layer (or sequence of layers) where they follow paths such that wave fronts reflecting successive from the same boundary are in phase with each other.

chirp, that is, a near-sinusoidal wave of slowly changing frequency. Subsequent processing compresses the chirp to a simple pulse by a correlation process. This method has certain advantages in the form of control of source parameters and environmental acceptability.

Until recently, such modern reflection methods had not been applied to study of the deep crust of the United States. Recently, a national program designated as COCORP (Committee for Continental Reflection Profiling), consisting of university representatives from Cornell, Houston, Wisconsin, Princeton, and others, was begun with National Science Foundation support. The intent is to use the seismic reflection methods of the petroleum industry to explore greater depths in the crust and uppermost mantle with high resolution. Two tests have been conducted so far, one in Texas and one in New Mexico, and they indicate that the method has outstanding potential for delineating detailed characteristics of the deep crust. Widespread exploration of the crust by this method will almost certainly result in a new level of understanding of the history of formation, the physical properties, and ultimately, of the potential usefulness of this vast, unexplored region. As a consequence, many elements of our society will benefit from this research.

FUTURE DEVELOPMENTS IN CRUSTAL SEISMOLOGY

In the next decade, we expect major new advances in our knowledge of the deep crust. Seismic reflection profiling will play a major role in this advance. Current methods explore to depths on the order of 30 to 50 km, including most or all of the crust, but the potential exists for more improvement. For example, larger vibrators capable of much stronger signals and of extending the frequency spectrum to lower frequencies are now being developed in the petroleum industry. Shear-wave vibrators are also being developed. New methods of signal telemetry will facilitate deployment of large arrays of many elements. Such instruments and the corresponding techniques offer the prospect of determining detailed spatial variations in deep structure using both compressional and shear waves. These data may be used to deduce such other physical properties of the crust as its strength, porosity, lithology, and even electrical resistivity, for example. In the latter case, it may be possible to infer electrical properties where crustal structure is either too complex to be resolved electrically or where conducting surface layers limit the resolving power of the electrical method at depth. It is expected that controlled-source reflection studies, such as those described in the previous section, will play an increasingly important role both in delineating the physical properties of deep crustal layers and in locating presently hidden natural resources.

ELECTRICAL CRUSTAL PROBING

Measurement of the electrical resistivity of the Earth is one of several major categories of techniques used in geophysics to explore the subsurface by physical means. Each of these categories has its individual strengths and weaknesses and is not competitive with the others; rather, all are supplementary to each other in the solution of problems in subsurface exploration. Electrical methods for probing the earth are highly diverse, so that there is great flexibility for applying specific techniques adapted to the requirements of specific problems.

While occasional attempts were made to use natural electric fields or measurements of grounding resistance as an aid in mineral prospecting prior to 1900 [11], the modern application of electrical prospecting methods stems from the second decade of this century. At that time Wenner [12] and the Schlumberger brothers [13] devised and made routine use of direct-current methods for measuring earth resistivity. A group of Swedish geophysicists [14] fielded an electromagnetic prospecting system for locating metal ores. These developments are based on the electromagnetic field equations of James Clerk Maxwell [15], who in addition to writing the general equations upon which all electrical prospecting methods are based, made detailed analyses of applications such as the use of four-electrode arrays to measure the electrical resistivity of extended media such as the earth.

In the decades that followed the work of Wenner, the Schlumbergers, Sundberg, and their colleagues, the use of resistivity surveys grew slowly, with the principal application being the search for highly conductive mineral deposits. By the 1930s, electrical surveys were also being applied to search for ground water, to geological engineering, and even to structural studies for oil exploration. Unfortunately, many of these attempts were unfruitful, and electrical prospecting methods did not enjoy the rapid growth in application experienced by several of the other categories of geophysical methods during the 1930s.

Theoretical Complexity

In retrospect, one gains the impression that successful use of electrical techniques four decades ago was probably inhibited because of the complexity of the mathematical behavior of electromagnetic fields. Until quite recently, with the advent of high-speed computers and analytic techniques, the equations on which field measurements were based were not solved numerically. Field methods were developed and field data were evaluated empirically using simplified or intuitive versions of the basic theory. Seismic and potential-field (gravity and magnetic) methods were pursued successfully with pragmatic, simple versions of the basic theory, without full recourse to the more complex aspects of the theoretical backgrounds of the methods. However, with electromagnetic field behavior, predictions based on incomplete versions of theory are often highly misleading; because of this, a number of paradoxes exist in the application of electromagnetic methods. It may be useful at this stage to examine two of these paradoxes, inasmuch as they help in understanding the difficulty in arriving at the most effective means for doing a specific electrical survey.

A well known paradox is the "paradox of anisotropy." Visualize an Earth structure in which the resistivity varies as a function of direction. This anisotropy often occurs in layered rock, where the resistance to current flow is less along the fine laminations than across them. One might reasonably expect high values if measurements are made across the beds along the direction of high resistivity, as in the case of measurements made in a vertical bore hole penetrating a sequence of flat-lying beds. On the contrary, Keller and Frischknecht [16] show analytically that the resistivity measured with a vertical array of electrodes (with current being transmitted from one end to the other of the array, vertically along the borehole) is that for current flowing horizontally through the layers. Even when this phenomenon is demonstrated mathematically in a straightforward manner, some tend to dismiss it as a mathematical artifact. In fact, the paradox can readily be shown in field measurements and demonstrates the important fact that electromagnetic fields can behave quite differently than we would expect intuitively.

Another paradox is little known and unnamed. It is often not understood at first, even by those with a fairly sophisticated understanding of electromagnetic fields. This paradox deals with the skin depth, or penetration depth, of currents in a sequence of layers with alternating high and low resistivities. It is well known that when an electromagnetic field propagates through a partially conducting body such as the Earth, it loses energy as it generates eddy currents and dissipates heat in that body. The eddy currents are generated more intensively at higher frequencies, so it is generally believed that the penetration of the electromagnetic fields decreases at higher frequencies. In a medium of fixed conductivity this is true, but in the layered sequences usually encountered in the Earth the situation is often different. The Earth's outer region typically consists of three zones with different electrical conductivities. The middle layer (at depths of about 5 to 25 km in the crust) is highly resistive compared to those above or below. When Earth resistivity surveys are made with direct-current methods, the current is screened from the interior of the earth by the highly resistive middle layer. With unusually large electrode separations, current can be forced into the third layer, but otherwise resistivity measurements made at the surface generally will not show the presence of the third layer, nor generally provide any information about the subsurface past the top of the middle layer. When an a.c. induction method rather than a d.c. method is used, the electromagnetic fields can penetrate the resistant layer to induce currents in the underlying

conductive rock without difficulty. Thus, contrary to general belief, greater penetration is obtained in this case by raising the frequency, since the alternating field can induce current flow beneath the resistive layer, while a direct current cannot penetrate. Despite the belief that d.c. methods can provide information from greater depths than a.c. methods, the opposite is often true.

Recent Advances

Major changes are taking place in the application of electrical methods to geophysical exploration, for several reasons. An important factor is the increased use of high-speed computer facilities at a great many locations, along with the availability of highly efficient numerical techniques that were unknown a few years ago. Thus, a practitioner of electrical surveys can now perform an exact evaluation of his data using electromagnetic theory (within the resolving power of the method). He need not be mislead by intuitive approaches that result in errors due to the paradoxical behaviors described above.

The second factor altering the whole complex of electrical prospecting is the growth in importance of geothermal energy as a supplement to more conventional energy resources. Because heating of rock to the temperatures required for production of commercial geothermal fluids reduces their resistivity by a significant factor (from 5 to 7, with resistivities in geothermal reservoirs being typically from 1 to 10 Ω-m), electrical surveys have become a primary method for geothermal exploration. Over the past 3 to 5 years, probably more effort has been expended on electrical surveying for geothermal systems than in the entire preceding half century for all applications. Moreover, geothermal prospecting requires greater precision than was demanded of previous applications. Geothermal reservoirs lie at greater depths than are of interest in mining, engineering, or ground-water applications. Not only must greater effort be expended on exploring to greater depths, but interpretations must be made with a high confidence factor because of the expense of drilling to test geophysical findings at great depth.

The interest in geothermal exploration has led to very rapid advances in both the instrumentation and field technology used to measure resistivity and in the analysis and interpretation of the data. In field studies, while conventional depth soundings (including the Wenner and Schlumberger arrays) are used to some extent, we are emphasizing new methods that allow large-scale measurements to be made with an economy of operation and an increased reliability of results. The principal new methods that are in use are the dipole mapping method [17], the quadrupole method [18-20], and the time-domain electromagnetic sounding method [21]. Other methods that appear to have considerable potential for use in exploration but which are only now being used experimentally include the geomagnetic deep sounding method [22, 23], the magnetotelluric method [24, 25] and telluric surveys.

Geomagnetic deep soundings, magneto-telluric soundings, and telluric surveys are based on the use of variations in natural electromagnetic fields as an energy source. The dipole, quadrupole, and time-domain electromagnetic methods are all based on the use of a controlled local source of energy. There are advantages and disadvantages to both approaches. With natural fields, no great effort is involved in obtaining high-intensity fields that penetrate to considerable depths, but measurements must be continued over a long enough period to obtain representative data from fields that change randomly in time. With controlled sources, measurements are made quickly, but considerable effort may be required in emplacing the source. Over the last few years, source equipment has been increased in capacity from the 5 to 20 kW that had been developed for use in minerals exploration to levels as great as 200kW [26]. The use of power levels of this magnitude permits rapid and accurate determinations of electrical structure to depths of 3 to 5 km, even in basins filled with highly conductive rocks (or to greater depths in more favorable circumstances).

FUTURE TRENDS IN ELECTRICAL PROBING

It appears likely that the size of power supplied will continue to grow as the need for studying electrical structure at great depths continues to

grow in importance. For example, in development of geothermal energy, considerable importance is attached to the development of hot, dry rock systems over the next decade. At present, it appears that exploration for such resources will be based on the recognition of areas of high temperature in the lower part of the crust and the upper mantle, where the high temperature renders rocks unusually conductive at shallow depths (10 to 20 km).

Future Equipment Trends

The use of very large sources for electrical surveys appears quite feasible in terms of present technology. The 200kW source mentioned earlier consists of a diesel engine driving an electrical generator with auxiliary rectifiers and pulse-forming circuits; the total weight is 6000 lb (2720 kg), so that the system is fully mobile under field conditions (a photograph of the system is shown in Figure 3). Over a limited range, the power capability of such a system can be increased readily, with a proportional increase in weight. Thus, it would be reasonable to build a power supply with a capacity of 0.5 to 0.75 MW merely by scaling the size of the engine, the electrical generator, and the truck to carry it. The weight would be 15,000 to 22,500 lb (6,800 to 10,200 kg) which would be readily transportable on a vehicle that could maneuver over moderately rough terrain. Unfortunately, a tenfold increase in power will normally only double the approximate depth to which an electrical survey can be carried. The weight and size of conventional diesel engines have, to date, made a tenfold increase in power supply capacity impractical.



Figure 3—This shows a truck-mounted 200 kw diesel-driven D.C. generator used as source for electrical prospecting.

Consideration is being given to the design of very high power sources based on novel prime sources of energy, and in fact a 50-MW power supply has been field tested for use in electromagnetic sounding recently in the U.S.S.R. One approach to building a 20-MW power supply is based on the use of an aircraft-type turbine engine to spin an electrical generator. Such systems are being constructed for powering remote facilities such as mines or remote communities. Such a generator would be built in two pieces and would require two vehicles, each capable of transporting approximately 30,000 lb (13,600 kg). It is likely that the converter equipment needed to produce the d.c. used in electrical sounding will weight perhaps 10 tons (9070 kg) and will require a third truck, but such an advanced system seems both technically feasible and desirable for exploration of the electrical nature of the deep crust, in view of the importance of electrical surveys in such applications as geothermal prospecting and earthquake prediction.

An even more imaginative approach to high-capacity mobile electromagnetic sources is the use of a magneto-hydro-dynamic (MHD) generator as a prime source. An MHD generator is one in which a stream of very hot ionized gas (3000 to 4000°K) is passed between the poles of a magnet. In accordance with Faraday's Law, this generates a voltage and hence a current in an external circuit. A gas stream with a cross section of a few square meters is capable of producing about 50 MW of power output. The heat developed in an MHD generator is very great, and for continuous operations, massive cooling facilities must be provided. In exploration, where the energy source is used intermittently, the need for cooling is greatly reduced, and it is likely that an MHD generator with up to 50-MW capacity would be built with a weight of 50,000 lb (22,680 kg).

It is clearly possible to increase the power capacity of generators by two orders of magnitude beyond 0.2 MW, which is the largest in use today. Such power supplies would make possible detailed studies of the electrical structure of the crust and upper mantle using the electromagnetic methods. The rate at which such systems are developed will depend to a large extent on the inten-

sity of interest in the properties of the crust, and the degree of success achieved in using electrical soundings for geothermal exploration and earthquake prediction.

The capability of an electrical surveying system to probe the earth is as much a function of the receiver sensitivity as it is of the source strength. At present, most receiving equipment employs analog or simple magnetic tape recording equipment. Processing is computer oriented, so that there are often significant delays in converting field data into a format compatible with digital computers. In a few cases, minicomputers have been used in the field, primarily as devices to apply digital filters and to carry out synchronous stacking. It seems likely that both with the controlled-source methods which use some kind of manmade source and the natural-field methods which depend on fluctuations in the earth's magnetic field (micropulsations), newly developed digital microprocessors will be used to sample, filter, and stack the received signals. Microprocessors are low in cost and are basically preprogramed, and this means that their operation is rapid and efficient.

Data Interpretation

At the present time, very rapid strides are being made in our ability to interpret field data, and it is likely that even more significant developments will be made in the next few years. Thus, with the advent of more powerful current sources and microprocessors, and with improved theoretical developments [27] our ability to develop the electrical prospecting method into a more readily useable and available tool will increase rapidly over the next decade or so.

Interpretation of electrical data is twofold. First is the conversion of the measured field quantities (e.g., voltages, currents, and electrode locations) to resistivity values as a function of lateral position and depth in the earth's crust. Second is the geological interpretation of the inferred resistivity distribution. Both encounter significant problems at the present time.

The interpretation of resistivity distributions in the Earth can be further subdivided. First is the direct approach, or "cut-and-try" method, in which one assumes a model, computes the field distribution it would produce for a given electrode distribution, and compares the computed field with the observed field data. The other is the inverse approach, in which field data are inverted to yield an Earth model directly. The direct approach is much simpler to handle mathematically than its inverse in which the electrical structure in the earth is inferred directly from the electrical measurements in the field. The first problem of significance to be solved by the direct approach was that of the response of a direct current sounding method over a flat layered earth [13]. While Stefanescu was able to obtain a solution, these early analytical results were of little value until the 1960s, when computers became generally available to overcome the numerical difficulties.

Direct solutions for time-varying electromagnetic fields are even more difficult to evaluate numerically, and only recently have the necessary numerical techniques been developed [27-29]. The first useful numerical results were provided by Wait [30].

Much effort has been spent on obtaining forward solutions for the d.c. and electromagnetic response of exotic Earth structures, such as buried spheres or cylinders. While interesting exercises, these computations are of limited value in interpreting field data because of their specialized nature. It has been shown that the forward problem of the response of Earth structures of arbitrary shape can be solved numerically in several ways, including the finite difference methods [31], finite element methods [32], transmission line methods [33], and moment methods [34, 35]. These methods are each capable of providing any required accuracy in calculation, but all are still quite expensive in terms of computer time.

If the forward problem can be solved, methods can be found, in principle at least, to do the inverse problem. Basically, inversion consists of making a guess as to the probable earth structure, followed by a numerical solution of the forward problem to see how closely the guess approximates observed data. Then, in contrast to the cut-and-try method, a series of derivatives of the error with respect to the model parameters are computed. These derivatives are used to set up a series of normal equations, which can then be

solved to find the "correct" model. Several mathematical techniques are available for solving the normal equations; they, in essence, involve minimizing the error. These include Marquardt's method [29, 36, 37], the Backus and Gilbert method [38-40] and the Fibonacci search model [21].

In the last several years, these inversion methods have produced spectacular results for sets of data that can be thought of as representing a layered Earth (an Earth model in which resistivity varies with depth only). No successful application of existing inversion techniques to two-dimensional models of the Earth have been reported. Such inversions presently require time-consuming computations and are too costly for any conceivable application to electrical surveying techniques. However, recent developments suggest it is reasonable to expect that better methods for refining models in the inversion process will be discovered to make two- and three-dimensional interpretations practical in the next few years.

Summary

In summary, the rapid advances made both in acquiring field data and in interpreting those data portend even more significant improvements in technology in the next few years. As these improvements are made, electrical surveying techniques will provide a powerful tool for studying crustal structure in detail. When details of the resistivity distribution in the crust are combined with data from improved seismic techniques for measuring such quantities as Poisson's ratio and seismic attentuation, we will undoubtedly be able to achieve major advances in the state-of-the-art for inferring such quantities as the porosity, fluid content, permeability, temperature, pressure, lithology, strength, and stress in the crust in addition to specific applications which will be discussed later.

LABORATORY STUDIES

Laboratory data are essential to interpreting the geophysical field measurements of seismic velocity, electrical resistivity, gravitational and magnetic attractions, and thermal data. The goal is to interpret such field data in terms of the underlying lithology and its porosity, microcrack content, permeability, fluid content, pore pressure, confining pressure, temperature, strength, and stress. These properties are not directly measurable from the surface. In principle, the procedure is to measure the physical properties of a suite of rocks with various porosities, fluid content, temperature, pressures, etc., and to catalog their physical characteristics under various conditions. Then, by making multidisciplinary field observations, it should be possible to interpret subsurface geology with greater precision than has ever before been possible, using these multidisciplinary laboratory data for control.

In combination with such data, a clear understanding of geologic pressures and principles is required before the resulting interpretations can be given meaning in geological terms (mineralogical, petrological, tectonic, structural, historical, economic, or stratigraphic).

Physical Properties of Rocks

Historical—Scientific interest in the properties of rocks and minerals began no doubt in antiquity. We can speculate that "engineering data" were accumulated and passed orally to succeeding generations. Scientific interest is clearly evident in the writings of the 17th and 18th centuries and increased steadily in the 19th and 20th centuries. Maxwell, the father of electromagnetic theory, was concerned with the electrical properties of crystals, W. L. Bragg with their response to X-rays, Voigt with their elasticity, and Fourier with their thermal conductivity.

The scentific exploration of the earth in the 1800s added considerable impetus to developing an understanding of the physical properties of rocks, particularly as functions of pressure and temperature. Lord Kelvin's debate with geologists on the age of the Earth, which began in 1862 [41], was based on thermal gradients from heat flow values measured in boreholes and thermal conductivities measured on rock samples. Thermal conductivity was important for estimates of the Earth's thermal flux from the interior.

Compressibility and other elastic properties were needed for the interpretation of seismic data on the velocities of shear and compressional waves in the Earth's interior. Adams and Coker [42] recognized the scientific potential of interpreting the seismic data in terms of rock types and began the collection of laboratory data on elastic properties.

During the past 30 years, the set of data on all physical properties of rocks has increased manyfold. The few measurements on thermal conductivity available to Lord Kelvin have increased to tens of thousands. The few data on the compressibilities of rocks reported by Adams and Coker have increased to thousands. Similar increases exist for the data set on velocity of elastic waves, electrical conductivity, and hydraulic permeability, to name only a few examples. But not only has the data set increased, our understanding of the principles has increased significantly also. Let us turn now to a few examples of the laboratory measurements of physical properties.

Compressibilities—The compressibilites of two rather different rocks are shown in Figure 4. To obtain those curves, electrical strain gauges are epoxied on small specimens, encapsulated in a rubbery material (sylgard), placed in a pressure vessel, and measured (strain as a function of pressure). Compressibility is the slope of the strain-vs-pressure curve. Note the very large difference in the two curves: one changes rapidly with pressure at low pressure, the other scarcely changes with pressure. *This striking difference in behavior* with pressure was first observed 50 years ago during the classic studies of Adams and Williamson [43] on compressibility and correctly interpretated by them. They suggested that the large rate of change at low pressure was due to microcracks in the rock that close with pressure. Their interpretation has stood the test of time. Today, with a scanning electron microscope, we can examine rocks with magnifications as high as 100,000X and actually see the microcracks that Adams and Williamson suggested on indirect evidence to be present. The microcracks are extensive in Westerly granite, but are (almost) completely absent in Frederick diabase.

Microcrack Control of Physical Properties—The microcracks affect many other physical properties. In Figure 5 we show examples of the velocities of compressional and shear waves. To

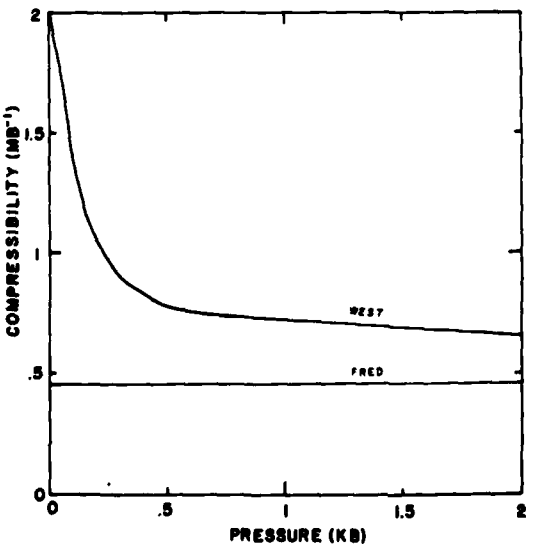


Figure 4–Effect of pressure on compressibility of rocks. Westerly granite contains many microcracks, which act as weak springs until closed. Frederick diabase contains no cracks.

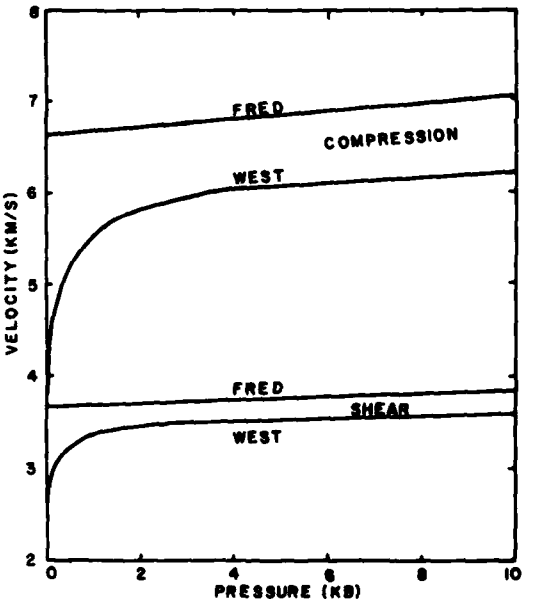


Figure 5–Velocity of elastic waves in rocks. The pressure closes most cracks by 2 kbar in Westerly, and the velocities at P > 2kbar are intrinsic. The Frederick diabase contains no cracks; therefore its velocities change very little at these low pressure. Redrawn from Birch [44] and Simmons [45].

measure velocity, we use an ultrasonic technique developed by Professor Francis Birch of Harvard and measure the delay of an elastic pulse traveling through a small cylinder of rock. Note that the velocities of the granite increase rapidly at low pressures but slowly at higher pressure. The velocities of the diabase increase uniformly with pressure over the whole pressure range. These features are readily understandable in terms of the microcracks present in the two rocks.

We can measure the volume of microcracks, using the same data that were used to obtain Figure 4. The key to measuring crack volume is recognition that the large changes at low pressure are due to the cracks (as pointed out by Adams and Williamson [43]). In Figure 6, if we extrapolate the linear portion to zero pressure, the intercept *is* the strain due to cracks and therefore the crack porosity. We can even obtain the distribution function for crack porosity, provided that strain is measureed with extremely high precision. The intercept $\zeta(P_c)$ of the tangent to the strain curve is the strain at zero pressure due to all cracks that close at pressures equal to or less than P_c. Such high-precision measurement on Westerly granite has given a crack porosity of 0.07-0.1%. The value for Frederick diabase is less than 0.0006%, the experimental error of the technique.

Microcracks in rocks are very common. When present, they dominate the physical properties. Microcracks in the crust are dynamic; they form, anneal, and then form again. They are produced by the same stresses that cause earthquakes. They anneal because they are thermodynamically unstable. The cycle may be repeated many times in any given region throughout the long periods of geologic time.

As an example of the direct observation of microcracks with a scanning electron microscope (SEM) and an optical microscope, we show in Figures 7A and 7B cracks in a billion-year-old

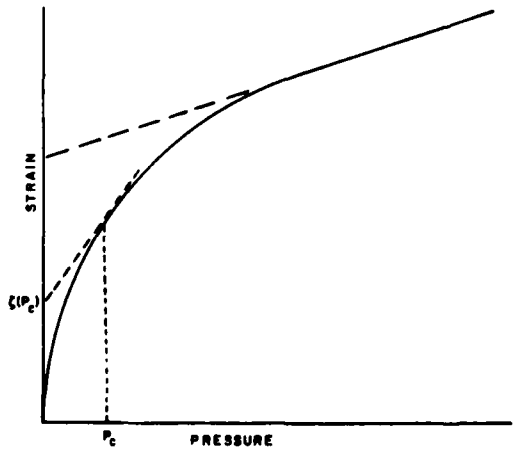


Figure 6–Schematic compression curve for rocks that contain microcracks. Extrapolation of linear portion to zero pressure yields microcrack porosity. The intersection of the tangent yields strain due to all cracks closing at $P \leq Pc$.

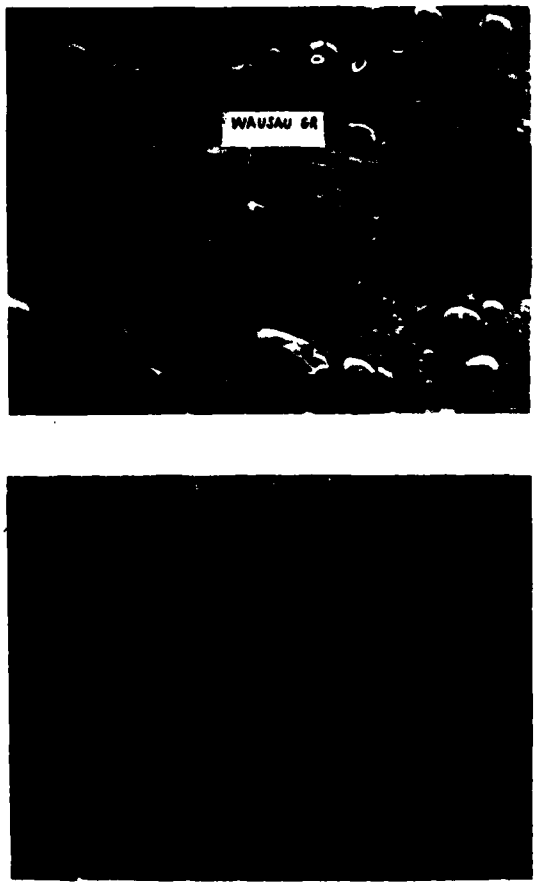


Figure 7–Photomicrographs of Wausau granite. The field-of-view overlap of the optical micrograph (B) is marked on the SEM micrograph (A). The mounds are produced during preparation of the specimen. With the SEM, we see only the surface; with the optical microscope, we see features throughout 10 mm. The two healed cracks appear as rows of holes in the SEM but as planes of bubbles in the optical microscope. The mineral is quartz.

granite from Wausau, Wisconsin. One crack is now open, and two cracks, formerly open, are now marked only by fluid-filled holes. Such open cracks are the ones that cause the physical properties to change rapidly with pressure in the laboratory and with depth in the Earth. The annealed cracks have very small effects on most properties.

Laboratory Methods

Advantages and Disadvantages—The study of the properties of rocks in the laboratory has several advantages. The rock can be characterized with respect to mineralogy, composition, texture, microcracks, defects, and so on. Each property can be measured as precisely as one wishes. Several properties can be measured on the same sample in order to search for empirical relations among properties. The conditions of pressure, temperature, and fluid pressure can be varied readily. Briefly stated, the physical properties can be measured precisely in well-characterized specimens under carefully controlled conditions.

There are also several disadvantages to the study of rocks in the laboratory. The state of stress in the Earth's crust is unknown and therefore cannot be modeled properly in the laboratory. The exact boundary conditions to be used on laboratory specimens are uncertain. Should the surfaces be free or constrained? Should surface tractions be controlled? The sampling problem is large. How can a few cubic centimeters be statistically representative of hundreds of cubic kilometers? Some rocks that occur in significant volumes at depth may be rather rare, or perhaps absent, at the surface. How can we obtain samples of them for laboratory work? The frequency of signals used for field measurements usually differs from the frequency used in the laboratory, and some properties depend strongly on frequency. Seismic signals in the Earth vary from 0.01 to 1000 Hz, but ultrasonic signals used in the laboratory range from 0.1 to 100 MHz. Fortunately, over this wide range, elastic properties exhibit little dispersion. Electrical signals in the field range from 10^{-3} to 10^{3} Hz, but laboratory measurements are readily made at frequencies of 10^{3} to 10^{7} Hz. Unfortunately, the electrical properties of some rocks are strongly dispersive over the range 10^{3} to 10^{7} Hz. Hence, the matter of frequency and dispersion must be examined for each property. So the laboratory study of the physical properties has both advantages and disadvantages. What then is its role? How can we use its advantages and avoid its disadvantages? First, we can obtain data that allow us to examine the effects of one parameter at a time. For example, in obtaining the data for Figures 4 and 5, we varied a single parameter (pressure) for each rock. We could perform similar measurements in which we varied only temperature, or pure fluid pressure, or initial porosity of the microcracks, and so on. Indeed, we have done many of these experiments in the past 30 years to isolate the effect of each variable on each physical property. In Figure 8 we illustrate this approach with the effect of composition on the elastic properties of crack-free rocks. From measurements of the velocity of samples with a range of composition, at a pressure of 10 kbar when all cracks are closed, Francis Birch showed that, to first order, the velocity of compressional waves is a function of density and mean atomic weight.

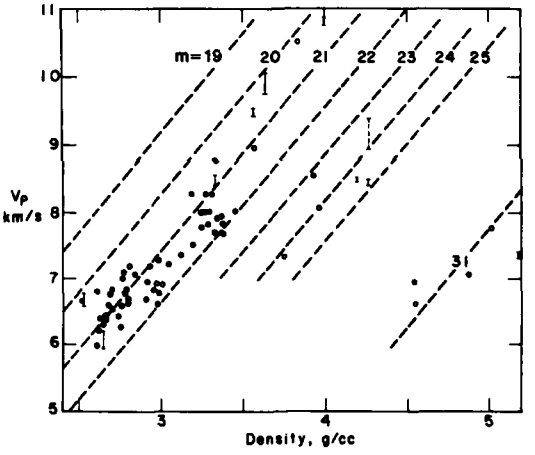


Figure 8–Effect of composition on the velocity of compressional waves in crack-free rocks. ρ is density, Vp is velocity, and m is mean atomic weight. After Birch [46].

Relations Among Physical Properties—A second valid use of laboratory data is for establishing

relations among various physical properties. These relations can then be used with field measurements on one property to estimate other properties. The desired property may be very difficult, impossible, or very expensive to measure. Consider, for example, the strength of rock, an important factor in the design of foundations for dams, large buildings, and undersea facilities; in the cost of highway excavations; and in the design of missile silos to withstand the pressure produced in nuclear explosions. For some applications strength is best measured directly on cores. For other applications it is best estimated from the velocities of compressional and shear waves.

In-situ Properties—A third use of laboratory data is for predicting the properties of various rock types *in-situ* as a function of depth. For the shallow crust, the key to such predictions, we believe, is the recognition that microcracks dominate the physical properties. Hence, a model of the microcracks as a funtion of depth can be used, eventually, to predict the variation of physical properties with depth. The microcrack model would be based on rock type, stresses, stress history, and the tectonic history of the region. The basis for each predictive ability is only now being developed; it is on the research front today and is being actively investigated.

Testing of Theory—A fourth use of laboratory data is for testing theoretical expressions. For example, the effects of cracks on various properties can be calculated for certain simple theoretical models. Penny-shaped and ellipsoidal models have been popular with theoreticians because the associated equations are tractable and the solutions are often expressible in relatively simple terms. Unfortunately, cracks with such simple geometry are rarely seen in rocks. Is the apparent match between calculated and measured values coincidental, perhaps merely the result of fitting experimental data with two adjustable parameters? How good or how poor *are* the theoretical solutions? Can they be improved with small perturbations in the models? We believe that the data on physical properties measured in the laboratory, on well-characterized samples under carefully controlled conditions, can be used to test the validity of theoretical models and also to improve the theoretical models.

NAVAL APPLICATIONS

We have outlined many of the reasons a knowledge of the crust is important to the Navy. It is our purpose in this section to amplify the anticipated advances (in each of the three areas described above) in terms of their likely impact on our understanding of crustal properties of importance to the Navy.

First, it is important to recognize that a multidisciplinary approach is necessary for inferring physical properties of the crust from geophysical observations made at the surface. Thus, while we can measure seismic velocities and attenuations, electrical conductivities, gravity, and magnetic and thermal fields, we cannot directly measure porosity, fluid content, temperature, strength, stress, permeability, or the lithology of the crust at depths below the surface.

Therefore, although certain geophysical techniques have greater resolving power than others, the problems to be attacked must be approached from a multidisciplinary viewpoint. For this reason, the specific applications will not be repeated three times in this discussion but rather will be discussed on a joint basis where this is possible.

Furthermore, it is important to reiterate that the interpretation of geophysical field data can be made only through a comparison with multidisciplinary values of the physically observable properties of rocks in the laboratory as a function of porosity, fluid content, pore and confining pressure, temperature, lithology, permeability, etc.

Geothermal Energy

Whenever knowledge of the Earth's interior is potentially of value, the high-resolution seismic methods are likely to be in demand. For example, in the case of geothermal energy sites, the structure around the site and particularly the configuration of the igneous pluton or magma body can be determined seismically. Several new seismic methods are currently in use. One such study uses microearthquake-generated, reflected compressional and shear waves to map magma bodies; another maps deep structures by plotting the spatial pattern of attenuation of seismic waves; a

third uses the seismic reflection profiling technique. In fact, microearthquakes, and even unusually high, continuous background noise, has been used for prospecting in geothermal areas. Development of a geothermal powerplant for Navy or general use will surely entail exploration of the earth by many or all of these techniques, as tailored by experts to fit the demands of particular sites. Such techniques are still in their developmental stage but will play increasingly important roles in the future. Currently, electrical resistivities are found to be quite low (generally on the order of 10Ω-m) over the central region of geothermal reservoirs. Combining seismic and electrical results with laboratory data, we expect to improve resolving power for evaluating reservoir temperature, fluid content, etc., and even reservoir lifetime.

Electrical Properties of the Crust

Under conditions of complex surface conductivity distributions, and especially where surface conductivity is very high, it is often quite difficult to resolve electrical observations into a meaningful crustal conductivity distribution. Because the seismic method generally has the edge in terms of resolving power, we expect to be able to convert seismic measurements of shear and compressional interval velocities to electrical conductivity distributions through relationships being developed in the laboratory. This will be possible in principle because of the intimate control exerted by microcracks over crustal properties.

New techniques using seismic shear and compressional waves may prove useful in evaluating the electrical properties of the ground planes of large low-frequency antennae. Such sites should be explored seismically for crustal anomalies in order to evaluate the distribution of ground conductivity and thus the antenna radiation patterns in areas of complex geology.

Earthquake Hazards

The earthquake is a potential military hazard. An earthquake of magnitude 8 struck Japan in 1944 during World War II and posed a considerable problem for Japanese forces. The Naval base at Bremerton, Wash., suffered several million dollars damage in 1965 from a magnitude 6.5 earthquake which occurred near Olympia, Wash., some 35 mi (56 km) away. Another notable case of damage occurred at the Kodiak Naval Base, caused by the magnitude 8.4 Anchorage, Alaska, Good Friday earthquake of March 27, 1963. The epicenter for this earthquake was located about 270 mi (435 km) from Kodiak island under Prince William Sound. Damage at Kodiak was associated primarily with a tsumani (seismically generated sea wave) caused by the earthquake.

For the military, it may be as important to know that an earthquake is *not* going to happen (in choosing the site of a naval base, for example) as to know that one is imminent. Major efforts are now underway and growing in the United States to develop ways of predicting earthquakes or otherwise lessening the hazard. The focal mechanism and seismicity studies noted on p. 331, ff. Figure 2 are a part of that effort. So are related studies of surface deformation by surveying and by application of strain meters, tilt meters, water-level recorders and other devices. The structures and geologic history of an earthquake-prone region are important in understanding the hazard. At present the VIBROSEIS (controlled source of mechanical vibrations) is being used to monitor possible changes of travel time to a reflector deep in the crust in California to test the hypothesis that changes in seismic velocity precede earthquakes.

In addition, the measurement of electrical resistivity across a fault zone prior to an earthquake produces characteristic changes in the resistivity across the fault due to the changing pore pressure in the zone of stress, as the fault pores adjust slightly prior to the earthquake. At this time, considerable hope is attached to the possibility that the electrical method may provide a strong indicator of stress conditions in the regions of an impending earthquake.

Seismic Communication

Although communication by means of seismic waves has the inherent disadvantage of low information rate, significant delay, and limited range, the Navy should be alert to developments

of more powerful sources and more sensitive detectors of seismic waves because some special situation might arise in which seismic communication would be desirable. Reconnaissance studies based on earthquake sources are of interest in this regard as a means of discovering channels of low attenuation or zones of focusing within the Earth.

Electromagnetic Communication

As mentioned earlier, communication through the crust may be possible electromagnetically. This will undoubtedly require that the electrical properties of the crust be approximately independent of lateral inhomogeneities in the lithology. This problem once again calls upon a multidisciplinary approach to resolving questions about the physical properties of the crust, which are particularly difficult to measure in areas of thick sedimentary cover, ot difficult surface topography, or of complex geology.

SUMMARY

The Navy must operate in the Earth's environment. The Earth interacts with and affects naval systems. The Earth is potentially useful as an energy source. It may act as part of operational (e.g., communication) systems. Earthquakes pose potential threats to naval bases. For these and related reasons, it is essential for the Navy to understand the fundamental properties of the solid Earth. In a broader sense, the solid Earth is even more important to study in the context of national defense because it is the storehouse of an infinitely complex and uneven distribution of mineral and energy resources that are essential to the functioning of the U.S. Navy.

Our national prosperity depends on the ability to make use of the Earth, and so does our defense. The contribution that ONR can make to a study of the Earth through research in the solid Earth sciences must be seen in this broader context, which includes an intricate network of studies of all aspects of the geology and geophysics of the Earth.

REFERENCES

1. M. C. Hendershott, "The Effects of Solid Earth Deformation on Global Ocean Tides," *Geophys. J. R. Astron. Soc.* **29**, 389–402 (1972).
2. H. H. Hess, *History of the Ocean Basin in Petrological Studies: A Volume in Honor of A. F. Buddington*, A. E. J. Engel, ed., 599 p., Geological Society of America, New York, 1962.
3. B. L. Isacks, J. E. Oliver, and L. R. Sykes, "Seismology and the New Global Tectonics," *J. Geophys. Res.* **73**(18), 5855–5899 (1968).
4. J. M. Bird and B. L. Isacks, eds., *Plate Tectonics, Selective Papers from the Journal of Geophysical Research*, American Geophysical Union, Washington, D.C., 1972.
5. Wyllie, *The Dynamic Earth: Textbook in Geosciences*, John Wiley and Son, New York, 1971.
6. J. Ewing and M. Ewing, "Sediment Distribution on the Mid-Ocean Ridges with Respect to Spreading of the Sea Floor," *Science* **156**, 1590 (1967).
7. J. Heirtzler et al., "Marine Magnetic Anomalies and the Geomagnetic Time Scale, *J. Geophys. Res.* **73**(6), 2119–2146 (1968).
8. S. LePichon, "Sea-Floor Spreading and Continental Draft," *J. Geophys. Res.* **73**(12), 3661–3697 (1968).
9. S. Mueller, ed., "Special Issue: The Structure of the Earth's Crust Based on Seismic Data," *Tectonophysics* **20**(1-4) (1973).
10. Steinhart and Meyer, "Explosion Studies of Continental Structure," Carnegie Institute of Washington, Publ. 622, Washington, D.C., 1961.
11. C. A. Heiland, *Geophysical Exploration*, 1013 pp., Prentice-Hall, New York, 1940.
12. F. Wenner, "A Method of Measuring Resistivity," Bull., NBS 1a: Paper 258, p. 469, 1915.
13. C. Schlumberger, *Etude sur la Prospection Electrique du Sous-Sol*, Gauthier-Villars, Paris, 1929.
14. K. Sundberg, "Principles of the Swedish Geoelectrical Methods," *Gerlands Beitrage zur Geophysik, Erganzungshefte* **1**, 298 (1931).
15. J. C. Maxwell, *Treatise on Electricity and Magnetism*, 1082 pp., Dover, New York, 1888.
16. G. V. Keller and F. C. Frischknecht, *Electrical Methods in Geophysical Prospecting*, 527 p., Permagon Press, Oxford, 1966.

17. G. V. Keller et al., "The Dipole Mapping Method," *Geophys.* **4053,** 451 (1975).
18. A. Morris, "Quadripole Mapping Near the Fly Ranch Geothermal Prospect, Northwest Nevada," M.Sc. Thesis T-1699, Colo. School of Mines, 100 p., 1975.
19. T. Tasci, "Exploration for a Geothermal System in the Lualualei Valley, Oahu, Hawaii," M.Sc. Thesis T-1743, Colo. School of Mines, 87 p., 1975.
20. D. Doicin, "Quadripole-Quadripole Arrays for Direct Current Measurements—Model Studies," *Geophys.* **41**(1), 79–95 (1976).
21. G. V. Keller, "A Comparison of Two Electrical Probing Techniques," *Geoscience Electronics,* in press (1976).
22. F. E. M. Lilley, "Magnetometer Array Studies: A Review of the Interpretation of Observed Fields," *Phy. Earth Planetary Interiors* **10**(3), 231–240 (1975).
23. M. C. Frazer, "Geomagnetic Deep Sounding with Arrays of Magnetometers," *Rev. Geophys. Space Phys.* **12,** 401–420 (1974).
24. L. Cagniard, "Basic Theory of the Magnetotelluric Methods," *Geophys.* **18**(3), 605 (1953).
25. Keeva Vozoff, "The Magnetotelluric Method in the Exploration of Sedimentary Basins," *Geophys.* **37**(1), 98 (1972).
26. N. Harthill, "The Time-Domain Electromagnetic Sounding Method," *Geoscience Electronics,* in press (1976).
27. D. P. Ghosh, "The application of Linear Filter Theory to the Direct Interpretation of Geoelectrical Resistivity Sounding Measurements," *Geophys. Prosp.* **19**(2), 192–217 (1971).
28. W. L. Anderson, "Fortran IV Programs for the Determination of the Transient Tangential Electric Field and Vertical Magnetic Dipole for a M-Layered Stratified Earth by Numerical Integration and Digital Linear Filtering," USGS Publ. PB 226 240/5, Denver, Colo., 1973.
29. J. J. Daniels, "Interpretation of Electromagnetic Soundings Using a Layered Earth Model," Ph.D. Thesis T-1627, Colo. School of Mines, 86 p., 1974.
30. J. R. Wait, "Mutual Coupling of Loops Lying on the Ground," *Geophys.,* **19**(2), 290 (1954).
31. I. R. Mufti, "Finite-Difference Resistivity Modeling for Arbitrarily Shaped Two-Dimensional Structures," *Geophys.* **41**(1), 62–78 (1976).
32. J. H. Coggon, "Electromagnetic and Electrical Modeling by the Finite Element Method," *Geophs.* **36,** 132 ff. (1971).
33. A. Dey et al., "Electric Field Response of Two-Dimensional Inhomogeneities to Unipolar and Bipolar Electrode Configurations," *Geophys.* **40**(4), 630–640 (1975).
34. Colin T. Barnett, "Theoretical Modelling of Induced Polarization Effects due to Arbitrarily Shaped Bodies," Ph.D. Thesis 1453, Colo. School of Mines, 239 p., 1972.
35. J. O. Parra, "Electromagnetic Scattering from Conductors in a Conductive Half-Space Near a Grounded Cable of Finite Length," Ph.D. Thesis T-1711, Colo. School of Mines, 166 p., 1974.
36. D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Non-Linear Parameters," *J. Soc. Indus. Appl. Math.* **11**(2), 431–441 (1963).
37. C. M. Crous, "Computer-Assisted Interpretation of Electrical Soundings," Colo. School of Mines, M.Sc. Thesis 1363, 108 p., 1971.
38. G. E. Backus and J. F. Gilbert, "Numerical Applications of a Formalism for Geophysical Inverse Problems," *Geophys. J. R. Astron. Soc.* **13,** 247–276 (1967).
39. J. R. Inman, Jr., Jisoo Ryu, and S. H. Ward, "Resistivity Inversion," *Geophys.* **38**(6), 1088–1107 (1973).
40. W. E. Glenn, Jisoo Ryu, and S. H. Ward, "The Inversion of Vertical Magnetic Dipole Sounding Data," *Geophys.* **38**(6), 1109–1129 (1973).
41. W. Thomson, "On the Secular Cooling of the Earth," *Math. Phys. Pap.* **3,** 295–311 (1890).
42. F. D. Adams and E. G. Coker, "An Investigation into the Elastic Constant of Rocks, More Especially With Reference to Cubic Compressibility," Publication 46, Carnegie Institute, Washington, D.C., 1906.
43. L. H. Adams and E. D. Williamson, "On the Compressibility of Minerals and Rocks at High Pressures," *J. Franklin Inst.* **195,** 474–529 (1923).
44. F. Birch, "The Velocity of Compressional Waves in Rocks to 10 Kilobars, Part 1," *J. Geophys. Res.* **65,** 1083–1102 (1960).
45. G. Simmons, "Velocity of Shear Waves in Rock to 10 Kilobars, Part 1," *J. Geophys. Res.* **69,** 1123–1130 (1964).
46. F. Birch, "The Velocity of Compressional Waves in Rocks to 10 Kilobars, Part 2," *J. Geophys. Res.* **66,** 2199–2224 (1961).



James M. Coleman is Director of the Coastal Studies Institute and Professor in the Department of Marine Sciences at Louisiana State University. Dr. Coleman has published more than 50 papers in the field of coastal dynamic processes, has presented numerous invited papers before professional societies, and has conducted many short courses, seminars, and field trips for industrial and professional groups in the United States and abroad. He received the A. I. Levorsen Award from the American Association of Petroleum Geologists for the best paper presented at the Coastal Association of Geological Societies in 1973 and the Louisiana State University Distinguished Research Master Award in 1976. In 1976, he was also honored as American Association of Petroleum Geologists Distinguished Lecturer and as a Distinguished Faculty Fellow of the Louisiana State University Foundation. Dr. Coleman earned B.S., M.S., and Ph.D. degrees in Geology from Louisiana State University. He is a member of the International Association for Sedimentology, the American Association of Petroleum Geologists, the Geological Society of America, the Gulf Coast Society of Economic Paleontologists and Mineralogists, and Sigma Xi.



Stephen P. Murray is the Assistant Director of the Coastal Studies Institute and an Associate Professor at Louisiana State University. He has written more than 20 articles and technical reports in the field of coastal physical oceanography and has received honors from Rutgers University, Woods Hole Oceanographic Institution, the University of Chicago, the National Science Foundation, and the National Research Council. Dr. Murray served as a lieutenant on active duty with the Army in 1950-1951. He earned a B.A. at Rutgers University, an M.S. from Louisiana State University, and a Ph.D. at the University of Chicago, where he completed his studies in Coastal Physical Oceanography. He is a member of the American Geophysical Union, the American Meteorological Society, and the Society of Sigma Xi.

COASTAL SCIENCES: RECENT ADVANCES AND FUTURE OUTLOOK

James M. Coleman and Stephen P. Murray

Coastal Studies Institute
Louisiana State University
Baton Rouge, La.

Man has been fascinated with the changing panorama of the world's coastlines from time immemorial. Coastal plains were the birthplaces of civilization, and many military conflicts have been waged for control of these productive, but normally inhospitable regions. The world's shorelines are a unique boundary, separating three domains: the land, the sea, and the atmosphere. Although the coastal plains and adjacent shallow continental shelves comprise only 5% of the area of the globe, the 450,000-km shoreline displays a wide variety of settings and is complicated by a number of interacting driving forces, such as winds, waves, currents, and tides.

The importance of the shallow-water continental margin came into focus as a result of World War II operations, and postwar research programs were supported strongly by governmental agencies and industry. Recently interest has peaked again, not only for basic scientific reasons but, also, because of the high potential for resource development on the continental shelves and the location of large industrial complexes next to huge new harbors, the national concern for proper management and environmental maintenance of the coastal zone, and the far-reaching consequences of international legal agreements on territorial rights.

The coastal domain can be defined in scientific terms as that region, both inland and seaward of the shoreline, in which the properties characteristic of the air-sea-land interface exert significant control on environmental conditions. Mr. W. Tolbert of the Naval Coastal Systems Laboratory gives the following operational definition to the same region: "The coastal domain extends from that region offshore where forces designed for open oceanic operations lose, through system degradation or tactical restraints a significant portion of offensive or defensive capabilities and inland to that region where forces afloat can no longer directly, excluding aerial support, provide effective combat support." In neither case is the definition limited by specific distances, water depth, or geographic features.

The coastal domain is composed of the coastal plain and the continental shelf and waters that cover it. The most common concept of a shoreline, to most observers, is the sandy strip that borders a land mass; however, on a global basis shorelines and continental shelves display a high degree of variability. Along rocky headland coasts, such as those that border the California coast of the United States, Spain, Norway, and much of the western South American coast, the coastal domain is extremely narrow. Oceanic wave spectra and ocean currents in these areas continue unmodified to within 1 km of the shoreline, and offshore weather patterns and airflow seem not to be affected until the high-relief

features of the immediate shoreline are encountered. Along low-relief muddy coastlines (Guianas, Gulf of Po Hai, etc.), however, offshore waters are shallow and offshore bottom slopes are low. Ocean-generated waves and currents are drastically modified as they propagate across the broad, shallow region. Oceanic and continental weather patterns, influenced by the heat balance of the broad shallow-water area and low-lying coastal plain, are also drastically modified as they enter the region, and local mesoscale weather patterns develop. Between these extremes, several other basic types can be defined. Thus, although the concept or variability in the Earth and ocean sciences is not new, geographic and temporal variability of processes and landforms reaches an extreme at the air-land-sea boundary.

Figure 1 illustrates schematic profiles across different coastal settings. Mud-bound coasts (Fig. 1A) span all latitudes of the earth and in the Americas constitute some 23 percent of the shoreline length; offshore slopes are low, and the width of the continental shelf ranges from 100 to 150 km. The shoreline normally displays broad, muddy tidal flats that are backed on the landward side by marshes, mangrove swamps, dike-protected agricultural plains (in tropical and temperate climates), or broad salt flats barren of vegetation (in arid climates). Relief in the coastal plain is normally low, and the plain may extend inland for distances up to 150-200 km. Quite often tidal creeks and local drainage channels form a complex maze across the plain.

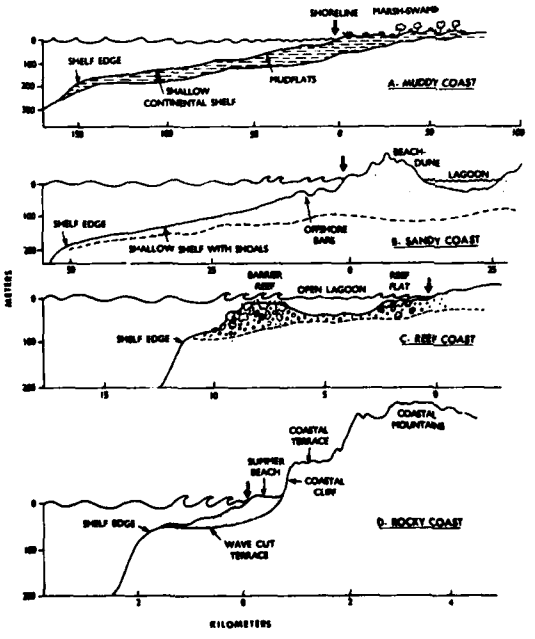


Figure 1–Schematic cross sections of various coastal settings

Sandy-beach shorelines (Fig. 1B) have received the highest amount of scientific attention in the past decades. In these settings, the shoreline is marked by an accumulation of sand and is constantly undergoing change. Beach width and slope change throughout the year in response to differing wave intensities and sediment supply. Some beaches display large eolian (wind-formed) sand dune fields immediately landward of the shoreline, whereas in regions of low sediment supply and low wind intensities dunes are absent. Seaward of the shoreline is a region commonly referred to as the surf zone, where incoming waves shoal rapidly and break. The process of breaking results in extreme energy dissipation and causes a near-constant motion of the sandy sediments. The interaction of waves and currents with bottom sediments produces a wide variety of nearshore topographic features (Fig. 1B) commonly referred to as offshore bars. These bars, displaying complex and varying patterns in different regions, persistently undergo changes in shape and magnitude. Seaward of the surf zone, the continental shelf slopes to a depth of approximately 200 m, where the shelf edge lies. The distance to the shelf edge is highly variable, but off sandy coasts continental shelf widths range from a few tens of kilometers to over 150 km. The sandy beach may abut directly against the coastal plain or it may be separated from the plain by shallow lagoons, in which case the offshore sand island is referred to as a barrier island.

Reef-bound coasts (Fig. 1C) are most common along tropical and temperate continental margins and along the shorelines of tropical islands in the Pacific, Atlantic, and Indian Oceans and the Caribbean Sea. In these settings coral and other biological assemblages dominate the shoreline. In most instances these continental shelves are narrow, and, seaward from the reef crest, oceanic depths are encountered within a few kilometers.

Ocean waves impinge and violently break on the reef crest, expending considerable amounts of energy. The reef crest may directly abut the continent or be separated from it by open lagoons, sounds, or tidal flats. The sounds are normally areas of quiet water and are floored with scattered reef patches or carbonate muds and sands. Topographic features or bottom roughness elements display extreme variations and as a result modify incoming waves and oceanic currents. Few other environments exhibit such drastic changes in process intensity as commonly occur in reefs.

Rocky or cliffed coasts (Figure 1D) make up a substantial part of the world's shorelines, but in terms of research investigations they have received a minimum of effort. Coastal mountains or high-relief elements very commonly are present adjacent to the shoreline, and within a few kilometers elevations may reach several hundred meters. Small beaches of relatively coarse material often accumulate at the foot of coastal cliffs. Beaches associated with rocky cliffs commonly display significant variations in width, slope, and topography on a seasonal basis. Normally a wave-cut terrace is present in the offshore area; sometimes it is covered by a thin veneer of sediments, but often the bare rock is littered with debris. The continental shelf is extremely narrow, and often within a kilometer the ocean bottom plunges to oceanic depths.

These four examples, although not inclusive of all coastal settings, serve to illustrate the inherent variability associated with the air-sea-land boundary. The extreme spatial and temporal changes in both processes and landforms that occur in the coastal regions are perhaps not equaled in complexity in any other environmental setting. Thus it is significant that in the past few decades research efforts have been increasingly oriented toward interdisciplinary studies on a wide variety of domestic and foreign shores. Research along foreign shorelines is of particular importance because domestic shorelines display only a limited number of coastal types. Increasing mobility of naval forces, electronic sophistication of armaments, and the rapidly changing political climate in the world demand that a better understanding of coastal processes be forthcoming so that required environmental support data will be available on a timely and global basis. Significant advances in coastal sciences have been made in the past 30 years, and milestones in process and landform studies will be summarized in the following sections.

COASTAL PROCESSES

Coastal science has evolved rapidly in the past few years with the recognition that it is a suite of processes, occurring at different intensities in different environments at reasonably distinct time and length scales, that exerts control on the movement and arrangement of water masses and sediment particles in coastal waters.

Physical processes operating in the coastal and shallow-water regions of the world can be conveniently categorized in three general areas. The first deals with atmospheric motions, while the other two, wave motions and current motions, are active in the water column itself. Obviously there are close and inseparable ties and feedback mechanisms that link these three categories of motion, but their initial separation allows an assessment of their relative importance to future research.

Atmospheric Motion

Meteorologists have long recognized concentrations of energy at various time scales in the atmosphere. Figure 2 shows a schematic distribution of familiar meteorological phenomena, most of which strongly affect the coastal zone. Turbulence and gustiness (or unsteadiness) in the airflow have been extensively studied for their role in the air-sea transfer processes. Although these phenomena are studied at scales of seconds and centimeters, their universal occurrence makes them critical parameters in wave generation and dissipation and the heating and cooling of surface waters. Unfortunately, most of the research to date [1] has assumed horizontally homogeneous conditions such as rarely occur in the coastal zone. Notable exceptions are the studies of Panofsky and Petersen [2] and Hsu [3] who investigated the effects on the wind profile of abrupt surface roughness changes such as occur at the shoreline.

Organized atmospheric motions between 100 m and 1 km, usually exemplified by tornadoes, appear at present to have little impact on coastal processes. In reference to the longer scales shown in Figure 2, we know of no serious attempt to determine the response of the coast and inshore waters to be thunderstorm systems so ubiquitous throughout most of the tropical and subtropical regions of the world. Our knowledge of coastal processes is still strongly colored by original studies along midlatitude European and American coasts, where thunderstorms are usually only at noise level, yet conditions over large stretches of coast in Asia may be controlled by incessant barrages of thunderstorm activity.

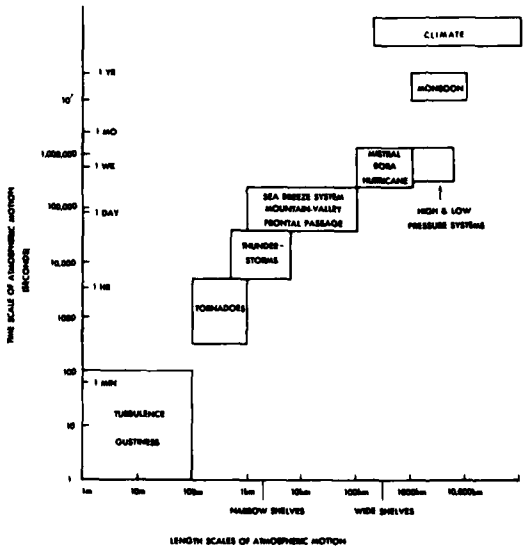


Figure 2–Time-length scales of atmospheric motions

The daily cycle of the sea-breeze system (Figure 2) is an important process operating on length scales of 10-100 km both along the coast and perpendicular to it. Studies by Johnson and O'Brien [4] and Hsu [5] show the mechanics of the airflow to be well understood, but only recently have Sonu et al. [6] documented the surprisingly coherent response of nearshore currents, waves, and beaches to sea breeze forcing. In many parts of the world's coast not yet studied (e.g., Chile), sea breeze conditions routinely reach storm levels and probably are the dominant coastal driving forces.

Pronounced local daily circulations occur where land-sea breeze and upslope-downslope wind regimes reinforce each other. An example is the Red Sea-Gulf of Aden system. Steep escarpments rise abruptly from the coastline in this region and generally fair skies allow a maximum of surface heating. During the morning, downslope-land breezes prevail, causing convergence over the water; by afternoon upslope-sea breezes prevail, causing divergence over the water and convergence at selected locations over the land. Rainfall and vegetation patterns are known to be controlled by this phenomenon, and it is extremely likely, although as yet unverified, that some coastal processes are dominated by this same system. The Pacific coast of Guatemala and the Mediterranean coast of Asia Minor have strong wind systems of this type and time scale [7].

Frontal passages are meteorological events that also operate at these time and length scales in middle and subtropical latitutdes. The impact on coastal and shelf waters of these abrupt wind shifts and intense air temperature changes associated with frontal passages is only beginning to be appreciated. Nowlin and Parker [8] reported on the chilling of shelf water after a frontal passage. The effect of a passing front on shallower coastal water will be even more dramatic and may even lead to conventional overturning, as observed in lakes.

In contrast to our ignorance of the smaller scale systems, tropical cyclones (hurricanes and typhoons) have been recognized as discrete forcing agents capable of modifying the water mass characteristics and current and wave fields over distances of hundreds of kilometers [9, 10]. Forristall [11] has studied the currents produced on the shelf by hurricane winds using the numerical model approach, and predicts speeds in excess of 3 knots.

Studies of large-scale local wind effects such as the mistral have shown how these intense winds, blowing down the Rhone and adjacent valleys, affect the deep waters of the Mediterranean over scales of hundreds of kilometers [12]. Coastal waters must also be strongly influenced by the mistral, both directly and in response to the temp-

erature transformation of the offshore water. Along the coast of the Adriatic Sea the bora, a large-scale local wind effect very similar to the mistral, is well recognized as a controlling agent of local and nearshore weather. Similar gravity-driven winds, draining off the glaciers in polar regions, are well known for nearly instantaneous initiation of severe weather conditions [13]. The future challenge is clearly to document and understand the control that these mesoscale weather phenomena exert on processes operating along coasts.

Operating over the same time scales but coherent over greater distances—thousands of kilometers—are the familiar high and low-pressure systems (anticyclones and cyclones) that control the weather in much of the middle and higher latitudes, especially in the winter half of the year. The systematic wind shifts and accelerations over large distances have allowed a good understanding of the response of beaches, waves, and inshore currents to the passage of these large and frequently intense systems, both on large continental lakes and exposed ocean shorelines [14]. Waves generated by these storms produce the sea and swell that have been rigorously investigated over the last 30 years by D. L. Inman and his associates in California. Much of our knowledge of wave mechanics and beach processes has emanated from this effort [15].

At the next larger time and space scales, monsoons, such as the East Indian monsoon, the North Australian monsoon, and the reversing Indian Ocean monsoon, are well known climatologically and meteorologically, but their impact on nearshore currents, inshore wave fields, and resultant coastal responses remains unknown and unidentified. Climate, the longest time scale in which the atmosphere operates, has traditionally involved qualitative and statistical studies, however, recent research by Resio and Hayden [16, 17] indicates promising progress in relating some of the mechanics controlling climate (such as Jet Stream behavior) to large-scale and long-term patterns in coastal processes.

In retrospect, despite the variety of atmospheric motion shown to be impacting the coastal environment, only the effects of large-scale synoptic pressure systems, and to a lesser extent the sea breeze system, have been thoroughly investigated. The significance of the impact of atmospheric forcing on other scales of motion needs considerable attention.

Current Motion

Currents in coastal waters similarly can be understood in terms of their occurrence at quasi-discrete time and length scales (Figure 3). Turbulence occurring at scales of centimeters and seconds is responsible for the diffusion of momentum, salt, heat, and sediment particles throughout the water column. Since the pioneering study by Bowden and Fairbairn [18], progress has been slow but steady in this difficult field. Niiler [19] notes that the parameterization of turbulent diffusion in nearshore waters is still poorly known, despite its importance to most theoretical and numerical models of shelf current systems.

Currents associated with wave orbital motions occur over distances of several meters and persist for the period of the wave, say 3-20s. Research in this instrumentally difficult area proceeded quite slowly until the recent advent of reliable electromagnetic current meters. Miller and Zeigler [20] successfully measured the internal velocity field in shoaling waves and identified distinct types of velocity fields associated with classes of wave geometry. Thornton and Krapohl [21] carried this work further by reporting velocity measurements and delineating their correlation to surface wave heights. Under swell conditions good

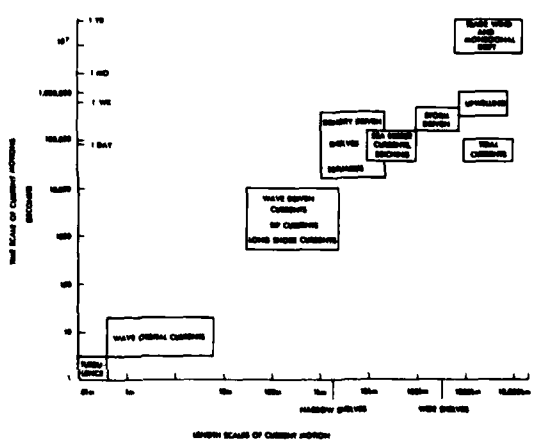


Figure 3–Time-length scales of current motions

agreement was seen between observation and linear theory.

The shoaling and breaking of the waves encroaching upon a shore generate other types of current that exhibit larger time and length scales than the waves themselves (Figure 3). Water carried onshore in the shoaling waves first moves along the shore (longshore currents) and then at specific locations returns offshore through the mechanism of the narrow seaward-flowing rip currents. The most significant advances in this field, since the pioneer work in the early 1950s, have arisen from the recent application of radiation stress concepts [22]. This approach has significantly increased our knowledge of the physics controlling longshore currents and rip currents, but much dependence is still placed upon the choice of eddy viscosity and bottom friction coefficients. Extension of this line of work showed that long waves trapped against the coast and traveling perpendicular to it (edge waves) [23] interact with incident waves to produce rip currents.

Sonu [24] on the other hand, presented detailed observations and calculations of longshore and rip currents on a beach exhibiting natural irregularities in its surface (rhythmic topography) and argued that rip currents were highly dependent on the bottom topography. Sonu's conclusion is supported by Noda [25], who combined the radiation stress concepts of wave-driven current with a complex bottom topography in a numerical model that successfully reproduced the basic points of the field observations. In a different vein, Dalrymple [26] shows that rip currents can also be generated by the intersection of wave trains of the same frequency arising either from different wave-generating systems or reflection from the beach of the incident waves. Because much progress has been made on this problem on sandy coasts, future work likely will turn to determining the basic characteristics of wave-driven currents along (a) long stretches of muddy coast around the world where energy loss caused by bottom friction appears to be significantly higher than on sandy coasts, (b) rocky coasts where abrupt shoaling produces almost instantaneous breaking and turbulence levels that are likely to be extreme, and (c) arctic coasts where offshore pack ice produces fetch-limited conditions and the freezing of slush ice on the surface makes the inshore water resemble a highly viscous suspension.

Currents driven by horizontal gradients in the density field (baroclinic effects) are especially important in coastal waters as a result of the lighter fresh to brackish water brought in the river, estuary, and bay effluents. Vertical mixing of the introduced continental runoff sets up shoreward pressure gradients and resultant flow in the bottom shelf waters, as evidenced by the circulation patterns found by Bumpus [27] and Harrison et al. [28], whose work shows shelf water moving up toward the coast and entering the estuaries. Density-driven currents are an essential part of estuarine circulations [29], and Gibbs [30] reported on two-layered estuarinelike circulation patterns off the mouth of the Amazon. A very interesting new aspect related to density gradients is the effect of sudden cooling of shallow shelf waters after cold front passages. The increase in density may lead to cascading over the shelf edge, as suggested by Stefansson et al. [31]. This process should be very amenable to investigation by time series satellite remote sensing. Similarly, it is quite possible that superheated, highly evaporated waters on shallow banks will reach sufficient density from evaporation to cascade off the banks, flow down the slope, and set up a thermal discontinuity at its equilibrium depth.

Currents driven by wind of the sea breeze systems exhibit a daily cycle and can be the dominant mode of current activity along low-tide coasts [6]. Murray [32] has shown how small vertical variations in the density profile can have a marked effect on the structure of the wind-driven current along the shore. Blanton (1974) showed that nearshore currents reversed about 6 after a wind shift, but farther offshore the reversal lagged by about 12 in summer and 36 in fall. The implication is that thermal structure is playing a strong role in the coastal dynamics.

Currents associated with seiching in harbors or lakes or on shelves are generally considered small in magnitude compared to tidal and wind-driven currents, but future studies will undoubtedly show specific cases and localities where such currents exert a dominant environmental control. Investigations of baroclinic effects in the coastal boundary layer have led to the search for high-

speed zones trapped along the coast (coastal jet), as predicted by Csanady in 1975 [33]. Similar effects emerge from Walin's study of the atmospheric forcing of stratified coastal waters in the Baltic Sea [34].

Storms associated with the midlatitude migrating high- and low-pressure systems drive intense though obviously intermittent currents along most of the shorelines and shelf area they traverse. Beardsley and Butman [35] found that intense but relatively short-duration wind events dominate the circulation over the New England shelf and can even account for most of the net flow. Murray [36] made very similar observations for the storm-driven flow over a section of the Gulf of Mexico shelf. Intense wind-driven cross-shelf motions have been shown by drogue tracks to traverse the total width of the shelf off the Delaware capes during cyclonic storms [37]. Cannon [38] has shown that currents in a Pacific Coast submarine canyon also respond strongly to the passage of pressure systems across the shelf and onto the coast. Shepard et al. [39] measured strong currents, along the axes of submarine canyons, that they related to progressive internal waves. Cross-canyon currents were found to be due to both cross-canyon winds and tidal forcing. Onshore winds, incident waves, standing edge waves, and gravitational driving by suspended sediment are also cited by Inman et al. [40] as important to the generation of currents in submarine canyons. Future work at these scales should investigate the frequency and magnitude of currents induced by the adjustment of the water mass to the atmospheric pressure gradients along the coast as the moving pressure cell intercepts the shoreline.

Tidal currents occur regularly at daily and semidaily tide scales, and although they exhibit length scales covering thousands of kilometers, they are frequently controlled by local coastal and shelf topography. Hendershott and Speranza [41] have examined theoretically the role of Coriolis force in producing rotary tidal currents in large bays of simple outline. Hart [42] used numerical techniques to study the current field in a tide-dominated sound that has two entrances separated by more than 100 km. Differences in amplitude and phase between the two entrances exert great control on the spatial variations of the currents. Hart's calculation of current flow agreed very well with NASA high-altitude imagery flown over the sound. Thus, coordination of remote sensing and numerical modeling may be a valuable technique in the future.

In a recent review article, Niiler [19] reported that upwelling, defined as a periodic intrusion of deep, fertile midocean water onto the continental shelf, with a compensating offshore flow of surface water, is now such a widely observed occurrence that it is probably an inherent part of shelf circulation. In a few regions, the intrusion is dramatic and generates intense biological productivity (e.g., off Oregon and California). Off the East Coast of the United States the water upwelled in summer is confined below a strong seasonal thermocline until strong vertical mixing characteristics of winter ensue [43]. During upwelling events off Oregon, surface waters are blown offshore to the right of a recently intensified wind and a broad inflow covers the lower half of the water column, according to Johnson [44] and Halpern [45]. Observations by Komar et al. [46] and Mooers et al. [47], however, suggested the possibility of a two-celled circulation pattern during a period of strong upwelling. Thompson's [48] numerical study showed how a two-celled pattern could develop from sinking of water near the boundary of the surface front. In the most recent contribution in a rather extensive collection of numerical studies devoted to the upwelling problem, Peffley and O'Brien [49] showed the importance of real bathymetry to understanding local characeristics of the upwelling cycle off Oregon.

At the largest scale of motion with respect to coastal processes, currents driven by trade winds and monsoons impact on much of the world's coast. Roberts et al. [50] have shown how such currents interact with and even dominate wave processes on a narrow island shelf. Lee [51] showed that when a major boundary current meanders near a narrow shelf (e.g., the Florida Current off Boca Raton), cyclonic spinoff eddies (10-30 km in scale) can ride up completely over the shelf and dominate the circulation. Lee showed that such eddies advect heat and salt into the coastal region and effectively flush the inshore waters as they are translated northerly at speeds of about 25 cm/s. Further studies to determine the mechanics of generation, their spatial and tem-

poral distribution, and decay time are recommended. Bang and Andrews [52] have also shown that an intense frontal system meanders and sheds eddies onto the continental shelf off southwest Africa.

Currents within a few kilometers of the coast can be greatly influenced by monsoonal wind patterns. Leetmaa and Truesdale [53] showed a quick reversal and broadening of the Somali Current off the east coast of Africa after the shift from the northeast to the southeast monsoon. A few of their data points approach the coast (where a current maximum is seen), but the implications to coastal processes are not drawn and must remain for future research.

Wave Motion

A major form of energy transported from the deep oceans into the coastal zone consists of periodic motions that transmit energy into the region without a significant input of oceanic water mass. The periodic motions span a frequency wavelength range of some nine orders of magnitude (Figure 4). At the short-wavelength, high-frequency end of the wave spectrum the capillary and short gravity waves occur. These waves contain very low levels of energy and do not significantly or directly affect coastal processes; however, they do play an important role in the transfer of momentum from wind stress into the water column. Recently, interest in these waves has increased because of the possibility of using high-frequency surface waves as indicators of surface windspeed (i.e., Marks and Stacy [54]).

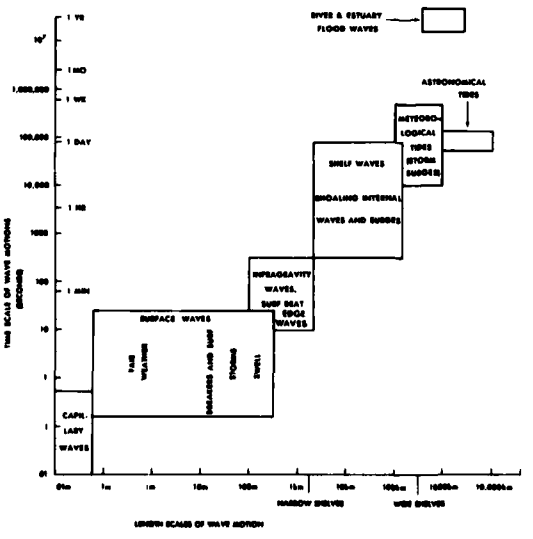


Figure 4–Time-length scales of wave motions

Surface waves with periods of between 1 and 15 s include wind waves, surf, and swell. This part of the spectrum contains the most significant amounts of energy affecting coastal processes. Surface waves cause mixing of nearshore waters and produce the major movements of coastal sediments. Collins [55] reviewed shallow-water wave spectral changes and concluded that, for open, relatively straight sections of coast, wave processes outside the nearshore zone are reasonably predicted on the basis of considerations of wave refraction, wave shoaling, and energy gain from the wind and loss to bottom friction. Results of ongoing research, however, indicate that other processes on the shelf may be strongly affecting wave propagation. These processes include wave scattering, wave-wave interaction and wave-bottom interaction.

The surf zone presents special problems for wave research. The theoretical approaches available for studying the problem are generally inadequate to deal with the basic properties of waves. Much recent effort has been made to improve the ability to measure surf zone wave characteristics to determine the effect of the wave-front celerity on run-up.

The results of these recent studies of the surf zone indicate that several processes are at work. Suhayda and Pettigrew [56] have shown that within the surf zone waves undergo transitions between bore and nonbore motions that are a function of nearshore bathymetry. Sawaragi and Iwata [57] have indicated that the energy lost in breaking goes primarily into turbulence, and little goes into bottom friction. Thus wave height in the surf zone is determined strongly by distance from the break point. Dingler [58] extensively studied the conditions controlling ripple formations as a function of bed material and flow conditions. He established criteria for the onset of grain motion

and the transfer from vortex ripple flow regime to sheet flow. Suhayda [59] has shown that run-up on beaches is dominated by low-steepness wave components, which can produce standing waves in the surf zone. Wood [60] has documented the large variability of wave celerity and wave height at a point in the surf zone over short intervals of time. The interaction of wave sand currents is recognized as an important but difficult problem; Dalrymple [26] recently presented a model for surface waves propagating over a linear shear current in which the interaction produces changes in both wavelength and wave shape. In another interaction study the concentration of suspended sediment by wave effects on the bottom has been investigated by Liang and Wang [61], who related the concentration through a power law to particle settling velocity, fluid particle velocity, and wave number.

Waves having a period between 30 sec. and 10 min., called infragravity waves, includes surf beat, edge waves, and tsunami. These waves have been studied recently because of the apparent role they play in beach dynamics. Suhayda [59] indicated that infragravity waves dominate run-up on beaches. These waves have been found to interact strongly with beach bathymetry [62] and appear to be possibly generated in the inner surf zone [63]. Guza and Davis [23] described mechanisms for generating edge waves on a plane beach. Guza and Inman [64] then indicated that low-frequency waves in edge wave modes may generate beach cusps. Short [65] has shown that the positions of multiple bar crests agree well with the predicted locations of antinodes of standing infragravity waves.

At scales of kilometers and hours, shelf waves have excited considerable attention in recent years. Hamon [66] first showed the effect on sea level along the coast of shelf waves that were apparently generated by synoptic pressure systems; sea level variations travel counterclockwise around Australia at speeds of 500-800 km/day and with amplitudes of 10-20 cm. Similar observations on the Pacific Coast of the United States have been made by Cutchin and Smith [67]. Theoretical progress on shelf waves has been made [68, 69] but the role they play in affecting the coast has not yet been investigated. Winant [70] has clearly shown the importance to coastal water of shoaling internal waves or surges manifested by sudden intrusion of cold water into the nearshore area. Possible implications include onshore sediment transport and significant levels of energy disruption. Cacchione and Southard [71] described in more detail the possible sediment entrainment and transportation by shoaling internal waves, and Cacchione and Wunsch [72] discussed laboratory experiments on the breaking and mixing of internal waves on a sloping bottom (shelf). It appears that the spatial and temporal characteristics of internal motion will find considerable application to coastal problems in the next few years.

Storm surges extend over hundreds of kilometers and usually are active for several days. Considerable effort has been expended in the past in attempting to understand the catastrophic surges produced by hurricanes and extratropical cyclones [73], and recent studies have been largely oriented toward numerical work on large computers. With respect to less dramatic processes, observational and numerical studies such as that of Galt [74], who investigated surges over the shelf produced by atmospheric pressure gradients, should provide new insights into coastal currents and sea level.

Hendershott [75] recently reviewed the progress in understanding deep ocean tides, and Niiler [19] pointed out that almost all the recent work has been devoted to the internal tide problem. Observations show that internal tides are clearly important in stratified coastal waters but that the complicated mechanisms controlling the feeding of energy to the baroclinic modes by barotropic modes are still unknown. Numerical models of coastal tide unfortunately must now face the reality of inhomogeneous water masses.

At the longest time scale we consider, i.e., periods of a year, flooding of rivers and estuaries produces yearly cycles in the velocity and density structures of estuaries, as well as significant variations in water depth. Seasonal injection of fresh water into coastal and shelf areas will markedly affect the dynamics and may lead to distinct seasonal patterns in the type and/or intensity of the locally dominant coastal process. Similarly, long-recognized seasonal variations in sea level have yet to be evaluated for their possible influence on the coastal systems as a unit.

COASTAL LANDFORMS

Interacting coastal processes erode, transport, and deposit available sediments to form highly variable coastal landscapes along the world's shorelines. The inherent complexity of coastal landforms has caused considerable problems when attempts have been made to formulate a practical and usable classification of the world's coasts. Early classifications of coasts were largely descriptive [76, 77] and suffered from loose definitions of categories and imprecision in application. Later classifications by Johnson [78] and Shepard [79] were primarily genetic in nature and, because sufficient data are not available for most coastal areas, it is difficult to apply these classifications on a worldwide basis. McGill [80] and Alexander [81] presented maps of the world's shorelines showing the distribution of major coastal landforms and shore features. In 1971 Inman and Nordstrom introduced a combined genetic and descriptive scheme containing two important aspects: (a) coastal features were statistically analyzed as well as presented in map form and (b) scale problems were overcome by systematizing coasts as to levels or orders of coastal landforms. One of the more recent attempts at coastal classification was by Dolan et al [82] and Hayden and Dolan [83]. Their classification was the first major attempt at grouping coastal features on the basis of forcing processes, material response, and biological response. The resulting classification, although complex and consisting of several orders or levels, conveyed a considerable amount of information concerning any given geographic area.

These attempts at classifying the world's shoreline landforms all point to the complexity of this region, and one of their main uses is to provide a framework on which to organize the individual studies completed in specific geographic regions. Development and availability of new sensors (atmospheric, wave, currents, etc.), expanded analysis capability via computers, and remote-sensing data gathering techniques have resulted, in the last two decades, in significant advances in the study of specific phenomena in several differing coastal environments. Recent significant research results in a few coastal settings will be discussed in the following sections.

Sandy Beaches

Sandy beaches are found throughout the world in all types of climate and tidal conditions; however, the most continuous and best developed beach-barrier islands are found in regions displaying moderate wave energy, wide continental shelves, intermediate to low tidal range, and a large continuing supply of sediment. Sandy beach deposits make up approximately 20% of the total shoreline in the Americas [82], and it is estimated that 13% of the world's shorelines display sandy beaches.

Beaches are extremely dynamic areas in which both subaerial deposits and bars in the surf zone are continuously undergoing change. Although a considerable amount of research has been conducted on beaches during the past several decades, one of the most significant lines of study has been concerned with the types and morphology of offshore bars, their rates of change, and concepts relating to their origin. Offshore bars exist seaward off nearly every sandy beach around the world and display a wide variety of configurations. The most common bar configurations are shown in Figure 5. The first type (Figure 5A) consists of multiple linear, parallel bars that are not connected to the beach. Often these bars will extend alongside, unbroken, for several hundred kilometers. Bar spacing and water depth over the bars are a function of offshore wave climate, sediment supply, tidal range, and local nearshore wind fields. This bar type is normally found in regions where waves arrive nearly parallel to the coast and where offshore wave energy shows a persistence in intensity all year.

A second type of bar configuration (Figure 5B) is characterized by periodic longshore undulations in the bar, commonly referred to as "rhythmic topography." These undulations can be present in both the inner bar and the outer bar; the alongshore spacing of these rhythms varies from 500 m to several thousand meters for the outer bar and from 100 m to several hundred meters for the inner bar. The outer bar tends to form a continuous series of symmetrical curves, whereas the inner bar is often skewed and is apt to develop discontinuous crests. The outer bar changes more slowly and remains fixed for long periods. The undulations in the outer bar display movement

parallel to the shoreline, rather than onshore-offshore. The inner bar, on the other hand, is extremely sensitive to rapid changes in wave direction or intensity, and it responds quickly by changing its configuration to the new wave field. This bar configuration leads to a constantly changing bar clearance and location along the beach. Another phenomenon associated with this bar configuration is the generation of regularly spaced rip currents in the surf zone. Mass transport of water by wave action and alongshore currents inside the breaker zone results in a water pileup in the bays between the rhythmic horns; this is relieved by a strong seaward-directed current commonly called a "rip current." It is important to note that rip currents are not steady but that seaward flow fluctuates, sometimes appreciably, with differing frequencies, and that these fluctuations depend on the nearshore wave field at a given time.

The third major offshore bar configuration (Figure 5C) consists of regularly spaced en echelon bars that are attached to the shoreline. The spacing between the points of attachment to the beach varies from about 1500 m to distances on the order of 10 km. From the point of attachment the bar crest trends offshore and parallels the shoreline at distances 400-800 m seaward. This bar configuration commonly forms where low waves arrive at the coast at relatively high angles to the shore and where alongshore currents are persistent in intensity and direction. Alongshore migration of the bars is normally rapid, and rates of up to 150 m/year are not uncommon. Such rapidity of migration can quickly outdate beach reconnaissance surveys and render practically useless the maps of nearshore topography.

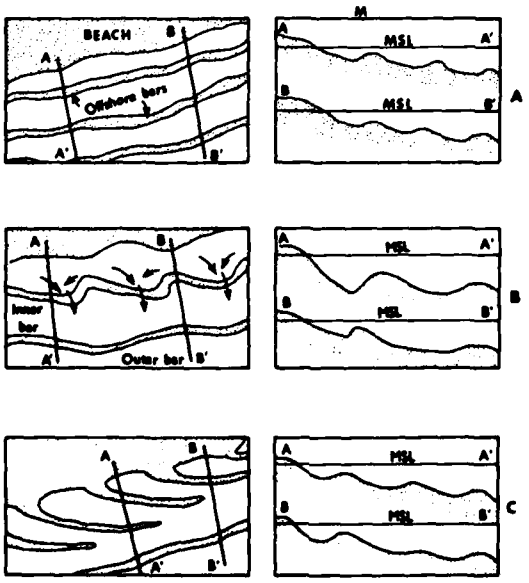


Figure 5–Generalized configurations of offshore bars. The left-hand diagrams are plan views, and the right-hand diagrams are the corresponding vertical cross sections.

The recognition of various types of nearshore bar topography and its change relative to dynamic processes has been well documented over the past 30 years. Simultaneous with the development of offshore bar morphology was research on concepts and theoretical considerations concerning the mechanisms responsible for the formation of offshore bars. In general, the formation of bars has been related to (a) mechanisms associated with breaking waves, (b) radiation stress arising from wave shoaling in the nearshore region, (c) formation of edges waves, (d) wave reflection and development of standing waves in the surf zone, and (e) formation of the bars by sediment transportation by alongshore currents. These concepts have been developed from laboratory and field studies and from theoretical considerations.

Another significant aspect of research that has evolved in the past decade from work on sandy beaches (which applies equally as well to several other coastal types) is the research conducted on aerosol generation in the nearshore region. Two types of atmospheric sea-salt particles affect the coastal zone: (a) surf-produced sea spray generated by mechanical dispersion of breaking waves and (b) aerosol generated in open ocean by bursting bubbles and carried landward by low-level winds. Use of various remote platform sensors and scanners (aircraft and satellites) has been increasing in the past decade, and the presence of low-altitude atmospheric sea salts degrades the quality of data obtained; it is therefore important that we substantially improve our understanding of aerosol-related processes.

Muddy Coasts

Muddy coasts represent a class of coastal features in which the major common attribute is that

fine-grained suspended sediment in various concentrations and electrochemical states is consistently present in nearshore waters and on the shorelines. This type of coast spans all latitutdes of the earth; however, many of the more extensive muddy coastlines are associated with and are found in the vicinity of large deltas. This coastal setting has received considerably less attention than many other coastal types, even though in the Americas it constitutes some 23% of the shoreline length. Some of the large expanses of muddy coastlines are found in the Guianas, Surinam, the Gulf of Mezen, the Gulf of Po Hai, the North Sea, India, the east and west coasts of Malaysia, and the Louisiana coast.

The landforms most commonly associated with the shoreline are (a) marsh and mangrove vegetation along the strand; (b) shell lag and organic debris accumulations; (c) broad, bare mudflats on which only low halophytic or salt-tolerant vegetation is scattered; and (d) tide flat surfaces formed by large biomass (serpulid worm reefs, shell banks, etc.).

High concentrations of fine-grained sediment in the water column and the generally high water content and weak nature of the bottom sediments cause drastic changes in coastal processes and their interaction with the bottom. The dynamic behavior of high suspended concentrations is strongly affected by electrostatic forces, and the sediment becomes subject to a different set of hydraulic flow conditions [84]. Sediment concentrations in nearshore waters display large variations in different geographic settings, as shown in Table 1, taken from Wells and Coleman (in preparation).

Table 1
Sediment Concentration

Location	*Concentration (mg/l)*		*Source*
	Maximum	*Minimum*	
Louisiana Coast	6.2×10^1	1.0×10^0	Manheim et al. [85]
East China Sea	7.0×10^1	5.0×10^0	Emery et al. [86]
Venezuela Coast	1.0×10^2	1.0×10^0	Van Andel and Postma [87]
Gulf of San Miguel	2.0×10^2	6.0×10^1	Swift and Pirie [88]
Dutch Wadden Sea	6.2×10^2	5.0×10^1	Postma [89]
Gulf of Thailand	9.7×10^2	1.0×10^0	NEDECO [90]
Gulf of Po Hai	1.0×10^3	1.0×10^2	Zenkovich [91]
British Guiana Coast	2.6×10^3	5.0×10^0	Delft Hydraulics Laboratory [92]
Surinam	3.8×10^3	1.4×10^1	Wells and Coleman (in preparation)

Rapid fluctuation in sediment concentration in nearshore waters also occurs and is a function of tide level, wave action intensity, and windspeed and direction. Figure 6 shows the variation in sediment concentration in surface waters at a site 1.2 km offshore along the muddy coast of Surinam during a portion of a tidal cycle. During the 5-h sampling period, turbidity in surface waters increased from 100 mg/l to 3800 mg/l at low tide and then decreased to 700 mg/l at the end of the sampling period. Near the bottom (often quite difficult to determine) suspensions as high as 166,000 mg/l were measured. This range in turbidity over such a short period of time is not unusual in such set-

tings and contrasts sharply with turbidity along sandy coasts, which may attain concentrations of 30 mg/l during extremely high wave action.

This amount of sediment in suspension is high enough to alter the dynamic viscosity of water. In some of the regions shown in Table 1, the viscosity during maximum suspension near the bottom would be 0.65 cm^2/s in the Gulf of Po Hai; 0.071 cm^2/s along the Surinam coast; and 0.018 cm^2/s along the Louisiana coast. Pure water at 20°C has a dynamic viscosity of 0.01 cm^2/s.

Many liquids with high suspension, such as those referred to above, do not obey the law of Newtonian fluids, and the effective viscosity itself becomes a function of the strain imposed. Krone [93] has established that San Francisco Bay muds display properties of non-Newtonian fluids. Thus the electrochemical state of the muds in suspension and on the bottom introduces conditions which may make invalid many of the exciting theories applied on "clear water" coasts to sediment erosion, transport, and deposition.

Energy dissipation of surface waves by muddy sediment-laden waters and interaction with a flexible, movable bottom reaches significant proportions in these environments. Recent work in East Bay, Mississippi River delta [94] has indicated that wave dissipation caused by interaction with a flexible mud bottom is much larger than bottom friction loss (by an order of magnitude). Simultaneous measurements at two sites in East Bay showed a 50% decrease in wave height between the two sites. Bottom friction could account for only a 5% reduction in height. Monitoring of the bottom motion during the experiment with a three-axis accelerometer indicated that movement of the mud occurred in a wavelike oscillatory fashion in response to various frequencies of surface waves and was the major factor responsible for the dissipation of waves.

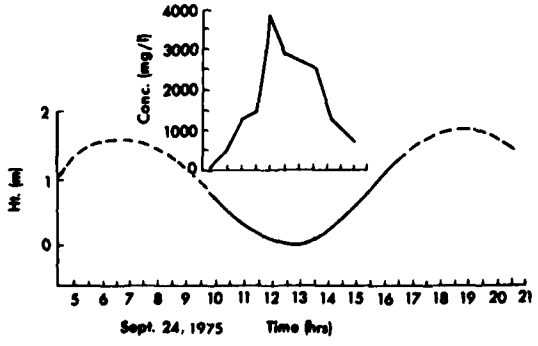


Figure 6—Sediment concentration in surface waters as a function of tide level. Site is offshore (1.2 km) Surinam.

In muddy coastal regions, standard hydraulic concepts of flow, sediment erosion and transportation, and interactions of water column and bottom must be modified considerably to include the effects of high sediment suspension and the flocculated nature of the sediments. Understanding of these dynamic interactions would then allow much greater insight into siltation problems in harbors, bottom stability, and rates of change on muddy coastal shelves.

Deltas and River Mouths

Deltas are low-lying plains composed of streamborne sediments deposited by a river at its mouth as it enters the sea. Such coastal features are widely distributed, form along the coasts of virtually every landmass on the globe, and occur in all climatic regions. Of the larger deltas in the world, 11 are located in the USSR, 7 are in Southeast Asia, 6 are in South America, 4 each are in Africa and North America, and 2 are in the Middle East. The introduction of large volumes of sediment and fresh water into marine waters and nearshore regions results in formation of highly complex coastal landforms and subaqueous topography. Rivers and river mouths provide, in many regions of the world, the only access to inland areas, and understanding the dynamics that control delta plains is critical to properly utilizing these transportation avenues. Some of the most densely populated regions in the world (for example, Bangladesh and Southeastern China) lie in delta regions because of the generally good agricultural land and abundant marine resources.

One of the initial systematic studies completed on deltas was by Samajlov [95], who discussed major deltaic processes and hydraulic regimes of river mouths and described the settings of some 65 river deltas. The most recent and comprehensive research on comparison of deltaic process and form was published by Coleman and Wright [96-97] and Wright et al. [98]. This study compared

(by various cluster and discriminant analysis techniques) several hundred process and form parameters in 55 deltas located throughout the world. Results from the study indicated that (a) attempts to classify delta landscapes on single of a few parameters were not meaningful, (b) deltas do cluster, however, into relatively discrete groups on the basis of *sets* of related morphologic or process variables, (c) delta landforms represent responses to forcing functions that are active not only within the delta but also within other component parts of a river system (drainage basin, receiving basin), and (d) the most conspicuous morphologic variations in deltas could be accounted for in terms of a few processes, such as river discharge regime, tidal range, river-mouth processes, shoreline wave energy, intensity of coastal currents, climate, and tectonics of the receiving basin.

River-mouth effluent processes are extremely important in controlling bar configurations at the mouth, water mass characteristics, and generation of density interfaces and internal waves, and also in causing major mass movement of bottom sediments downslope. River-mouth plumes respond to relative contributions of outflow inertia, turbulence, bottom friction, buoyancy, and marine forces. Effluent behavior varies significantly with river stage, and four discrete dynamic regions can be identified: (a) region 1, which extends from the mouth to four channel widths seaward, is characterized by buoyancy-dominated lateral effluent expansion, vertical thinning, and vertical entrainment of underlying saline waver; (b) region 2, which is situated over the bar, shows maximum attainment of densimetric Froude numbers, breaking of internal waves, and intense mixing; (c) region 3, lies approximately 6 to 10 channel widths seaward, Froude numbers decrease to subcritical values, and depths of the interface increases; and (d) region 4, which extends from 10 channel widths seaward to the outermost limit of the effluent, exhibits rapid expansion under the influence of buoyancy, and is subject to mixing by marine forces. The dynamics of effluent mixing and the acceleration and deceleration of the flow controls sediment dispersal and hence is responsible for river-mouth bar formation and sediment transport along the delta coast. The presence of bars of varying configurations has caused considerable navigation problems to both commercial traffic and military operations. Rapid sedimentation and erosion, commonly found at river mouths and estuaries, has caused considerable difficulty in mine warfare operations. Patterns of river-mouth bars associated with differing effluent processes have been described by Coleman and Wright [97], Nelson [99], and Wright et al. [98].

Associated with river effluents are a variety of types of density gradients and periodic phenomena (internal waves) that drastically affect acoustical transmission and reflections. Wavelike phenomena have been observed by various remote-sensing techniques (LANDSAT, high-altitude IR scanners) at the mouths of nearly all rivers. Water masses at river mouths show pronounced multiple sharp salinity interfaces and temperature steplike structures. Temperature changes of several degrees Celsius and salinity changes of 10-15 ‰ can often occur over a vertical interval of 1 m or less. Such a magnitude of density interfaces is rarely found in the deep ocean, and their presence would significantly affect sound propagation. Figure 7A shows a corrected, low-pass-filtered thermistor record taken off the mouths of the Mississippi River, and a scale expansion of the first 15 minutes of the record is shown in Figure 7B. The most conspicuous features are the high-frequency temperature oscillations that occur throughout the entire record. Peak-to-peak amplitudes of these oscillations reached 3.6 K. Power spectrum analysis of the data indicated narrow and distinct spectrum peaks with periods from 16 to 33 s. The relatively high frequency internal waves constitute the fundamental oscillations; however, it is apparent from Figure 7 that lower frequency variations in the thermal record are present. Bursts of high thermal variations occur at intervals of 10-12 min (these bursts correspond to surface expressions of wavelike phenomena seen on remotely sensed imagery); they are believed to be associated with pulsations in flow that originate within the distributary.

The hydraulic conditions operative at river mouths and effluent mixing mechanisms exert a strong influence on the pattern of sediment dispersal at the river mouths, and sedimentation rates on the shelf seaward of the delta are ex-

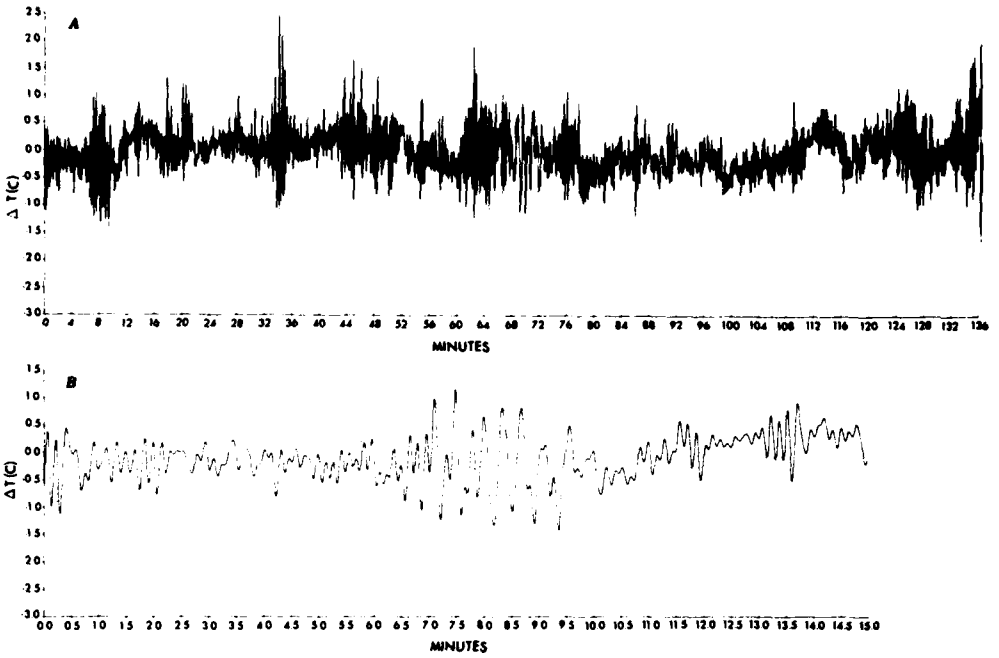


Figure 7—(A) Corrected low-pass-filtered temperature record from near pycnocline in South Pass, Mississippi River. (B) Expanded scale of first 15 min of record shown in (A).

tremely high. In the Mississippi Delta, sedimentation rates as high as 70 cm/year are not uncommon. The high rate of sedimentation does not allow pore waters to escape, and often the sediments are underconsolidated and display extremely weak shear strengths to depths in excess of 300 ft (92 m). The fine-grained clays contain high percentages of sedimentary gases, primarily methane and CO_2, which are formed by bacterial decomposition of organics [100, 101]. Methane concentrations as high as 2 ml/l have been recorded in these deposits. High methane content in the sediments caused appreciable problems. Gases in bottom sediments are a common phenomenon in many fine-grained continental shelf deposits and could severely degrade acoustic operations and render invalid the existing acoustic simulation models. Abundant gas, high water content, and weak strength also give rise to rapid and large-magnitude mass movements of marine bottom sediments. In the shallow-water portions of the delta (<100 m), rotational slumping, shallow diapiric intrusions, radial graben faulting, and sediment degassing and dewatering caused by surface-wave-produced bottom pressure perturbations cause downslope movements at rates and magnitudes that severely endanger bottom-laid or bottom-mounted structures. In deeper waters off deltas (>100 m to the upper continental slope) large-scale arcuate fault systems that cut and displace modern sediments are present on the shelf. These features have lateral dimensions of up to 10-15 km and extend from the surface to depths of several hundred meters. A second major kind of mass movement on the outer shelf and upper continental slope is represented by large, massive mudflows. The lobate seaward leading edge may extend for distances up to 40 km, and the thickness of the creeping mass of material can approach 60-70 m. Downslope movement rates of up to several hundred meters per year have been documented. This movement can be hazardous to bottom-emplaced structures or bottom-tethered objects. Mass movement of bottom sediments of differing magnitude have been documented in a large number of regions, for example off the Magdalena, Orinoco, Ganges-Brahmaputra, Nile, Niger, and Mississippi River deltas.

Coral Reefs and Atolls

Reefs built by coral and associated organisms are characteristic of tropical waters and commonly comprise a high percentage of the coastlines between latitudes 30°N and 30°S in the Pacific, Indian, and Atlantic Oceans, in the Caribbean and Red Seas, and in the Persian Gulf. The complex biological systems associated with reefs are largely controlled by temperature, salinity, turbidity, light intensity, nutrient availability, and zonation of physical processes.

A considerable amount of research on reefs has been oriented toward the biological aspects, but in the enthusiasm of discovering new aspects of how reef organisms function in their complex ecosystems, the role of physical forces on coral reefs has received little attention. Munk and Sargent [102] and von Arx [103] provided the initial investigations of dynamic processes (wave and currents) and their gross interactions with reef systems. Inman et al. [104] studied the sediment budget on the island of Kauai in the Hawaiian Islands; this was one of the first attempts to quantify the contributions of carbonates to the nearshore regime and the effects of wave action on the disperal of sediment. Reef morphology and wave processes have received some attention recently in papers by Tait [105] Hernandez and Roberts [106], and Roberts [107]. These studies show good agreement of gross geomorphic features with wave energy distribution.

The first comprehensive dynamics experiments in a fringing reef system were conducted on Grand Cayman in 1972 and in Barbados in 1973 [50]. These experiments indicated that deepwater waves are significantly modified by the high roughness elements of the reef tract as they propagate across the reef. A 20% reduction in deepwater wave height across the outer shelf resulted from the combined effects of friction, scattering, and reflection (a rate significantly greater than that occurring on sandy coasts). At the fringing reef crest, energy loss resulting from breaking produces a 75% reduction in wave height and is accompanied by substantial modification to the wave spectrum, including the introduction of multiple low-frequency peaks in the spectrum. Current measurements across the narrow fore-reef shelf show a pattern indicating strong interaction with bottom roughness elements (reef morphology). Unidirectional high-velocity currents (speeds of 50 cm/s) that have a diurnal tidal periodicity occur at the seaward edge of the fore-reef shelf. On the shallow fore-reef shelf, currents are considerably weaker (roughly 30% of the strength of those in deep shelf areas) and show a great deal more directional variability. This rapid attenuation of currents over a narrow shelf is attributed largely to lateral frictional effects associated with extreme bottom roughness. The resulting morphology showed strong correlation to individual processes operative around the reef. Evaluation of the relative importance of wave and current forces across the fore-reef shelf, based on field measurements, shows that wave forces contribute significant energy to the shallow shelf and that current forces apply a similar amount of energy to the deep shelf (Figure 8). The total force (Figure 8) across the reef shelf, therefore, is maintained at high levels, and at a depth of 21 m the combined wave-current force is the same as for a depth of approximately 3 m near the fringing reef crest. It is possible that these high current forces could be responsible for the development of the flourishing reefs that commonly occur in deep water at the margins of island shelves throughout the tropics.

Cliffed Coasts

A cliff is an abrupt break in slope; its slope is usually steep, generally greater than 15° to vertical, and its height is highly variable. Coasts with long, more or less continuous and actively changing sea cliffs vary tremendously in appearance, according to lithology, rock structure, exposures to wave attack, climatic conditions, and geomorphic history. Although estimates vary, approximately 40-42% of the world's shorelines are cliff-bound rocky coasts. A sea cliff combines a retreating cliff face, an undercut notch, and a bench that is eroded across bedrock near the shoreline but farther seaward normally becomes a depositional wave-built terrace. Cliff materials that slump off the cliff are normally transported in various ways to form small pocket beaches or are carried offshore and incorporated in the offshore wave built terrace. Marine erosion of cliffed coasts takes place mainly during storms and is

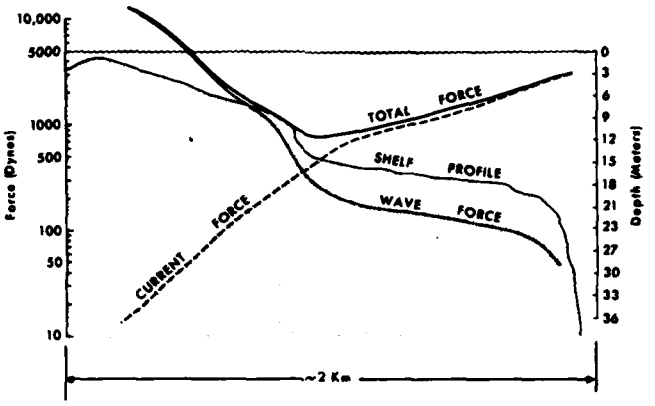


Figure 8–Distribution of wave and current forces across the fore-reef shelf of a coral reef.

achieved largely by wave action. The sheer weight of water, the hydraulic compression and release of air in pockets, joints, and cracks, and the abrasive action of water laden with rock debris all combine to produce mechanical erosion [108]. Other types of processes also play a role in cliff recession: water table processes such as leaching and differential cementation, aerosol-induced chemical weathering and breakdown of rock material, buildup of pore water pressure in sedimentary rocks leading to creep and creep rupture, freezing of pore water, and catastrophic events associated with tectonic movement.

The recent experimental and quantitative research on wave-induced cliff erosion by Horikawa and Sunamura [109] and Sunamura and Horikawa [110] is especially noteworthy. The work of these investigators along the east coast of Japan indicates average cliff recession of 0.7 m/year for the long term; submarine bedrock is being eroded downward at a rate of 0.02 m/year in very shallow water, and erosion rate decreases exponentially with increase in water depth. Material contributed to the littoral transport system from cliff recession and submarine erosion is estimated at 3.4×10^5 m^3/year, or about 24% of the total amount of material moving along the coast. This work emphasized the significant contribution of sediment supplied by cliff erosion to the nearshore waters. Such documentation points out the short-term dynamics of a coastal setting that in much of the literature is commonly thought of as a rather stable feature.

Landslides are especially important agents in contributing to cliff erosion, especially in tectonically active regions, humid climates, and regions where heavy wave action produced abundant aerosols that are carried onto the adjacent cliffs. Several lithologies are especially susceptible to landslide activity. They include layered sedimentary rocks (clays, mudstones, porous volcanics), sensitive or thixotrophic clays, and platy or foliated rocks. Heavy rainfall in the tropics and steeply dipping porous strata alternating with mudstones result in a build up of water pressure power, reduction in soil strength, and landslides ensue.

Estuaries

Estuaries, because of biological productivity, sheltered anchorages, and use as transportation arteries, have been of concern to man longer than the deep oceans. Despite this time advantage, serious research into the dynamics of estuarine circulation began only in the early 1950s, with studies centered at the Chesapeake Bay Institute and the University of Washington. As summarized in Dyer [111] much of the work to date has been concentrated in midlatitude coastal plain estuaries, where moderate tides and abundant rainfall produce the partially mixed type of estuary. The Chesapeake Bay system and the Mersey in England are probably the best understood of this type. Progress here is to the point where numerical models [112] are being used to study

long-time and large-scale variations in the salinity and velocity field. The numerical studies, however, are still severely hindered by lack of precise knowledge of the physics of the mixing and dispersion and diffusion processes in the channels.

Interest is now clearly focusing on the wide variety of estuarine types seen around the world's shorelines. Fiords have received somewhat less study than partially mixed estuaries, but the basic governing principles have been formulated [113] and they are of considerable topical interest. Gade [114] recently presented a statistical model for intermittent influx of new water into Norwegin sill fiords, and Long's [115] analysis explains the behavior of the halocline in a fiord under varying rates of freshwater influx. Further studies of the mechanics of fiord circulations are clearly warranted and should provide considerable advancements in the next few years. The large lagoon systems or bar-built estuaries typical of much of the low tidal regions of the world have, on the other hand, been largely neglected. Lee and Booth [116] provide rare insight into the processes controlling mixing and renewal of a large coastal lagoon in Florida. Wind-induced circulation, and to a linear extent tides, dominate the exchange mechanics, and renewal times of 1-3 months result. On the other hand, Kjerfve [117] also studied a wide, shallow-water body and found tidal effects to be dominant and surface wind stress to be only a modifying factor, at least during the summer regime. It appears that the surface area, windspeed, and fetch all play a critical role at times in the dynamics of these broad, shallow systems. Dyer and Ramamoorthy's [118] study of the Vellar Estuary, on the east coast of the Indian subcontinent, is especially interesting inasmuch as they describe the transition of this shallow channel from a highly stratified salt wedge type to one that is moderately stratified as the river flood drops off over a period of 25 days. An interesting numerical study, of Cienfuegos Bay, on the south coast of Cuba [119] shows the marked effect of wet season-dry season variability but omits hydrography. In the wet season, density effects resulting from river runoff control the circulation, but in the dry season a wind-driven surface current and a subsurface compensatory countercurrent, with zones of upwelling and downwelling, are present.

Other exotic density effects resulting from excess evaporation in arid regions have been studied by Bye and Whitehead [120], who explained the unusual salinity distribution in the Spencer Gulf, South Australia, with a theoretical model of flow in a narrow channel connecting two basins of water of varying density. It appears that estuarine research, though moving slowly owing to the few scientists involved in it, is approaching an understanding of the wide and fascinating variety of dynamical situations which can arise from the natural variations in topographic control, rainfall, temperature, river discharge, and tidal effects that drive the systems.

SUMMARY

Coastal sciences, as a field of research endeavor, is relatively young. Prior to the 1940s, research along the world's coastlines was conducted primarily by individual scientists who were scattered in various universities, government agencies, and private industries. Much of the work was descriptive reconnaissance. By the mid-1950s, research efforts were slightly more coordinated, and a larger number of scientists, particularly hydrodynamicists, were developing initial concepts concerning process-form interactions in nearshore regions. Two or three university-based institutes, whose major emphasis was conducting coastal research in a multidisciplinary fashion, had been formed. Several governmental agencies (Corps of Engineers, Coast and Geodetic Survey, various wildlife and fisheries groups and naval laboratories) were also concentrating research efforts in coastal regions. Funding agencies, such as the Office of Naval Research, National Science Foundation, and Corps of Engineers, were beginning to make continuing fund commitments for longer term and systematic research efforts. In the mid-1960s, a large number of institutions, university researchers, governmental agencies, and private industries, both domestic and foreign, were actively involved in various aspects of research along the world's shorelines. Development of new sensors and analysis capabilities made possible a rapid and significant advancement of our knowledge in this area. A skimpy but global data base now

existed, and efforts were beginning to be directed toward trying to explain the mechanisms operating in nearshore waters and the atmosphere. Recently (late 1960s to the present), in an era of rapid utilization of the world's shorelines, international concern for proper management and environmental maintenance of the coastal zone has led to a virtual explosion of interest by governmental agencies, private industrial concerns, and institutions that were and are involved in coastal zone research. Funding is at a record high; however, a large percentage of the funding has been used for coastal zone management projects rather than for research concerning basic mechanisms operating in the coastal region. The rapid advancement in national coastal zone management could not have been achieved in such a short period of time without the basic coastal research program that preceded it, during the past 2 or 3 decades. The Geography Programs, Office of Naval Research, played a major role in providing the opportunity and continued funding for initiating this basic research, which 20 or 25 years later has proved to be invaluable and requisite for rapid response to this national commitment.

The coastal sciences have made significant advances in the past 30 years. The early development of the concept of process response as applied to the coastal zone was a significant step. This concept built a foundation on which later researchers viewed the coastal zone as an integrated system in which there was a linkage and feedback between topography and landforms and the various interacting dynamic processes. Numerous quantitative documentations of both landforms and processes were being made by the early 1960s, and there was a general awareness of the need to measure and assess processes and forms in the many differing types of coastal settings. By the mid-1960s coastal scientists were statistically analyzing landforms to ascertain their variability, and field testing of hydraulic and atmospheric theory was actively being carried out. These activities led rapidly to the development of mathematical simulation models, which are presently being tested and modified in a wide variety of coastal environments.

A second major development in the field of coastal sciences was the realization that low-altitude atmospheric processes are greatly modified as they approach the shoreline from either the sea or the land. This basic modification to macroscale weather systems gives rise to unique microscale meteorological patterns that have dimensions on the order of 1 km vertically and 50-100 km horizontally. Coastal meterologists, therefore, realize that synoptic weather prediction techniques are not adequate to predict many coastal weather patterns. By the mid-1960s coastal meteorologists were actively engaged in field measurement programs designed to test and develop theoretical considerations for predicting microscale weather patterns. During this period the density of data required to document these changes was not readily available in many instances; however, the advent of remote data acquisition (remote-sensing imagery, data telemetry) rapidly increased the density of data available. Presently several physical models have been developed and are being applied and field tested. Undoubtedly first-approximation prediction schemes will be possible in the very near future.

A third major advance in the coastal sciences has been our ability to understand and predict coastal wave and sea-state conditions. The initial breakthrough was the application of existing spectral theory (developed in the field of pure mathematics) to ocean waves. Using this technique, coastal scientists were able to identify components of wave motion that were unique to shallow coastal waters (edge waves; surf beat; wave reflection, refraction, and defraction; etc.). Knowledge of those components rapidly improved the ability to develop predictive schemes based on a sound understanding of physical principles rather than on empirical relationships. With the advent of extensive use of the computer, wave forecasting and ship routing became a routine technique in many coastal regions. Extensions of these techniques into other coastal settings (reefs, muddy coasts, cliff coasts, etc.) are expanding our knowledge of the physical processes in these environments.

A fourth major advance was the documentation of the fact that coastal water masses display characteristics that are not simply a small-scale analogy to deep-ocean water masses. Coastal waters are characterized by mixing of water masses having extreme differences in salinity and temperature, sediment concentrations, and elec-

trochemical properites. Processes of mixing and diffusion display large variations over short time periods and small length scales. The early documentation of these phenomena led to specially designing and conducting field experiments to measure specific interactions. Model simulation followed immediately, and predictive capability in some instances has been realized. The mixing of coastal waters develops unique density interfaces and, combined with irregular bottom roughness elements, can cause severe problems in acoustic transmission and reflection. It is highly likely that future research in acoustic modeling will depend heavily on the research in coastal water masses that was conducted in the 1960s and 1970s.

Future coastal research will likely continue along lines similar to those described in this paper. There is a definite need for continuing detailed studies on specific landforms and specific processes. In addition, further studies on variability of global coastal processes and landforms need to be continued. A major research area that deserves serious attention deals with mechanisms of sediment transport in large rivers and on the continental shelves. Most of the work in the past has dealt only with sediment movement in the nearshore bar region. Mass movement of sediment on shelves is highly important and potentially detrimental to bottom-mounted or bottom-tethered systems, yet little is known concerning the mechanisms. Transport of large quantities of fine-grained mud occurs in many coastal shelf regions, and the modes of movement have not been well documented. Much of the work to date has simply been semiquantitative in nature, and concerted efforts in sediment transport would be rewarding and exciting.

A considerable amount of research has been oriented toward understanding the mechanisms responsible for forming various coastal landforms, but little research has been focused on the longevity of the features once formed or the factors responsible for their decay and deterioration. Studies oriented toward documenting the time-space history of landform deterioration would expand our knowledge concerning future utilization of the coastal zone.

Detailed studies along short stretches of coastline in the past few decades have repeatedly shown the presence of quantitatively important dynamical events that display large temporal and spatial scales and that clearly have not been generated by local winds, tides, or other local effects. In such instances, the generating mechanisms of these phenomena are not resolvable by experiments designed on the scale of a few hundred meters along the shore and a few hundred meters off the coast. Thus a major thrust of the coastal scientist in the future should be aimed toward designing larger scale field experiments to document some of these large-scale processes and their interaction with the sea bottom. Remote in-situ telemetry sensor systems and aerial remote-sensing techniques allow these scales to be studied without reliance on large, expensive oceanographic cruises using several ships. Present remote-sensing techniques, however, allow documentation of surface distribution of parameters only at a given instant in time. A major research effort in the future should be oriented toward closer coordination and combining three-dimensional simulation models with analysis of remote-sensing data. The imagery and its analysis can be used to calibrate and test the model results as well as provide data input parameters, and the model, once calibrated, can provide subsurface information concerning the water mass and can be used to interpolate parameters during those periods when remote-sensing imagery is not available.

The coastal sciences are young, yet in a relatively short period research advances have had a significant impact on planning by civilian and military strategists. The research has developed a valuable human resource, and scientists have responded efficiently and effectively to national emergencies in keeping the nation, and especially the Navy, ahead of the international research frontiers in coastal processes.

REFERENCES

1. J. C. Wyngaard, "Progress in Research on Boundary Layers and Atmospheric Turbulence," U.S. National Report, 1971-1974, to International Union of Geology and Geophysics, *Rev. Geophys. Space Phys.* **13** (3), 716–719 (1975).
2. H. A. Panofsky and E. L. Peterson, "Wind Profiles and Change of Terrain Roughness at RISO," *Quart. J. Roy. Meteorol. Soc.* **98**, 845–854 (1972).
3. S. A. Hsu, "Measurement of Shear Induced Roughness Length on a Beach," *J. Geophys. Res.* **76**, 2880–2885 (1971).
4. A. Johnson, Jr., and J. J. O'Brien, "A Study of an Oregon Sea Breeze Event," *J. Appl. Meteorol.* **12** (8), 1267–1283 (1973).
5. S. A. Hsu, "Coastal Air Circulation System: Observations and Empirical Model," *Monthly Weather Rev.* **98** (7), 487–509 (1970).
6. C. J. Sonu et al., "Sea Breeze and Coastal Processes," *EOS Trans. Am. Geophys. Union* **54** (9), 820–833 (1973).
7. G. P. Atkinson "Forecasters' Guide to Tropical Meteorology," U.S. Air Force, Air Weather Service, Tech. Rept. 240, 1971.
8. W. D. Nowlin, Jr., and C. A. Parker, "Effects of a Cold Air Outbreak on Shelf Waters of the Gulf of Mexico," *J. Phys. Oceanogr.* **4** (3), 467–486 (1974).
9. T. Ichiye, "Circulation Changes Caused by Hurricanes," *Contributions on the Physical Oceanography of the Gulf of Mexico*, L. Capurro and J. L. Reid, eds., pp. 229–258, Gulf Publishing Co., Houston, Texas., 1972.
10. T. Ichiye and H. Kuo, "Numerical Study on the Circulation and Sea Level Change of an Ocean Due to a Moving Storm," *Assessments of Currents and Hydrography of the Eastern Gulf of Mexico*, T. Ichiye et al, eds., Texas A & M Univ., Dep. of Oceanography, Rep. 601, 1974.
11. G. Z. Forristall, "Three Dimensional Structure of Storm Generated Currents," *J. Geophys. Res.* **79**, 2721–2729 (1974).
12. H. Stommel, A. Voorhis, and D. Webb, "Submarine Clouds in the Deep Ocean," *Amer. Scientist* **59** (6), 716–722 (1971).
13. W. J. Wiseman, Jr., et al., "Alaskan Arctic Coastal Processes and Morphology," Louisiana State University, Baton Rouge, Coastal Studies Institute, Tech. Rep. 149, 1973.
14. R. A. Davis, Jr., and W. T. Fox, "Coastal Processes and Nearshore Sand Bars," *J. Sediment, Petrol.* **42**, 401–412 (1972).
15. D. L. Inman and B. M. Brush, "The Coastal Challenge," *Science* **181**, 20–32 (1973).
16. D. T. Resio and B. R. Hayden, "An Integrated Model of Storm-Generated Waves," University of Va., Dep. of Environmental Sciences, Tech. Rep. 8, 273, 1973.
17. D. T. Resio, "Recent Secular Variations in Mid-Atlantic Winter Extratropical Storm Climate," *J. Appl. Meteorol.* **14** (7), 1223–1234 (1975).
18. K. F. Bowden and L. A. Fairbairn, "Measurements of Turbulent Fluctuations and Reynolds Stresses in a Tidal Current," *Proc. Soc. Lond.*, A **237**, 422–438 (1956).
19. P. Niiler, "A Report on the Continental Shelf Circulation and Coastal Upwelling," *Rev. Geophys. Space Phys.* **13** (3) (1975) (U.S. National Rep. to Internat. Union of Geodesy and Geophys.).
20. R. L. Miller and J. M. Zeigler, "The Internal Velocity Field in Breaking Waves," *Proc. Ninth Conf. on Coastal Engr.*, Am. Soc. Civil Engr., New York, 1964.
21. E. B. Thornton and R. F. Krapohl, "Water Particle Velocities Measured under Ocean Waves," *J. Geophys. Res.* **79** (6), 847–852 (1974).
22. M. S. Longuet-Higgins and R. W. Stewart, "Radiation Stress and Mass Transport on Gravity Waves with Applications to Surf Beats," *J. Fluid Mech.* **13**, 481–504 (1962).
23. R. T. Guza and R. E. Davis, "Excitation of Edge Waves by Waves Incident on a Beach," *J. Geophys. Res.* **79** (9), 1285–1291 (1974).
24. C. J. Sonu, "Field Observations of Nearshore Circulation and Meandering Currents," *J. Geophys. Res.* **77** (181), 3232–3247 (1972).
25. E. K. Noda, "Wave-Induced Nearshore Circulation," *J. Geophys. Res.* **79** (27), 4097–4106 (1974).
26. R. A. Dalrymple, "A Mechanism for Rip Current Generation on an Open Coast," *J. Geophys. Res.* **80** (24), 3483–3487 (1975).
27. D. F. Bumpus, "A Description of the Circulation on the Continental Shelf of the East Coast of the United States," in *Progress in Oceanography*, vol. 6, pp. 111–156, Pergamon Press, New York, 1973.
28. W. Harrison et al., "Circulation of Shelf Waters off the Chesapeake Bight," U.S. Dep. of Commerce, ESSA Prof. Paper 3, 82 p., 1967.
29. K. F. Bowden, "Circulation and Diffusion," in *Estuaries*, G. H. Lauff, ed., pp. 15–36, American Association for the Advancement of Science, Washington, D.C., 1967.
30. R. J. Gibbs, "Circulation in the Amazon River Estuary and Adjacent Atlantic Water," *J. Mar. Res.* **28**, 113–123 (1974).
31. U. Stefansson, L. P. Atkinson, and D. F. Bumpus, "Hydrographic Properties and Circulation of

the North Carolina Shelf and Slope Waters," *Deep Sea Res.* **18** (4), 383–420 (1971).
32. S. P. Murray, "Speeds and Trajectories of Currents Near the Coast," *J. Phys. Oceanogr.* **5** (2), 347–360 (1974).
33. G. T. Csanady, "Hydrodynamics of Large Lakes," *Ann. Rev. Fluid Mech.* **7**, 357–386 (1975), Annual Reviews, Inc., Palo Alto, Cal.
34. G. Walin, "On the Hydrographic Response to Transient Meteorological Disturbances, *Tellus* **24**, 169–186 (1972).
35. R. C. Beardsley and B. Butman, "Circulation on New England Continental Shelves: Response to Strong Winter Storms," *Geophys. Res. Lett.* **1** (4), 181–184 (1974).
36. S. P. Murray, "Observations on Wind, Tidal and Density Driven Circulation in the Vicinity of the Mississippi River Delta," in *Shelf Sediment Transport*, D. Swift, D. Duane, and O. Pilkey, eds., pp. 127-142, Dowden, Hutchinson, and Ross, Stroudsberg, Pa., 1972.
37. W. J. Wiseman, Jr., S. P. Murray, and H. H. Roberts, "High Frequency Techniques and Over-the-Horizon Radar in Coastal Research," *Proceedings of the Russell Symposium on Coastal Research*, Louisiana State University, Baton Rouge (in press).
38. G. A. Cannon, "Wind Effects on Currents in Juan de Fuca Submarine Canyon," *J. Phys. Oceanogr.* **2** (3), 281–283 (1972).
39. F. P. Shepard, N. G. Marshall, and P. A. McLoughlin, "Currents in Submarine Canyons," *Deep Sea Res.* **21**, 691–706 (1974).
40. D. L. Inman, C. E. Nordstrom, and R. E. Flick, "Currents in Submarine Canyons: An Air-Sea-Land Interaction," *Annu. Rev. Fluid Mech.* **8**, 275–310 (1976).
41. M. C. Hendershott and A. Speranza, "Co-oscillating Tides in Long Narrow Bays: The Taylor Problem Revisited," *Deep Sea Res.* **18** (10), 959–980 (1971).
42. W. E. Hart, "A Numerical Study of Currents, Circulation, and Surface Elevations in Chandeleur and Breton Sounds, Louisiana," Louisiana State Univ., Ph.D. dissertation, 140 p., 1976.
43. W. C. Boicourt, "The Circulation of Water on the Continental Shelf from Chesapeake Bay to Cape Hatteras," Johns Hopkins Univ., Ph.D. dissertation, 197 p., 1973.
44. D. R. Johnson, "Relationship Between Currents and Hydrographic Fields in a Small Region within the Coastal Upwelling System" (abstract), *Trans Am. Geophys. Union* **56**, 12 (1974).
45. D. Halpern, "Summertime Surface Diurnal Period Winds Measured Over an Upwelling Region near the Oregon Coast," *J. Phys. Res.* **79** (15), 2223–2230 (1974).
46. P. D. Komar, L. D. Kulm, and J. C. Harlett, "Observations and Analysis of Bottom Turbid Layers on the Oregon Continental Shelf," *J. Geol.* **82**, 104–111 (1974).
47. C. N. K. Mooers, C. A. Collins, and R. L. Smith, "The Dynamic Structure of the Frontal Zone in the Coastal Upwelling Region off Oregon," *J. Phys. Oceanogr.* **6** (1), 3–21 (1976).
48. J. D. Thompson, "The Coastal Upwelling Cycle on a β-Plane: Hydrodynamics and Thermodynamics," Florida State Univ., Ph.D. dissertation, 141 p. 1974.
49. M. B. Peffley and J. J. O'Brien, "A Three-Dimensional Simulation of Coastal Upwelling off Oregon," *J. Phys. Oceanogr.* **6** (2), 164–180 (1976).
50. H. H. Roberts, S. P. Murray, and J. N. Suhayda, "Physical Processes in a Fringing Reef System," *J. Mar. Res.* **33**, 233–260 (1975).
51. T. Lee, "Florida Current Spin Off Eddies," *Deep Sea Res.* **22** (11), 753–766 (1975).
52. N. D. Bang and W. R. H. Andrews, "Direct Current Measurements of a Shelf-Edge Frontal Jet in the Southern Benguela System," *J. Mar. Res.* **32** (3), 405–417 (1974).
53. A. Leetman and V. Truesdale, Changes in the Currents in 1970 off the East African Coast with the Onset of the Southeast Monsoon," *J. Geophys. Res.* **77**, 3281 (1972).
54. W. Marks and R. Stacy, "Prediction Models for Correlation of Laser Sea Return with Wind Profile," *Proc. Am. Soc. Photogram., Oct. 2-5, 1973, part II*, pp. 737–759 (1973).
55. J. I. Collins, "Prediction of Shallow-Water Spectra," *J. Geophys. Rec.* **77** (15), 2693–2707 (1972).
56. J. N. Suhayda and N. R. Pettigrew, "Observation of Wave Height and Wave Celerity in the Surf Zone," *J. Geophys. Res.* (in press).
57. T. Sawaragi and K. Iwata, "Wave Deformation after Breaking," *Proceedings of the 14th Conference on Coastal Engineering*, Am. Soc. Civil Engr., June 24-29, Copenhagen, pp. 481–499 (1974).
58. J. R. Dingler, "Wave-Formed Ripples in Nearshore Sands," Scripps Institute of Oceanography, Ph.D. dissertation, 136 p., 1974.
59. J. N. Suhayda, "Standing Waves on Beaches," *J. Geophys. Res.* **79** (21), 3065–3071 (1974).
60. W. Wood, "Wave Analysis System for the Breaker Zone," *Proceedings of the International*

Symposium on Ocean Wave Measurement and Analysis, Sep. 9-11, New Orleans, La., vol. 1, pp. 774–789, Amer. Soc. Civil Engr., New York, 1974.
61. S. S. Liang and H. Wang, "Sediment Transport on Random Waves," University of Delaware, College of Marine Studies, Tech. Rep. 26, 1973.
62. J. N. Suhayda, "Determining Nearshore Infragravity Wave Spectra," *Proceedings of the International Symposium on Ocean Wave Measurement and Analysis*, New Orleans, Sep. 9-11, New York, 1974.
63. E. Waddell, "Dynamics of Swash and Implication to Beach Response," Louisiana State University, Baton Rouge, Coastal Studies Institute, Tech. Rep. 139, 49 p., 1973.
64. R. T. Guza and D. Inman, "Edge Waves and Beach Cusps," *J. Geophys. Res.* **80** (21), 2997–3012 (1975).
65. A. D. Short, "Multiple Offshore Bars and Standing Waves," *J. Geophys. Res.* **80** (27), 3838–3840 (1975).
66. B. V. Hamon, "Continental Shelf Waves and the Effects of Atmospheric Pressure and Wind-Stress on Sea Level," *J. Geophys. Res.* **71**, 2883–2893 (1966).
67. D. L. Cuthin and R. L. Smith, "Continental Shelf Waves: Low-Frequency Variations in Sea Level and Currents over the Oregon Continental Shelf," *J. Phys. Oceanogr.* **3** (1), 73–82 (1973).
68. J. K. Adams and V. T. Buchwald, "The Propagation of Continental Shelf Waves," *Proc. R. Soc. Lond. A* **305**, 235–250 (1968).
69. A. E. Gill and E. H. Schumann, "The Generation of Long Shelf Waves by the Wind," *J. Phys. Oceanogr.* **4** (1), 83–90 (1974).
70. C. D. Winant, "Internal Surges in Coastal Water," *J. Geophys. Res.* **79**, 4523–4526 (1974).
71. D. A. Cacchione and J. B. Southard, "Incipient Sediment Movement by Shoaling Internal Gravity Waves," *J. Geophys. Res.* **79**, 2237-2242 (1974).
72. D. A. Cacchione and C. Wunsch, "Experimental Study of Internal Waves over a Slope," *J. Fluid Mech.* **66**, 223-239 (1974).
73. C. L. Bretschneider, "Storm Surges," *Advances in Hydrosci.* **4**, 341–417 (1967).
74. J. A. Galt, "A Numerical Investigation of Pressure-Induced Storm Surges over the Continental Shelf," *J. Phys. Oceanogr.* **1** (2), 82–91 (1971).
75. M. C. Hendershott, "Ocean Tides," *Trans. Am. Geophys. Union* **54** (1), 76–86 (1973).
76. A. Penck, "Morphologie der Erdoberfläche," Bibliothek g. Handbücher, herausgegeb. v. Fr. Ratzel, J. Engelhorn, Stuttgart 8°. I. Bd. XIV u. 471, S. II. Bd. X u. 969 S., 1894.
77. W. M. Davis, *Physical Geography*, Ginn and Co., Boston, 1898.
78. D. W. Johnson, *Shore Processes and Shoreline Development*, Wiley, New York, 1919.
79. F. P. Shepard, "Revised Classification of Marine Shorelines," *J. Geol.* **45**, 602–624 (1937).
80. J. T. McGill, "Map of Coastal Landforms of the World," *Geog. Rev.* **48**, 420–405 (1958).
81. C. S. Alexander, "A Method of Descriptive Shore Classification and Mapping as Applied to the Northeast Coast of Tanganyika," *Ann. Amer. Ass. Geogr.* **56** (1), 128-140 (1966).
82. R. Dolan et al., "Classification of the Coastal Environments of the World: Part I, The Americas," University of Va., Dep. of Environmental Sciencies, Tech. Rep. 1, 163 p., 1972.
83. B. Hayden and R. Dolan, "Classification of the Coastal Environments of the World," University of Va., Dep. of Environmental Sciences, 167 p., 1975.
84. A. T. Ippen, "Sedimentation in Estuaries," in *Estuary and Coastal Hydrodynamics*, A. T. Ippen, ed., McGraw-Hill, New York, 1966, pp. 648–672.
85. F. T. Manheim, J. C. Hathaway, and E. Uchupi, "Suspended Matter in Surface Water of Northern Gulf of Mexico," *Limnol. Oceanogr.* **17**, 17–27 (1972).
86. K. O. Emery et al., "Geological Structure and Some Water Characteristics of the East China Sea and Yellow Sea," ECAFE, Tech. Bull. 2, pp. 3–43, 1969.
87. T. H. Van Andel and H. Postma, "Recent Sediments of the Gulf of Paria," *Verhandel. Koninkl. Ned. Okad. Witensch.* **20**, 1–245 (1954).
88. D. J. P. Swift and R. G. Pirie, "Fine-Sediment Dispersal in the Gulf of San Miguel, Western Gulf of Panama: A Reconnaissance," *J. Mar. Res.* **28**, 69–95 (1970).
89. H. Postma, "Transport and Accumulation of Suspended Matter in the Dutch Wadden Sea," *Netherland J. Sea Res.* **1**, 191–240 (1961).
90. *A Study on the Siltation of the Bangkok Port Channel*, 3 vols., 474 p., NEDECO, The Hague, The Netherlands, 1965.
91. V. P. Zenkovich, *Processes of Coastal Development*, 738 p., Interscience, New York, 1967.
92. Delft Hydraulics Laboratory, *Demerara Coastal Investigation*, 240 p., 1962.
93. R. B. Krone, "A Study of Rheological Properties of Estuarine Sediments," U.S. Army Corps of Engineers, Comm. on Tidal Hydraulics, Tech. Bull. 7, 1963.

94. J. N. Suhayda et al., "Marine Sediment Instability: Interaction of Hydrodynamic Forces and Sediment Movement," *Proceedings of the Offshore Technology Conference*, Houston, Tex., Pap. 2625, Offshore Technology Society, Dallas, Tex., 1976.
95. I. V. Samajlov, *Die Flussmundungen*. (Veb. Hermann Haack), Gotha, Germany, 647 p., 1956.
96. J. M. Coleman and L. D. Wright, "Analysis of Major River Systems and Their Deltas: Procedures and Rationale, with Two Examples," Coastal Studies Inst. Tech. Rep. 95, Louisiana State Univ., 125 p., 1971.
97. J. M. Coleman and L. D. Wright. "Modern River Deltas: Variability of Processes and Sand Bodies," in *Deltas, Models for Exploration*, M. L. Broussard, ed., 555 p., Houston, Tex., Geol. Soc., 1975.
98. J. M. Coleman and L. D. Wright, "Research Techniques in Deltas," *Proc. Russell Symposium, Louisiana State Univ., 1976* (in press).
99. B. W. Nelson, "Hydrology, Sediment Dispersal, and Recent Historical Development of the Po River Delta, Italy," in *Deltaic Sedimentation, Modern and Ancient*, Soc. Economic Paleontol. and Mineral., Spec. Publ. 15, pp. 152–184, 1970.
100. T. Whelan et al., "The Geochemistry of Recent Mississippi River Delta Sediments: Gas Concentrations and Sediment Stability," *Proceedings of the Seventh Offshore Technology Conference, Houston, Texas*, pp. 71–85, 1975.
101. H. H. Roberts, D. W. Cratsley, and T. Whelan, "Stability of Mississippi Delta Sediments as Evaluated by Analysis of Structural Features in Sediment Borings," *Offshore Tech. Conf. Proc.*, Pap. OTC 2425, Houston, Tex., 1976.
102. W. H. Munk and M. S. Sargent, "Adjustment of Bikini Atoll to Ocean Waves," U.S. Geological Survey, Prof. Pap. 260-C, pp. 275–280, 1954.
103. W. S. von Arx, "Circulation Systems of Bikini and Rongelap Lagoons," U.S. Geological Survey, Prof. Pap. 260-B, pp. 265–273, 1954.
104. D. L. Inman, W. R. Gayman, and D. C. Cox, "Littoral Sedimentary Processes on Kauai, a Subtropical High Island," *Pacific Sci.* **17**, 106–130 (1963).
105. R. J. Tait, "Wave Set-Up on Coral Reefs," *J. Geophys. Res.* **77**, 2207–2211 (1972).
106. M. L. Hernandez and H. H. Roberts, "Form-Process Relationships on Island Coasts," Louisiana State Univ., Coastal Studies Inst., Tech. Rep. 166, 76 p., 1974.
107. H. H. Roberts, "Variability of Reefs with Regard to Changes in Wave Power Around an Island," *Proceedings of the 2nd International Coral Reef Symposium, Great Barrier Reef Comm., Brisbane, Australia*, pp. 497–512, 1974.
108. E. C. F. Bird, *Coasts: An Introduction to Systematic Geomorphology*, 2d ed., vol. 4, 246 p., MIT Press, Cambridge, Mass., 1970.
109. K. Horikawa and T. Sunamura, "Field Investigations of Coastal Erosion at Byobugaura and Taitomisaki, Chiba Prefecture," University of Tokyo, Dep. of Civil Engineering, Coastal Engineering Laboratory, Tech. Rep. BT-2, 128 p., 1970.
110. T. Sunamura and K. Horikawa, "A Quantitative Study on the Effect of Beach Deposits upon Cliff Erosion," *Coastal Engr. in Japan* **14**, 97–106 (1971).
111. K. R. Dyer, *Estuaries, a Physical Introduction*, 140 p., J. Wiley, London, 1973.
112. K. F. Bowden and P. Hamilton, "Some Experiments with a Numerical Model of Circulation and Mixing in a Tidal Estuary," *Estuarine Coastal Mar. Sci.* **3**, 281–301 (1975).
113. M. Rattray, Jr., "Some Aspects of the Dynamics of Circulation in Fjords," in *Estuaries*, G. H. Lauff, ed., American Association for the Advancement of Science, Washington, D.C., 1967.
114. H. G. Gade, "Deep Water Exchanges in a Sill Fjord: A Stochastic Process," *J. Phys. Oceanogr.* **3** (2), 213–219 (1973).
115. R. Long, "On the Depth of a Halocline in an Estuary," *J. Phys. Oceanogr.* **5** (3), 551–554 (1975).
116. T. Lee and C. G. H. Booth, "Circulation and Exchange Processes in Southeast Florida's Coastal Lagoons," Rosenstiel School of Marine and Atmospheric Science, Miami, Florida, Spec. Rep. 5, 1975.
117. B. J. Kjerfve, "Dynamics of the Water Surface in a Bar-Built Estuary," Louisiana State Univ., Baton Rouge, Ph.D. dissertation, 90 p., 1973.
118. K. R. Dyer and D. Ramamoorthy, "Salinity and Water Circulation in the Vellar Estuary," *Limnol. Oceanogr.* **14**, 4–15 (1969).
119. M. Tomczak, Jr., and C. G. Diaz, "A Numerical Model of the Circulation in Cienfuegos Bay, Cuba," *Estuarine Coastal Mar. Sci.* **3**, 391–412 (1975).
120. J. Bye and J. A. Whitehead, Jr., "A Theoretical Model of the Flow in the Mouth of Spencer Gulf, South Australia," *Estuarine Coastal Mar. Sci.* **3**, 477–481 (1975).

Walter Orr Roberts is a Professor of Astro-Geophysics at the University of Colorado, a trustee of the Max C. Fleischmann Foundation, Director of the Program in Science, Technology, and Humanism at the Aspen Institute for Humanistic Studies, and a Research Associate at the National Center for Atmospheric Research. As a Harvard graduate student in 1940 he established the Climax, Colo., solar coronagraph station of Harvard College Observatory. Dr. Roberts was a Research Associate at the Harvard College Observatory from 1948 to 1969. He became Director of the High Altitude Observatory in 1946, and he was the first Director of the National Center for Atmospheric Research. His research interests include the solar corona, solar spicules and prominences, the origin of geomagnetic disturbances, the influence of variable solar activity on the Earth's ionosphere and weather, and the effects of climate on world food production. He has published extensively in domestic and foreign journals. Dr. Roberts received an A.B. from Amherst College, M.A. and Ph.D. degrees from Harvard University, and numerous honorary degrees. He is a member of Phi Beta Kappa, Sigma Xi, and a number of scientific societies. He has acted as trustee to corporations, universities, and foundations and served on many boards and advisory committees, including those of the National Academy of Sciences. He has received, among other awards, the Cleveland Abbe Award of the American Meteorological Society and the Hagkins Medal of the Smithsonian Institution.



SUN-EARTH RELATIONSHIPS AND THE EXTENDED FORECAST PROBLEM

Walter Orr Roberts

Aspen Institute for Humanistic Studies
Professor of Astro-Geophysics, University of Colorado
Research Associate, National Center for Atmospheric Research
1919 Fourteenth Street, Room 811
Boulder, Colo.

From the dawn of the human intellect, men and women have anxiously scanned the skies for signs of change in the weather. Every living thing is affected by weather, and particularly by the extremes of wind, drought, flood, heat, and cold. Whole civilizations have been altered by large-scale, long-lasting changes in the climate. Entire races of people in ancient times were forced to migrate from their traditional homelands to more favorable lands in time of sustained drought.

In modern times the impact of weather and climate is not lessened. To be sure, individuals in favorable circumstances can be almost completely sheltered from blizzard cold and searing heat, but even these favored few suffer the economic impact of adverse weather. The vast majority of Earth's people are more vulnerable. In the semiarid lands, where changes are expecially large and frequent, populations are constrained against protective migration by political borders and by ownership patterns that leave little of the planet's land surface free. For the world's poor, weather and climate have a desperately severe impact. When drought strikes, as in the Sahel belt of Africa or western India, millions go hungry, cattle die, and children suffer malnutrition. Disease and premature deaths result.

No wonder, then, that every sign has been sought for predicting weather change. No wonder, either, that superstitions and false arts in weather forecasting have risen to lift from people the anxiety of uncertain future weather. And no wonder that leading scientists have, since the rise of learning, diligently pursued the extended-range weather forecasting problem. Few products of science and technology could possibly have more relevance and value to humanity.

I shall explore in this paper one of the many avenues of research on weather and climate, namely that having to do with variations in the Sun's emissions to space and their effects on the Earth's weather and climate. I shall endeavor to show that in spite of grave difficulties and uncertainities, there is bright promise of progress ahead, and that this promise may, just possibly, aid us in improving forecasting beyond the approximate 5-day limit of the best present global numerical forecast modeling techniques.

STATE OF THE PROBLEM

The Earth's weather machine is an exquisitely complex affair, in which many processes are simultaneously at work. Some, if not most, involve nonlinear interactions. This makes it extraordinarily difficult to identify cause-and-effect relationships from statistical-historical studies of the weather system.

Clearly, however, the weather system is principally driven by the Sun. The dust of volcanoes and the waste heat from factories, cities, and powerplants can produce only minor perturbations, although sometimes these have substantial human impact. However, if the Sun's life-giving radiation were to alter by just a few percent we would expect large changes in weather and climate. The patterns of glaciation would change, the distribution of rainfall would alter, and many other changes would develop as the weather machine adjusted to a different solar driving force.

Weather changes occur, of course, even in the absence of several-percent changes in the Sun's output. The changes occur in all aspects of weather: frequency of rain, strength of winds, snowfall, temperature, tornado occurrence, hurricane paths, humidity, etc. Moreover, changes occur on all time scales. Compared to long-term averages, there are anomalous days, weeks, months, years, decades, centuries, millenia, and geological eras.

So sensitively balanced are the living things of earth that these changes, even those of small percentages are nevertheless almost always important. A 6-week hot, dry spell can reduce the nation's corn output markedly, as it did in 1974. A short period of intense rain can produce a flood disaster, as in Rapid City, South Dakota on June 9, 1972, or in the Big Thompson River of Colorado on August 1, 1976.

Yet the causes of these anomalies remain obscure. It is even possible that there is no "cause," in the usual sense. Perhaps these extremes, with their disastrous local effects, are simply a part of the normal statistical fluctuations of the complex dynamical system that makes up the weather machine. In any event, they are a part of the weather-climate system that we must expect to live with forever. And it is not clear how well we shall succeed at their prediction.

One series of wheels and levers in the global weather machine involves the Sun's variable activity. I am convinced that this has, on occasion, significant impact on weather. The way this impact comes about, and its future potential for the extended-range forecast problem remain to be seen. There is enough promise, however, to justify more systematic research in the years to come.

From the time of Galileo's first solar observations around 1610, sunspots have been systematically studied as an evidence of some variable happenings on the Sun. For well over a century we have known of the roughly 11-year cycle in the sunspots' average numbers and sizes. For a half century or so, we have known that the magnetic polarity of the sunspots reverses from 11-year cycle to cycle, giving us, in reality, 22-year quasi-cyclical behavior. More recently, John A. Eddy in 1976, gave strong evidence for a long period of an almost spot-free sun in the late 1600's and early 1700's [1]. Is it mere coincidence that the "little ice age" was most severe during this time?

The Office of Naval Research during its entire 30 years has effectively pursued research on solar variability and the associated space and terrestrial effects. This field of research has grown enormously during this time. Today many sponsors and participating institutions produce an incredible wealth of knowledge. Space technology has brought the most revolutionary advances. We now regard the Sun not simply as a constant and steady source of light and heat, but also as an emitter of a vast array of earth-affecting particles and radiations. The radio wave, ultraviolet, and X-ray radiations incident on earth fluctuate irregularly by orders of magnitudes, and some of the pulses have abrupt onsets measured in seconds.

The majestic solar corona is a constantly changing thing, and the associated solar wind is a gusty wind as it blows past the Earth. Solar flares blast relativistic particles into space, disrupt magnetic field lines, and emit X-rays and ultraviolet. These phenomena of the variable sun have profound effects in the earth's upper atmosphere, as described in another article in this volume [2].

My concern in this paper, however, is strictly with the lower atmospheric weather and climate effects of these solar fluctuations. The problem I shall address in this paper is the degree to which there is promise of extending the weather and climate forecast range by appeal to solar variability as a causative factor.

At present solar activity is not regarded by many experts as an important element in forecasting the short-term detailed character of weather. So far as I know, no governmental operational

weather forecast group concerns itself routinely with observations of solar flares, or even of solar-associated ionospheric phenomena such as geomagnetic activity and auroras. Though it may be premature to consider solar activity as a practically useful factor in weather forecasting, there is growing evidence that the lower stratosphere and troposphere respond within 24 hours to certain solar-related phenomena, and that the effects appear significant in synoptic map analyses out to at least a week after the onset. Effects that persist a week or more in synoptic data are of special interest for the extended-range forecast, where predictive skill usually drops essentially to zero after 5 days. The responses to solar activity appear to cover most of the Northern Hemisphere and may be global. (Adequate tests in the Southern Hemisphere do not exist.) The responses, moreover, are of such magnitude as to be significant for forecasting if they can be understood.

As in more conventional approaches to weather forecasting, it is useful in the Sun-weather field to distinguish between (a) short-term forecasting of the detailed state of the global atmospheric system and (b) monthly, seasonal, annual, and longer-term forecasting of the averages of the weather parameters that characterize the climate of a region or the globe. I shall divide my discussion in this way, below.

It should be stated at the outset that many leaders in climate and weather research remain, even to this time, highly doubtful that solar activity significantly influences weather. Most recently, B. J. Mason, Chief of the Meteorological Office in the United Kingdom, expressed his skeptical view pungently in a public meeting of the Royal Meteorological Society [3]. Andrei Monin, distinguished Soviet weather expert, after some sharp comments about the danger of "heliogeophysical enthusiastics," went on to say:

> ". . . the greatest attention should be devoted to the question of whether there is a connection between the earth's weather and the fluctuations in solar activity. The presence of such a connection would be almost a tragedy for meteorology, since it would evidently mean that it would first be necessary to predict the solar activity in order to predict the weather; this would greatly postpone the development of scientific methods of weather prediction. Therefore arguments concerning the presence of such a connection should be viewed most critically."

To these skeptics I respond that the evidence supporting this elusive clue appears stronger and stronger. Moreover, as time passes the massive amounts of research in conventional meteorology do not appear to be giving bright promises of extending the detailed forecast range much beyond the 5-day present practical average limit of skill. Thus, all clues to new operative mechanisms, even if elusive and poorly understood, deserve forceful pursuit as well as critical review. To add strength of effort to this field would be very much in the spirit of venturesome search that has so long characterized the Office of Naval Research. The Office of Naval Research set the tone for the National Science Foundation and established operating philosophies that were critically important in the development of the post-World War II burst of creative scientific research in the United States. This spirit of innovative pursuit of new clues sometimes gets lost from view in the "big science" scene of today, where so many well-established programs are capable of further extending their research efforts in the more conventional domains, and thus consuming whatever incremental funds may become available.

The Evidence for Sun-Weather Effects

The history of the search for clues to weather prediction from solar variability is long. The path to our present position is a slow one, with many false branches and a twisted route. Most early attention was directed toward long-term solar weather relationships, usually with the long-term fluctuation of sunspot activity. The trend of the 11-year or 22-year quasi-cycle in sunspot numbers has often correlated for extended terms with meteorological variables, only to randomize or reverse phase when independent new data emerged against which to test the finding. Progress in solar-weather studies appears to be coming

first not from the long-term data analysis, though, but from the short term.

It now appears, as I shall describe below, that unassailable evidence is at hand to establish the reality of certain short-term weather responses to solar variables. It is therefore important today for researchers, in some group or other, to reexamine the spotty but voluminous scientific literature of the past to see where and how the best of the historical findings fit into a theoretical framework capable of explaining or at least consistent with the new evidence in which we have confidence today. By such measures we may perhaps discover, in a synthesis of past and present work, clues to the physical processes at work. These are now clouded in mystery so deep as to engender the skepticism that many well-qualified persons feel.

In the next pages I shall attempt a brief and selective summary of evidence that I consider most secure in establishing confidence that the variations of the sun affect the troposphere and stratosphere.

The Earlier Research

Most early research on Sun-weather effects involved the long-term climatic variables and the 11-year and longer variations of solar activity usually measured by sunspot frequency and size.

W. Herschel in 1801 produced one of the earliest quantitative results that appears to stand the test of time and new data [5]. Herschel found that in the rainy regions of the tropics, there is a small but seemingly significant trend of the average temperatures downwards during periods of increasing sunspot activity, and that average temperatures rise as sunspot activity decreases. The result was confirmed and extended by W Koeppen in 1873 [6]. This result merits reanalysis with modern data.

Within the last 100 years many workers entered the field of Sun-weather research. Among these were, to name but a few, H. Clayton, G. Walker, A. Girs, C. G. Abbott, S. Hanzlik, F. Baur, H. Willett, H. Wexler, and V. Rubaschev.

In the middle of the 19th century, speculation began about visual observations indicating that sudden cirrus cloud covers often developed over the whole sky after nights with brilliant auroras. Such effects, if really connected to the aurora, are relevant to short-term Sun-weather relationships, since auroras are now known to have a causal link with solar activity.

Duell and Duell [7], in a classic work on Sun-weather relationships, supported the idea of a cirrus cloud mechanism with both observational and theoretical evidence. They cited early observational data by H. Fritz and others obtained in the latter part of the 19th century, which seemed to link auroral activity and cirrus formation. They also cited work by Archenhold [8], who found a relationship among sunspots, geomagnetic storms, and solar haloes. C. G. Abbot in 1948 furnished additional evidence that enhanced sky brightness correlated with geomagnetic storms [9].

Barber [10] subsequently reported that on 50 cloudless days, over a short period, there were systematic increases in the scattered zenith sky light on days of moderate or strong magnetic storms. Strangely enough, these suggestive results have not been tested with new time periods or with the powerful polarizing photometers and other technologies of today. Such confimation is urgently needed. If it is verified that cloud or haze particles are produced by solar activity, it may be that we have found a first plausible Sun-weather causal mechanism, through the modulating effect of the cirrus on the infrared budget of the earth.

Many workers have examined temperature and pressure patterns, drought recurrence, cyclone formation, blocking high-pressure cells in the Pacific and Atlantic, and a wide range of other meteorological variables in empirical-statistical connections between solar activity and weather. Some of these works have indicated small but seemingly real Sun-weather effects. However it has usually been possible to argue that the results are not statistically significant. Lacking any sound theoretical basis on which to expect a Sun-weather effect, most objective analysts have remained skeptical. Moreover, much of the published work has been poorly done, with sloppy statistical methods and often with complicated results that have been grossly overinterpreted. This has given the field of Sun-weather research a bad reputation among careful scientists. Things are, however, beginning to change.

Modern Evidence of Short-Term Relationships

The first solid modern statistical work on short-term Sun-weather research is, in my opinion, the landmark 1956 paper of R.A. Shapiro [11]. In this, Shapiro examined data from a grid of North American surface barometric-pressure data for the years 1899–1945. He calculated a "persistence correlation index" for all days of large geomagnetic disturbance, measured by large values of the so-called c_1 geomagnetic disturbance index. The persistence correlation index involves correlating the mean value of the barometric pressure at a given station on days 0, 1, and 2 after a disturbed geomagnetic day with the mean pressure of days 3, 4, and 5 at the same station, then averaging over all stations and all disturbed days. A high persistence correlation index means small average 3-day change, or high persistence; low indices mean large average changes over 3 days, or low persistence. Shapiro then calculated the same index for the days preceding and following the geomagnetic disturbances, and plotted the results of the average persistence index as a function of the number of days before and after the day of large geomagnetic disturbance. Both the peak and the minimum are estimated to be significant at or better than the 95% confidence limit.

Figure 1 is from the original paper, and shows all years from 1899 to 1945, except for the 3 years centered on sunspot maximum. The figure shows a rise in persistence (small average barometric pressure change) following the geomagnetic disturbance, peaking at the third and fourth days. Following the broad peak is a long, steady decrease of persistence continuing out to day +14. As can be seen, no significant trends showed up in the days before the geomagnetic disturbance.

The result is consistent with an earlier study by R. Shapiro at the 500-mbar level (with a much smaller time period) in 1953, near the 1952–1953 sunspot minimum, when there were pronounced 27-day recurrent storms [12]. Figure 2 is adapted from the 500-mbar result, which led Shapiro to undertake the more massive work summarized in Figure 1. It exhibits the drop in persistence correlation index in the days following eight geomagnetic disturbances in 1953.

The persistence relationship to disturbed days showed up when the 1899–1945 data set was subdivided in various ways, lending confidence to the conclusion. Shapiro found that eliminating peak sunspot years, as in Figure 1, improved the significance of the result.

Correlations of similar nature in Europe, as given in Figure 3 [13], showed a significant persistence peak, like that of Figure 1, in the first days

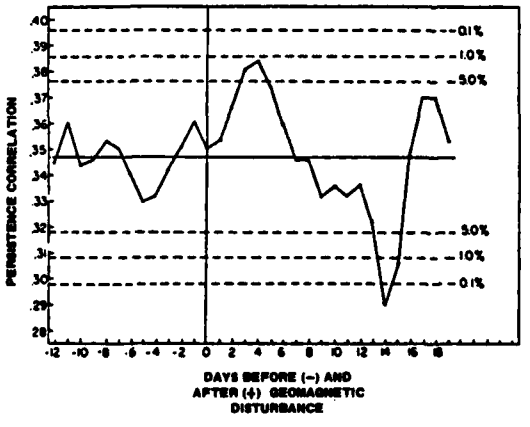


Figure 1—Average sea level persistence correlation over North America as a function of the number of days before (–) and after (+) a geomagnetic disturbance. Twelve years of maximum sunspot activity have been eliminated. Each mean represents 399 cases covering 35 years.

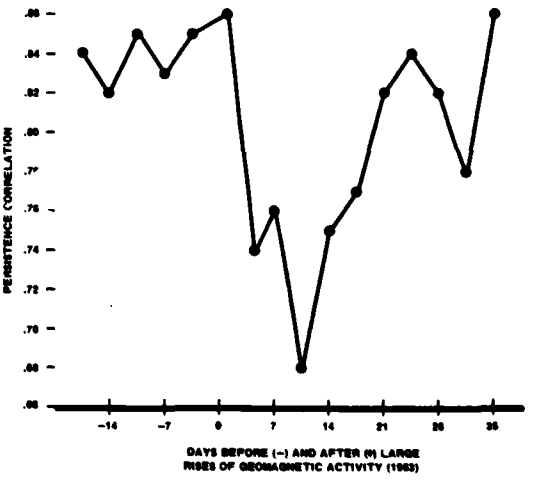


Figure 2—Average 500-mbar persistence correlation as a function of the number of days before (—) and after (+) large rises of the geomagnetic disturbance index K_p during 1953. Adapted from Table 1, Ref. 12.

after a geomagnetic disturbance, and they exhibited a similar slow decline later. However, the minimum after the slow decline was of marginal statistical significance.

These papers of Shapiro, when coupled with earlier suggestive but inconclusive results by Craig [14] and by B. and J. Duell [7] apparently showing opposite signs of barometric pressure changes following quiet and disturbed days, convinced me that there was solid reason to probe deeper into the relationships of solar activity, geomagnetic disturbance, and weather.

In the mid 1950s I gathered together a small group for research in the problem at the High Altitude Observatory, University of Colorado. I was encouraged and assisted by the Office of Naval Research and the Canadian Meteorological Service, and in particular by John N. Adkins of ONR and Patrick McTaggert-Cowan and Andrew Thompson of CMS. My first colleagues included David Woodbridge and Theodore Pohrte of the Colorado School of Mines and Norman MacDonald of the High Altitude Observatory.

In 1955 we spent many hours poring over copies of new 300-mbar Northern Hemisphere jet stream and pressure-height maps of western hemisphere weather between 0° and 180° longitude. We became convinced that there were marked instances of abrupt and large breakdowns of winter and spring zonal wind flow to meridional flow, after the eastward migration and growth of low-pressure troughs first identified in the Gulf of Alaska area a few days after strong auroras and magnetic storms. In 1956 we decided that it was worthwhile to try to quantify these highly subjective observations.

Our first step was to devise some measurement parameters. Woodbridge [15] developed a "trough index," which measured the ratio of the width of the trough to the depth: $I_t = W/D$. The width was measured from the inflection point of the southward bend of the contour line, as shown in Figure 4, to the corresponding inflection on the northward bend. The depth was simply the distance from the line joining the inflection point to the trough line at the southmost penetration of the isoheight line. In practise we used two height lines, as shown in Figure 4, and averaged the ratios. Closing troughs and cutoff lows were similarly measured. We felt that this trough index represented a rough but objective measure of the strength of cyclonic activity of the trough system.

The results of 3 years of this effort are summarized in Figure 5 and Table 1. Figure 5, from a 1960 paper by Macdonald and Roberts [16] shows the trend of the trough index on the day of first appearance (day=0) of the 300-mbar low-pressure system in the Gulf of Alaska "target area" defined as the sector between longitude 120° and 180°W, north of latitude 40°N. Every trough that could be identified in this area during the winter half year (October 1 to March 31) was included in the analysis. For each day of the trough's recognizable life we measured its trough index value. Sometimes we could follow trough systems for up to 2 weeks as they migrated east across North America and sometimes even the Atlantic Ocean.

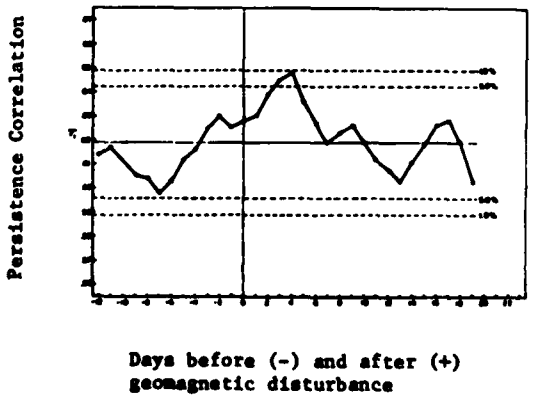


Figure 3—Average sea level persistence correlation over Europe as a function of the number of days before (-) and after (+) geomagnetic disturbance.

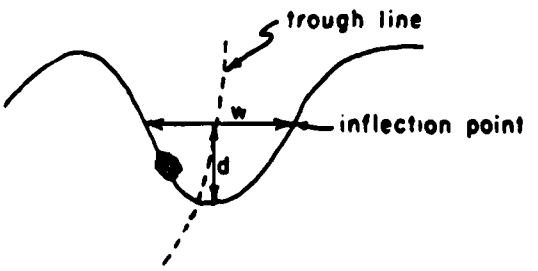


Figure 4—Method used for measuring the trough index. The final trough index was determined by averaging the calculations made at the 30 000-ft (9150 m) contour height and the 29 200-ft (8900 m) contour height.

Figure 5 shows average values of the trough index on the first and succeeding days for all winter half-year troughs in the periods 1956–1957, 1957–1958 and 1958–1959.

The dotted curve of Figure 5 shows the averaged index for the 52 troughs during this time that were preceded by strong auroras or sudden magnetic storms. It consists of all troughs whose first appearance in the Gulf of Alaska area came on days 2, 3, or 4 after the geomagnetic and auroral event. The solid line corresponds to all others.

Most notable is that the dotted line on day 5 after first recognition exhibits a trough index about 50% greater than the values of troughs not preceded by the geomagnetic and auroral outbreak.

It is of interest, in retrospect, that we failed to pursue or call attention to the fact that the geomagnetic and auroral troughs exhibited, when first indentified, lower trough indices than the others. Recent work by Olson, Roberts, and Zerefos [17] suggests that this difference is real and significant to us as we seek to understand what is happening physically.

To quantify the significance of these findings we worked up the contingency table shown as Table 1. In this table the first column shows the number of troughs preceded by the geomagnetic and auroral event. The troughs are divided into three classes, depending on the largest size they reached during their measurable lifetimes. The class intervals were arbitrarily chosen to produce a large, medium, and small class, each of which contained about one-third of the total numbers. The second column shows the numbers of troughs not preceded by a geomagnetic and auroral event on the second, third, or fourth day prior.

The contingency table exhibits a striking departure from the random expectations (which are shown in parentheses). There are three times as many large troughs as small ones in the geomagnetic and auroral column, and significantly more small troughs in the column of troughs not preceded by such events. The results we deemed statistically significant to a very high degree. By χ^2 analysis we concluded that the random probability of a table so skewed as this is on the order of 10^{-6}.

These results, which were also exhibited in subsets of the total set, indicated to us that solar activity, either directly or through association with the geomagnetic and auroral activity, was somehow causing the cyclonic activity of troughs in the Gulf of Alaska near this time to reach a larger ultimate intensity than would otherwise

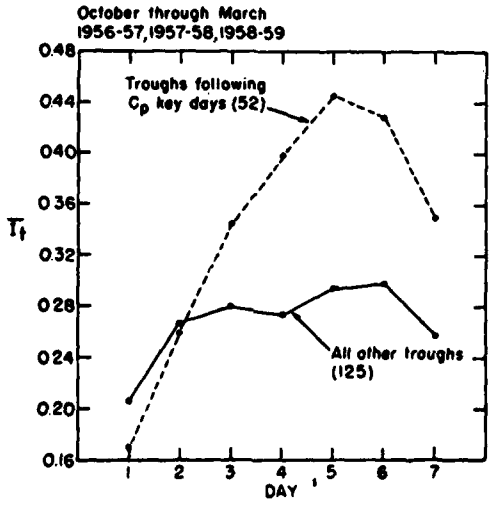


Figure 5—Average trough amplitude I_1 for the 7 days after first appearance in the target area for all troughs that did not move off the maps (past 0 longitude) during the 7 days. Dotted curves represent troughs that first appeared in target area on second, third, and fourth days after geomagnetic storm dates. Solid curves represent all others.

Table 1

Number of Troughs of Various Sizes After Geomagnetic Disturbances and not After Geomagnetic Disturbances

Trough Size	*After Geomagnetic Disturbance*	*Not After Geomagnetic Disturbance*	*Total*
Large	34(18)	31(47)	65
Medium	8(16)	50(42)	58
Small	10(18)	54(46)	64
Total	52	135	187

have been so. Moreover, other studies showed that the sensitivity of this trough discrimination was sharply centered on the third day after the geomagnetic and auroral activity. Including days 1-5, for example, did not enhance the discrimination.

During this same span of time, from the mid 1950s to the mid 1960s, Sun-weather research accelerated abroad, particularly in the USSR. Important contributions were made by B. Sazanov, L. Rakipova and many others, including C. Schuurmans of the Netherlands.

E. Mustel, head of the Astronomical Council of the Academy of Sciences of the USSR, became a leading contributor to research on Sun-weather relationships in his country, where substantial numbers of workers devoted their efforts to this field. Mustel in 1972 published an important summary of his work with various collaborators over more than a decade [18]. From this I have selected Figure 6. This figure shows winter (December through Februray) data from 1890 to 1967. The solid circles show stations of the upper latitudes of the northern hemisphere where surface barometric pressures exhibited an average increase at about 2 to 4 days after a large, isolated geomagnetic disturbance. The open circles represent stations with pressure drops from about 1 to 5 days after a geomagnetic disturbance.



Figure 6—Hemispheric distribution of the change in atmospheric pressure after a geomagnetic storm for the months of December through February and the years 1890–1967. The black circles correspond to an increase in pressure, and the open circles correspond to a decrease in pressure.

This work of Mustel again points strongly towards *winter season, high-latitude* weather responses to geomagnetic and auroral events. As with our own work, however, there is no good clue to what physical mechanism is operating.

Some Recent Results on Short-Term Relationships

In the 1970s I returned again to the problem of Sun-weather research, collaborating with R. H. Olson. To meet criticisms of subjectivity in our earlier work, Olson and I developed a new, more objective index to measure cyclonic trough intensity and a new and completely objective technique to identify geomagnetic and auroral events. To compute the trough index, we first produced 300-mbar maps at half-day time intervals for the Northern Hemisphere, representing in the maps the contours of the values of the absolute vorticity. We then measured the area of high positive (cyclonic) vorticity associated with each low-pressure trough first recognized in the Gulf of Alaska area defined as before. We chose as our index the area where the vorticity exceeded 20×10^{-5} s^{-1} *plus* the area where the vorticity exceeded 24×10^{-5} s^{-1}. We called the index the "vorticity area index" (VAI). Our results covered winter half years from 1964 to 1971.

In the new study we used the value of the VAI averaged over the first 3 days in the Gulf of Alaska (rather than the largest trough index obtained during the whole of the recognizable life of the trough). We also adopted a criterion of the geomagnetic and auroral event based entirely on the A_p geomagnetic disturbance index, thus averting any possible effect of weather on the visibility of auroras as a biasing factor in the statistics.

The contingency table for the new study, comparable to Table 1, is given as Table 2 from a 1973 paper by Roberts and Olson [19].

In Table 2 we show in the first column, just as in Table 1, the numbers of troughs that entered the Gulf of Alaska area on days 2, 3, and 4 after sharp increases in geomagnetic activity. The second column, however, is here only those troughs preceded by 10 geomagnetically quiet days before their entry into the Gulf.

The association of strong troughs with geomagnetic activity and weak troughs with

Table 2

Number of Wintertime Troughs that Attained an Average Vorticity Area Index of Large, Medium, or Small During First 3 Days of Trough Life in North Pacific East of 180° Longitude (Numbers in ( ) are randomly expected numbers.)

Trough Size	*Troughs Preceded By Sharp Geomagnetic Rise*	*Troughs Preceded By 10 Days of Geomagnetic Quiet*	*Total*
Large	45(30)	28(43)	73
Medium	27(30)	46(43)	73
Small	22(34)	60(48)	82
Total	94	134	228

geomagnetic quiet is again very striking. Table 2 is a fully independent and more objective confirmation of the earlier results. As Hines [20] has pointed out, however, it is in principle possible that the results of both tables could be explained by an effect on geomagnetic indices produced by strong lower stratospheric cyclones transmitting sufficient energy through gravity waves to the ionosphere. If this energy were to make a small but significant increase in the geomagnetic indices, it could bring about the inclusion of some extra large troughs in column 1, thus giving the statistical association. In this instance the table would reflect a meteorological influence on the ionosphere (a weather → geomagnetism → weather relationship, rather than a solar activity → geomagnetism → weather relationship).

The next major step in establishing that we are dealing with a Sun-weather effect came about as a result of collaboration of our group with a group at Stanford University under John M. Wilcox. In this study we added up, for each half-day weather map interval, all the vorticity area indices for the 500-mbar level (data at this level are somewhat more homogeneous than those at 300 mbar) for the Northern Hemisphere north of 20°N. We then used as a criterion of solar activity the time when solar magnetic sector boundaries were swept past the earth by the solar wind.

Figure 7, from Wilcox [21], summarizes the latest results. The upper graph shows the average behavior of the hemispheric vorticity area index averaged over days before and after the 50 sector-boundary passages published in the first analysis. The lower graph shows results from 81 new sector passages not included in the first analysis. It is clear that the two curves are essentially alike, lending confidence to the result. Vari-

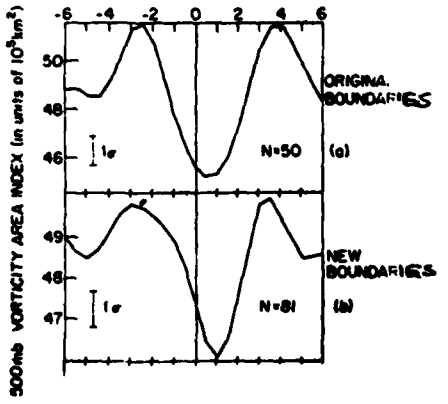


Figure 7—Average response of the hemispheric vorticity area index about times when solar magnetic sector boundaries were swept past the Earth by the solar wind. The upper figure represents the original boundaries analyzed; the lower figure shows 81 new sector boundary identifications.

ous other subsets of the data, likewise, show the same "signature." The signature, moreover, is repeated when we use only sector boundaries obtained from spacecraft data, thus eliminating the possibility of a weather → geomagnetism → weather correlation. I emphasize, as Wilcox has done, that the sector passage is a precise and convenient timing mark for the organization of solar activity such as flares, sunspot groups, solar surface magnetic structure, and solar wind. The boundaries themselves are almost certainly not a causal factor in the hemispheric VAI changes. The fact that the VAI begins to dip before the sector passage does not imply a weather response prior to the solar event, because systematic solar-terrestrial effects involving the solar wind, flares, etc., precede as well as follow the sector boundary passage.

There has been a good deal of argument about the size of the error bars in Wilcox's graph, as an estimate of the statistical significance of the result. This may be an interesting mathematical argument, but I prefer not to estimate the significance from the size of the error bars, but from the fact that almost all subsets of the data display essentially the same signature.

The most thorough and convincing statistical analysis of the VAI-sector work, however, has been done by Hines and Halevy [22]. They derived the signature information independently from our data, then subjected it to various extremely detailed tests. They also requested new solar sector-boundary values from which to make additional tests independent of earlier data. From this they concluded, "Reports of Sun-weather correlations have been greeted with skepticism by many. . . . We find ourselves obliged, however, to accept the validity of the claim by Wilcox et al., and to seek a physical explanation."

One other quite independent line of research on Sun-weather work merits attention because of its importance in the search for explanatory physical mechanisms. This is the several recent and extremely interesting works of R. Reiter [23, 24] that relate to an apparent effect of strong solar flares in the terrestrial atmospheric potential gradient, in the distribution and frequency of thunderstorms, and in the injection of stratospheric air into lower levels. These results of Reiter's work suggest that solar H_α flares are followed by 1- to 4-day-long rises in the daily mean value of the atmospheric potential gradient and of the earth-air current density. This, he suggests, affects the frequency and size of thunderstorms, and these are, of course, important to the overall energetics of global atmospheric circulation. Reiter concludes that in the principal world thunderstorm activity centers, thunderstorm activity increases within a few days after the solar flares. Bossolasco and collaborators [25] have reached similar conclusions from independent analysis.

Markson [26] has conducted independent analysis of solar sector boundaries and thunderstorms and has shown apparently significant associations. He concludes that there is evidence (a) for a long-term secular effect in worldwide thunderstorm activity, which varies inversely with solar activity over the sunspot cycle and may result from changes in the atmospheric ionization from galactic cosmic rays, which inversely correlate with solar activity, and (b) short-term effects characterized by increases in the earth-to-ionosphere current flow and by increased thunderstorm activity for several days following solar flares which, he believes, provide ionization in the air column between thunderstorm tops and the ionosphere as a consequence of energetic solar particles associated with flare emissions.

These relationships between solar activity and thunderstorms are not established to the degree of confidence of the sector-boundary and vorticity relationships described above. Establishing their reality will be, however, an extremely valuable advance if it can be achieved. This effect, if real, holds promise, in an otherwise very bleak picture, of providing a plausible physical mechanism by means of which the very small total energies associated with solar activity effects at the earth can modulate the general circulation.

Long-Term Drought and Temperature Recurrence Trends

Another possible meteorological response to variable solar activity is the apparent long-term trend in the occurrence of droughts in the Great Plains section of North America. There is also some evidence that droughts in parts of the Soviet Union follow the same general time pattern. However, in this section I shall deal only with the

North American case, since I consider this the best-established example.

The seeming dependence of drought in the Great Plains on the 22-year double sunspot cycle has been mentioned by several authors, e.g., Borchert [27], Marshall [28], Palmer [29], Roberts [30], Thomas [31], and Willett [32]. These and other authors have pointed to a periodicity of 20-22 years in these droughts Some authors [33] have traced the relationship as an unbroken series of droughts extending back in time to 1800.

The droughts, which are most pronounced in the western Great Plains, seem to recur near the time of every other sunspot minimum, and specifically at the time of the minimum in solar activity when the magnetic polarity of the leading sunspots in the sun's northern hemisphere is changing from north-seeking to south-seeking. Thus the recurrence period is roughly 20-22 years. Table 3 shows the dates of the midpoints of the four most recent Great Plains droughts, along with the dates of the appropriate sunspot minima.

Table 3

Midpoints of Severe Great Plains North American Droughts and the Dates of the Apparently Associated Sunspot Minima

Sunspot Minimum	*Midyear of Drought*
1889	1892
1912	1912
1933	1934
1954	1953

The association shown in the table suggests that if the recurrence tendency persists we will face a drought in the mid-1970s. The actual onset of such a drought is still not clear, but preliminary evidence supports the idea that we have already entered a fairly serious drought. In 1974 there was a 6-week extended dry spell in June and July in Iowa, Kansas, Nebraska, and Oklahoma, which did several billion dollars worth of damage to crops. The summer of 1975 was fairly normal, with the exception of a rather damaging dry spell in Iowa. The winter of 1975–1976 has been relatively dry in Kansas and Eastern Colorado, with considerable damage to the winter wheat crop. At the time of this writing, it appears that the summer of 1976 had a drought of fairly significant proportions, with corn and soy bean yields being adversely affected, and if the next few summers display a drought cycle, this will coincide closely with the appropriate sunspot minimum.

The current sunspot minimum has been delayed past its expected date. Most solar forecasters predicted it for 1974 or 1975. However the recent outbreak of flares in March and April of 1976 demonstrates that the decaying cycle is not completely quiet yet; I shall return to a discussion of this point later in the paper.

Along with the 20- to 22-year recurrence of drought, there has been increasing evidence in recent years that the sunspot cycle influences surface temperatures in North America. The difficulty of finding convincing sun-related cycles in meteorological data has been pointed out by Monin and Vulis [34] and by Gerety et al. [35]. However, in spite of the weakness of the signals, there are some reports of approximately 11- and 22-year periodicities in North American temperature data. Mather [36] shows a rather strong 20-year periodicity in January surface temperatures in the Delmarva Peninsula. His data are in general agreement with the drought cycle, i.e. when the Great Plains are having drought, the Eastern Shore of Maryland, Delaware, and Virginia is having unusually mild Januaries.

In addition, the use of more sophisticated statistical techniques, such as maximum-entropy spectral analysis, has enabled some analysts to isolate rather sharp solar signals. For instance, Currie [37] found a cycle of 10.5 years in North American surface temperatures. Mock and Hibler [38] found a 20-year periodicity in January temperatures in eastern North America, similar to the results of Mather.

These studies, though still short of providing us with conclusive results, emphasize the importance of intensified effort to identify variable average solar-cycle activity as a forcing function in decadal time-scale climatic trends.

THE OBSTACLES

Two principal obstacles stand in the way of accelerated research commitments to the field of Sun-weather research. The first has to do with the complexity, inconclusiveness, and poor quality of much of the research in the field. This has led to widespread doubts that the purported findings merit further effort. As one researcher has said, the effects, if real, are very complicated; otherwise they would have been discovered long ago.

In my view, there are now sufficiently solid empirical results at hand to assure any qualified person who looks into it objectively that solar activity does influence the lower stratosphere and troposphere materially, and that this effect is large enough that inclusion of these processes may, when fully understood, improve extended-range weather forecasting and allow advances in the prediction of climatic changes and fluctuations.

Thus I argue strongly for the establishment in one or more universities or other research centers of at least one well-supported, ably directed, long-term research program that commits the efforts of a half dozen full-time researchers to a vectored effort to perform critical studies, preferably using inductive inference methods, as recommended by Platt [39], to gain an understanding of the physical mechanisms that produce solar-weather effects.

The second obstacle is the trivial amounts of energy available in solar activity fluctuations when compared to the energetics of the general circulation of the earth's atmosphere. Because of this there is, for example, no real prospect of finding a sufficiently strong brute-force heating mechanism to produce the trough cyclogenesis reflected in our Gulf of Alaska results. We must, thus appeal to "trigger effects" based on instabilities, or to time-modulation effects, in which the energy for a trough change is at hand and the process ready to go when the solar activity simply accelerates or delays the time sufficiently to produce the observed time-associations. This latter mechanism has been proposed by Hines and Halevy [22]. However, as they point out, there is still the need to explain the mechanism by which this time-modulation can occur.

It is well known that ionospheric conditions at levels above 80 km are essentially controlled by solar and cosmic effects. At lower levels, to which most solar activity effects fail to penetrate (cosmic rays being a notable exception), the atmospheric density is far higher and the solar effects fall orders of magnitude below the necessary energy levels for direct effects on the denser lower atmosphere.

In the face of these two principal obstacles, most meteorological centers in the United States have been unwilling to commit substantial research efforts to the problem. I believe now, however, there is justification for heightened effort. It may well be, in another decade or so, that extended forecasting will be unthinkable without consideration of solar activity.

STEPS FOR THE FUTURE

Physical Mechanisms

It is most important for the future to gain an understanding of the physical mechanisms involved in Sun-weather relationships. It may be, as is probable with climate changes, that no single mechanism is at work, but several; this will not make the search easier. Each of several proposed mechanisms merits intensified study.

Variations of the Solar Constant—It is obvious that changes in the total solar radiation incident on the earth would affect weather and climate. Changes of a few percent would probably change glaciation, ocean temperatures, and land weather materially. Large changes of the type common in other stars would destroy the biosphere. Therefore, a search for small but real changes of the solar "constant," on the order of 0.5% to 2%, merits dedicated observation. Widespread recognition of this need appears to be emerging.

Also important is to study the magnitude of possible solar output fluctuations in near-visible ultraviolet solar radiation. This will require spacecraft study and is a high-priority item.

Solar Activity and Atmospheric Scattering or Blanketing—Sudden cirrus clouding, if a change of state were triggered by solar activity, could release latent heat of condensation and freezing, possibly of sufficient energy to be of atmospheric dynamical significance. Moreover, atmospheric scatterers or clouds can produce thermal blanket-

ing, through increased infrared opacity. This can be sufficient under some circumstances to change the heating of the atmosphere near cloud level by as much as 1°C/day, a significant amount dynamically. If such clouds can be triggered by solar activity, that would be a plausible Sun-weather mechanism.

The observational fact of sudden formation of such blanketing after solar activity outbreaks needs to be subjected to observational test. It is an urgent question in the search for mechanisms.

Ozone Destruction—Recent researches have shown that ozone destruction by a major proton flare can be significant, as shown by Angione et al. [40]. Moreover, there are some indications of solar-cycle ozone changes. Because of ozone's important role in atmospheric radiation balance there is justification for added effort here.

Atmospheric Electricity and Thunderstorms—As mentioned earlier, the global thunderstorm activity, if solar activity modulated, could be the clue to trigger effects of genuine importance in the lower atmospheric circulation. R. Markson (1973 and private communication) has pointed out that ionizing radiation reaching the tops of large tropical thunderstorms could increase the current flow from these storms that maintains the ionospheric potential and that this increased earth-atmosphere potential difference could, due to resultant changes in the atmospheric potential gradient near growing convective storms, enhance thunderstorm numbers or intensities. This is a highly promising direction to pursue observationally and theoretically.

Other Mechanisms—Various other mechanisms have been suggested, such as infrared heating of the lower atmosphere by radiation from the ionosphere during periods of disturbance. Similarly, dynamical interconnections among the ionosphere, mesophere, and stratosphere merit attention in the context of Sun-weather research.

One perplexing finding that emerges from the various Sun-weather studies is that the stratospheric-tropospheric response is very often extremely rapid. Our own studies of geomagnetism and flare effects in the hemispheric vorticity area index show that maximum atmospheric response often occurs on the first day. Other workers have suggested responses within a few hours. This is puzzling because it probably rules out modulation of the incoming and outgoing atmospheric radiation as a link in the causality chain. However this needs stronger inferential analysis.

New Indices

The time has come, in my view, to abandon simple efforts to "prove" that solar activity affects weather. I consider this step completed by the Hines-Halevy [22] work and other recent results. The urgent matter now is to direct critical statistical experiments to the elimination of some possible causal mechanisms and the discovery of new ones.

In this regard, indices like the hemispheric vorticity area index, have largely served their purpose. They were designed to be maximally objective, insensitive to data inhomogeneities, and global. We now need indices that get closer to known dynamical processes of the atmosphere and that are specific to geographical regions or phenomena that are of special importance in atmospheric dynamics.

As Wilcox has frequently stated, improved synoptic data of less conventional atmospheric parameters are an important need in seeking and testing mechanisms in the Sun-weather field. My own work with R. H. Olson on our proposed cirrus cloud radiation mechanism has, for example, been seriously impeded by lack of homogeneous daily data on the infrared flux to space over the Gulf of Alaska. If we had been able to have three successive winter half-years of course- or moderate-resolution data over this area, we could probably have confirmed or rejected this process as a relevant mechanism.

Among the desired synoptic data are measures of atmospheric electrical parameters, aerosol content with height, atmospheric radiation data, etc.

Case Studies

At 11 a.m. on September 1, 1859, a great solar flare, visible in white light as very few flares are, burst into view and was observed by R. C. Carrington of the Royal Observatory in England. The magnetic-field recorders at Kew Observatory fluctuated briefly with this event. Two days later a violent magnetic storm was recorded at the Kew

Observatory. Balfour Stewart, Kew director, concluded that terrestrial magnetic disturbances could be caused by flares. The idea was so unbelievable at the time that most researchers dismissed it as coincidence. Yet this event was instrumental in launching the modern history of Sun-Earth research and in verifying the now well-established connections of flares with geomagnetism.

It may be that the analog of the Carrington flare, so far as Sun-weather work is considered, occurred on March 23, 1976. Prior to that date solar activity had dipped to a low level, and many, myself included, felt that the eighth recurrence of the roughly 22-year recurrent Great Plains U.S. drought was at hand. The phase of the sunspot cycle (minimum phase of alternate spot cycles) that coincided with the Great Plains droughts of the 1950s, the dust bowl years of the 1930s, and earlier droughts in this region was at hand. Great Plains soil moisture was low. Much of the winter wheat was drought-destroyed. Continued hot, dry weather and high winds would put the Great Plains in serious trouble. Then suddenly on March 23, 1976, a huge X-ray flare broke forth near the east solar limb. On March 27 there was a brilliant and widespread aurora and a large geomagnetic storm. This was solar activity of the dying sunspot cycle, but it was large activity characteristic not of sunspot minimum but of high solar activity.

We have not yet made a careful synoptic weather study, but a cursory examination of the 500-mbar level suggests that in the Gulf of Alaska area the circulation changed from highly zonal to more cyclonic (meridional) very soon after the March 26 geomagnetic storm. Strong westerly winds appear, also in cursory study, to have ceased on the Great Plains, and heavy rains came in the northern and southern reaches of the region.

It is my contention that an intensive cooperative retrospective world geophysical interval might reveal, through a case study, important clues to the lower atmosphere's response to a large, isolated solar outbreak. Perhaps we could find at what levels and in what parameters the responses (if indeed they occurred) appeared. The interval might cover 6 weeks on either side of the March 23 flare and the abrupt commencement of solar activity.

Other individual case studies may also be of value. I include as a candidate for retrospective Sun-weather studies an intense, isolated solar outbreak of August 4, 1972, which might, because it was in the Northern Hemisphere summer, give us prospect of ascertaining whether there are Southern Hemisphere Sun-weather effects. A similar outbreak on July 3, 1974, would also be a useful subject. Other similar events can be identified, including some notable events from earlier times.

We are at the stage in Sun-weather research at which individual case studies may lead to fruitful conclusions regarding the operative physical processes. Nor is other Sun-weather work being forcefully pursued in this country. Little is being done today in the Sun-weather field, due both to lack of available funds and to still widespread skepticism among many, if not most, of the leaders and pacesetters in meteorology. Obviously the stakes are high in extended-range weather forecasting and in climate prediction. If solar activity can contribute to advancement in either area, it will amply justify greatly expanded effort.

REFERENCES

1. J. A. Eddy, "The Maunder Minimum," *Science* **192**, 1189–1202 (1976).
2. H. Friedman, "Solar-Terrestrial Physics," in *ONR's 30th Anniversary—Science, Technology, and the Modern Navy*, ONR, Arlington, Va. 1976.
3. J. Gribbin, *Nature* **259**, 367–368 (1976).
4. A. Monin, *Weather Forecasting as a Problem in Physics*, MIT Press, Cambridge, Mass., 1973.
5. W. Herschel, "Observations Tending to Investigate the Nature of the Sun, in order to find the Causes of Sumptoms of its Variable Emission of Light and Heat,*Roy Soc. Phil. Trans.* (1801).
6. W. Koeppen, "Uber Mehrjahrige Perioden der Wiherung, Insbesondere uber die 11-Jahrige Periode der Temperatur," *Zeit der Osterreichischen Gesellsch. fur Meteorol.* **8** (1973).
7. B. Duell and G. Duell, "The Behavior of Barometric Pressure During and After Solar Particle Invasions and Solar Ultraviolet Invasions," *Smithsonian Miscellaneous Collections* **110**(8) (1948).

8. G. Archenhold, "Untersuchungen ueber den Zusammenhang der Haloerscheinungen mit der Sonnenaktigheit," *Gerland's Beitr. Geophysik.* **53**, 395–475 (1938).
9. C. G. Abbot, "Magnetic Storms, Solar Radiation, and Washington Temperature Departures," *Smithsonian Miscellaneous Collections* **110**(6), (1948).
10. D. R. Barber, "Changes in Brightness, Polarization, and Colour of the Zenith Day Sky Accompanying Geomagnetic Activity," *J. Atm. Terr. Phys.* **7**, 170–172 (1956).
11. R. Shapiro, "Further Evidence of a Solar-Weather Effect," *J. Meteor.* **13**, 335–340 (1956).
12. R. Shapiro, "A Possible Solar-Weather Effect," *J. Meteor.* **11**, 424–425 (1954).
13. R. Shapiro, "A Comparison of the Response of the North American and European Surface Pressure Distributions to Large Geomagnetic Disturbances," *J. Meteor.* **16**, 569–572 (1959).
14. R. A. Craig, "Surface Pressure Variations Following Geomagnetically Disturbed and Quiet Days," *J. Meteor.* **9**, 280–290 (1952).
15. D. D. Woodbridge, T. W. Pohrte, and N. J. MacDonald, "A Possible Effect in 300mb Circulation Related to Solar Corpuscular Emission," Institute for Solar-Terrestrial Research, High Altitude Laboratory, Tech. Rep. No. 3, 1957.
16. N. J. MacDonald and W. O. Roberts, "Further Evidence of a Solar Corpuscular Influence on Large-Scale Circulation at 300 mb," *J. Geophys. Res.* **65**, 529–534 (1960).
17. R. H. Olson, W. O. Roberts and C. S. Zerefos, "Short-Term Relationships Between Solar Flares, Geomagnetic Storms, and Tropospheric Vorticity Patterns," *Nature* **257**, 113–115 (1975).
18. E. R. Mustel, "On the Reality of the Influence of Solar Corpuscular Streams Upon the Lower Layers of the Earth's Atmosphere," USSR Academy of Sciences, Astronomical Council, Moscow, Pub. No. 24, 1972.
19. W. O. Roberts and R. H. Olson, "New Evidence for Effects of Variable Solar Corpuscular Emission on the Weather," *Rev. Geophys. Space Phys.* **11**, 731–740 (1973).
20. C. O. Hines, "Wind-induced Magnetic Fluctuations," *J. Geophys. Res.* **70**, 1758–1761 (1965).
21. J. M. Wilcox, "Solar Structure and Terrestrial Weather," *Science* **192**, 745–748 (1976).
22. C. O. Hines and I. Halevy, "Reality and Nature of a Sun-Weather Correlation," *Nature* **258**, 313–314 (1975).
23. R. Reiter, "Increased Influx of Stratospheric Air into the Lower Troposphere after Solar H_2 and X-ray Flares," *J. Geophys. Res.* **78**, 6167–6172 (1973).
24. R. Reiter, "Increased Frequency of Stratospheric Injections into the Troposphere as Triggered by Solar Events," *J. Atmos. Terr. Phys.*, to be published in 1976.
25. M. Bossolasco et al., "Solar Flare Control of Thunderstorm Activity," Institute Universitario Navale di Napoli, Institute di Meteorologia E. Oceanografia, pp. 213–218, 1972.
26. R. Markson, in "Possible Relationships between Solar Activity and Meteorological Phenomena," Goddard Space Flight Center Symposium Report, NASA SP-366, 171, 1973.
27. J. R. Borchert, "The Drought of the 1970's," *Ann. Ass. Amer. Geog.* **61**, 1–22 (1971).
28. J. R. Marshall, University of Kansas, Ph.D. Thesis, 1972.
29. W. C. Palmer, "Meteorological Drought," U.S. Weather Bureau, Res. Pap. No. 45, 1965.
30. W. O. Roberts, in "Possible Relationships Between Solar Activity and Meteorological Phenomena," pp. 3–23, Goddard Space Flight Center Rep., NASA SP-366, 1973.
31. H. E. Thomas, "The Meteorological Phenomena of Drought in the Southwest," Geological Survey Pap., Prof. Pap. 372-A, U.S. Geological Survey, 1962.
32. H. C. Willett, in Encyclopedia of Atmospheric Sciences and Astrogeology, pp. 869–878, R. W. Fairbridge, ed., Reinhold, New York, 1967.
33. L. M. Thompson, "Cyclical Weather Patterns in the Middle Latitudes," *J. Soil Water Conservation* **28**, 87–89 (1973).
34. A. S. Monin and I. L. Vulis, "On the Spectra of Long-Period Oscillations of Geophysical Parameters," *Tellus* **23**, 337–345 (1971).
35. E. J. Gerety, J. M. Wallace, and C. S. Zerefos, "Sunspots, Geomagnetic Indices and the Weather: A Cross-spectral Analysis Between Sunspots, Geomagnetic Activity and Global Weather Data," *J. Atm. Sci.*, to be published, 1976.
36. J. R. Mather, *Climatology, Fundamentals and Applications*, McGraw-Hill, New York, 1974.
37. R. G. Currie, "Solar Cycle Signal in Surface Air Temperature," *J. Geophys. Res.* **79**, 5657–5660 (1974).
38. S. J. Mock and W. D. Hibler, "The 20-Year Oscillation in Eastern North American Temperature Records," *Nature* **261**, 484–486 (1976).
39. J. R. Platt, "Strong Inference," *Science* **146**, 347–353 (1964).
40. R. I. Angione, E. J. Medeiros, and R. G. Roosen, "Stratospheric Ozone as Seen from the Chappuis Band," *Nature* **261**, 289–290 (1976).

MATERIAL SCIENCES



John B. Wachtman, Jr., is Chief of the Inorganic Materials Division of the National Bureau of Standards. He directs a research program on measurement techniques, standards, and data relating to the chemical reactions, processing, characterization, and physical properties of inorganic materials including ceramics, glass, optical materials, and electronic materials. His personal research has been on mechanical properties of hard materials. He has received distinguished awards from the National Bureau of Standards, the Department of Commerce, and the American Ceramic Society. He is currently consulting with the Materials Program of the Office of Technology Assessment having been OTA Program Manager in 1974 and 1975. Dr. Wachtman earned a B.S. and an M.S. in Physics at Carnegie-Mellon University and a Ph.D. in Physics at the University of Maryland. He is a member of the National Academy of Engineering, past President of the Federation of Materials Societies, and President-elect of the American Ceramic Society.



James R. Johnson is Executive Scientist and Director of the Advanced Research Programs Laboratory, Central Research Laboratories, of the 3M Company. Dr. Johnson is author of more than 50 publications, holds 20 patents, and has received distinguished awards for contributions to his profession. He received B.S., M.S., and Ph.D. degrees in Ceramic Engineering from Ohio State University. He is a member of the National Academy of Engineering and a Fellow and past President of the American Ceramic Society.

CERAMICS IN THE FUTURE

John B. Wachtman, Jr.
National Bureau of Standards
Washington, D.C.

James R. Johnson
3M Company
St. Paul, Minn.

Abstract: Ceramics in the broad sense of inorganic, nonmetallic materials already play many vital roles in military and civilian technology. The first 30 years of the Office of Naval Research's existence coincided with great progress in the development of advanced ceramics for special applications such as computers, optics, electronics, etc. Important developments in ceramics for bulk uses have also occurred, e.g., refractories for the basic oxygen process and glass reinforcing fibers.

Prospects for future development are discussed in terms of a matrix structure that considers promising scientific opportunities as one dimension and promising technologies as another. The discussion is developed in an overall context of concern with energy, declining supplies of some high-grade ores, general pollution effects, and specific concern for toxic substances.

It is concluded that further advances in high-technology ceramics with important practical payoffs should occur. In addition, high volume use of advanced bulk ceramics seems possible. The extent of such use may be determined as much by public attitudes as by strict technical advantage.

Ceramics are described, in the broad sense, as materials that are inorganic and nonmetallic (in terms of bonding, not electrical conductivity). It includes not only traditional ceramics, but also glass, portland cement, and many electronic, magnetic, and optical materials. Ceramics have strong ionic or covalent bonds. Although exceptions can be found, ceramics frequently have high melting points, are resistant to chemical attack, and exhibit good wear and creep resistance but undergo brittle fracture. Combinations of these and other chemical and physical properties deriving from their bond character give them both unique value and sharp limitations for many applications.

Many advances in ceramics have occurred during the 30 years of existence of the Office of Naval Research. Some new developments have used ceramics directly, while others have used them as components of various composites. Improved ceramics include the family of pore-free polycrystalline ceramics, beginning with alumina in the early 1960s. One of the first major applications was as a container for sodium vapor in high-efficiency lights. Subsequent developments have included both polycrystalline active laser materials and windows, lenses, and prisms for high-intensity lasers. Ceramic radomes are another example. Still another is the oxide fuel which is the basis for all present commercial and military nuclear reactors. Ferrite cores for computer memories played a vital role in several generations of large computers, and silicon-based electronic chips are the heart of the present rapid development of minicomputers. The practical commercial use of the basic oxygen steel-making process required the development of a new class of refractories—the tar-bonded magnesium oxides. Examples of ceramics in composites include fiberglass-reinforced plastic, which has

revolutionized the construction of small boats. Carbide cutting tools, composed of a ceramic such as tungsten carbide in a metal matrix such as cobalt, have largely replaced steel tools in machinery. All-ceramic tool materials, such as aluminum oxide, also play a specialized role and are likely to increase in use.

Other examples of recent developments in ceramics could be given, but the purpose of this paper is to look ahead. If the authors could describe in detail the future development of new technical materials they would be prophets indeed. Disavowing this intention, the writers instead have attempted to identify broad trends and scientific opportunities and to discuss these in terms of a series of areas of promise in specific types of applications.

Actual commercial development, of course, depends on much more than technical advances. In addition to the question of cost, there are the questions of altering basic concepts of engineering design and breaking with tradition. Two aspects of this latter question are especially pertinent to the future of ceramics. One is the question of first cost vs life-cycle cost. The use of porcelain enamel could extend life and lower life-cycle cost in certain applications such as mufflers. The other is the tradition of design and use with structural materials. Recently developed ceramics coupled with new design, fabrication, and use procedures could lead to wider structural application as will be discussed later. Professional and public attitudes may be as important as technical and economic factors in determining the future of ceramics in some applications.

GENERAL TRENDS AFFECTING THE FUTURE OF CERAMICS

Two major trends affecting the future uses of materials are the rising costs of energy and raw materials and the increasing dependence of the United States on imports of fuels and many nonfuel minerals. The outlook for many materials is that future production will depend on mining enormous volumes of low-grade ores, with necessarily higher cost, energy expenditure, and environmental impact [1]. It appears that ceramics may be favored in this competition because of both plentiful raw materials and some advantage in energy costs associated with production [2-4].

The basic argument for the plentiful supply of raw materials for ceramics is that the components of most ceramics are among the most abundant in the earth's crust. The most common elements in the earth's crust are listed in order of occurrence in Table 1 [5]. For comparison, the major elemental components of a variety of ceramics are listed in Table 2 [6, 7]. A striking correlation is apparent; the major components generally include the 10 most common elements. To this must be added the fact that nitrogen, the only component of the promising sialon family of ceramics not in the first 10, is available in essentially inexhaustible amounts from the atmosphere. There are, of course, important ceramics that use less common elements, such as chromium-containing refractory brick.

There is much more to the question of raw materials than crustal abundance. To be economically usable, material must be available in proper chemical form, in sufficient concentration, and in an appropriate location. There does seem to be a general relationship, developd by McKelvey and updated by Erickson, between reserves and crustal abundance [8]. This relationship states that the resource potential R of an element (in metric tons) is related to the crustal abundance A (in parts per million) by $R = 2.45A \times 10^6$. However, several qualifications of this approximate, empirical formula exist. For example, it refers to resources currently recoverable (with present technology and under recent economic conditions), and deviations associated with the inherent geochemical nature of a particular element may exist. Nevertheless, the relationship does suggest that the long-term prospects for ceramic raw materials are generally good.

The rising cost of energy and the increasing dependence of the United States on imported fuels has focused attention on the energy required to produce materials. Table 3 lists estimates by Hayes [3] and Samples [9] of the total energy required to produce 1 ton of each of a variety of metals and nonmetals. This table shows that the energy required to produce ceramic products is generally less than for the commonly used metals. The difference is likely to increase as higher grade

Table 1

The Relative Abundance of the Principal Elements in the Earth's Crust, Expressed in Percent by Weight [5]

Element	*Percent by Weight in Earth's Crust*	*Element*	*Percent by Weight in Earth's Crust*
O	46.0	Ce	0.0043
Si	27.0	La	0.0039
Al	8.3	Nd	0.0026
Fe	5.8	Co	0.0025
Ca	5.2	Y	0.0024
Mg	2.8	Li	0.0021
Na	2.3	Nb	0.0019
K	1.7	N	0.0018
Ti	0.64	Sc	0.0018
C	0.28	Ga	0.0018
Mn	0.13	B	0.0013
P	0.12	Pb	0.0012
Sr	0.048	Sm	0.00067
F	0.045	Gd	0.00067
S	0.040	Th	0.00058
Ba	0.039	Pr	0.00057
Cl	0.028	Br	0.00044
V	0.014	Dy	0.00041
Zr	0.013	Er	0.00027
Cr	0.011	Yb	0.00027
Zn	0.0094	As	0.00022
Ni	0.0089	Sn	0.00017
Rb	0.0078	U	0.00017
Cu	0.0063		

Table 2

Major Components of Some Commonly Used or Promising Ceramics

General Name	*Typical Major Chemical Components [5,6]*	*Typical Uses*
Common Brick	Si, Al, O	Building
Window Glass	Si, Ca, Na, O	Building
Refractories	Al, O; Al, Si, O; Mg, O; Al, Si, Zr, O; Cr, Mg, Fe, Al, O; Si, C	Furnaces Crucibles Molds
Porcelain Enamel	Si, Al, Ca, K, O; sometimes also Na, Pb, and/or B	Protective coatings
Chemical Stoneware	Al, Si, Na, O	Chemical processing
Laboratory Glassware	Si, B, Na, O	Chemical processing
Electrical Porcelain	Al, Si, Mg, O	Power distribution, electronics
Dielectrics	Ti, Ba, O	Electronics
Semiconductors	Si, Al, P; Ge, Ga, As, etc.	Electronics
Silicon Carbide	Si, C	Resistors, abrasives
"Sialons"	Si, Al, N, O	Bearings, turbine blades (potential)
Optical Devices	Si, Na, Ca, O; IN, Pb; CdS; Si, AlP; Al, Ba, As; etc.	Communications, Solar cells
Magnetic Ceramics	Fe, Al, Y, O; etc.	Memories
Portland Cement	Si, Al, Ca, O	Construction

metal ores are depleted and more energy is required to work lower grade ores. Depletion of lower grade ores for ceramics should occur much more slowly.

Net energy analysis thus indicates that there is a real driving force for the substitution of ceramics for other materials. This generalization is subject to many qualifications, however. Comparison of the figures given by Hayes and by Samples shows that net energy analysis is still subject to considerable variation in results. These energies are not thermodynamic values but are the sum of estimates for a series of processes of varying efficiency. For example, Samples gives the following

Table 3

Energy Requirements for Selected Materials According to Hayes [3] and Samples [9]

Commodity	*Energy Required [3] 10^6 BTU/ton*	*Commodity*	*Energy Required [9] 10^6 BTU/ton*
METALS			
Steel Slab	24	Cold Rolled Steel	53
Aluminum	244	Rolled Aluminum	220
Zinc	65	Rolled Zinc	91
Lead	27	Lead	44
Copper, Refined	112	Rolled Copper	131
Chromium, Low Carbon Ferroalloy	129		--
Magnesium	358		--
Manganese, Electric Furnace	52		--
Titanium	408		--
Uranium, Acid Circuit	776		--
NONMETALS			
Quicklime	8.5		--
Portland Cement	7.6		--
Common Brick	3.5		--
Glass Containers	17.4		--
Refractory; Basic Brick	27		--
Refractory; Fireclay	4.2		--
	--	Vinyl Chloride	64

breakdown of his total figures of 53 x 10^6 BTU/ton for cold rolled steel (all in units of 10^6 BTU/ton): 1.4 for mining, 1.6 for coking, 20.0 for the blast furnace, 4.7 for the steel furnace, 2.1 for other materials for the steel furnace, 16.2 for hot rolling, 5.7 for cold rolling, and 1.3 for all transportation. Most of the difference between the two figures for steel in Table 3 arises because energy required for rolling and transportation is not included in the smaller figure. The question is not which figure is "right," but precisely what comparison is being made and which energy requirements must be included for this purpose of the comparison. A second important qualification is that the net energy comparison should be made on the basis of the amount of material required for the function to be performed rather than on an equal-mass basis. For example, if the function is to carry a load, the

strengths of the materials in question affect the masses required, and a detailed calculation, including stress analysis, is needed before a proper net energy comparison can be made.

With these qualifications in mind, it still seems clear that the relatively low energy requirement of ceramic materials favors their increasing use where possible in place of materials requiring greater energy, in view of the likelihood of still further increases in the price of energy.

Many institutional factors affect possible increased use of ceramics. Perhaps the most striking of these is the effect of regulations intended to protect and enhance environmental quality and human safety. For example, it seems probable that there will be continuing attempts by the Federal Government to regulate the production and control of hazardous substances. As more is learned about long-term trace toxicity, new laws and regulations with great impact on the competition among materials may be established. In this case, one might be tempted to speculate that the general tendency would be to favor ceramics because of their relatively inert chemical nature. On the other hand, consideration of the health hazards of asbestos and a few ceramic dusts such as fine silica suggests the danger in generalizing. The whole issue of toxicity and its impact on materials is likely to hold some surprises for unsuspecting materials users.

FRAMEWORK FOR TREATMENT OF POSSIBLE FUTURE TRENDS IN THE USE OF CERAMICS

New uses of ceramics requires commercial as well as purely technical innovation. The process of overall innovation is generally regarded as requiring a match between the pull of a need (including the perception of a market) and the push of a technical opportunity. Our treatment of possible future trends in the use of ceramics will be developed in these terms. Detailed treatment would take us into the area of proprietary product development. Instead, we will treat the subject in terms of broad categories of technical opportunities and of practical needs.

The conceptual scheme is illustrated in Figure 1, which displays a matrix whose rows are technical opportunities and whose columns are practical needs. Both lists are certainly incomplete; the authors hope their selection includes at least some of the most important items and gives a broadly correct view of the most significant trends. Our procedure is first to discuss major areas of scientific and technical promise and then to discuss areas of need in terms of the potential of ceramics to meet the needs.

Figure 1a.

	Categories of Practical Need (See list below)					
Categories of Scientific Promise						

Figure 1b. Categories of Scientific Promise

Powder Preparation
Processing
Property-Molecular Structure Relationship
Property-Microstructure Relationships
Fracture Behavior
Heterogeneous Reactions

Figure 1c. Categories of Practical Technical Need

Energy Systems
Transportation Systems
Environment Systems
Communication Systems
Metallurgical and Other Processing Systems
Structural/Composite Systems
Waste Management Systems
Electrical Systems
Information Systems
Medical Systems

Figure 1—Framework for discussion of possible trends in the use of ceramics.

THE TECHNICAL PROMISE OF CERAMICS

Powder Preparation

Kingery has noted that, while there are exceptions to any general statement about ceramics, there is much merit in the traditional idea of ceramics as the product of powder processing [10]. Most ceramics are made by a process of blending powders, followed by extrusion or a cold compaction stage (slip casting or cold pressing), and concluding with heat treatment with or without pressure. Other forming processes are used, including melting and casting, but sintering remains the predominant process.

Powders of nominally the same composition can differ greatly in their suitability for sintering. A very small average particle size is generally desirable, to allow the production of a small grain size and consequently high strength in the final product. A suitable distribution of particle sizes is needed to minimize porosity in the cold compact and assist the sintering process. For some applications (such as optical and electronic devices) a very high purity may be desired. In addition there is the phenomenon of highly active powders, which sinter at lower temperatures and more completely than nominally similar powders prepared by different means. The phenomenon is not completely understood but is thought to involve defect structure, anion impurities, and/or adsorbed chemical species [11]. All of these are difficult to measure, and basic work in this area is needed. It seems very likely that substantially improved powders are possible and that advances in characterization tools (including advances in spectroscopy and signal-to-noise enhancement) open the way to better understanding of powder reactivity and, in turn, to substantially improved ceramics. The potential impact extends across the entire range of polycrystalline ceramics because there is probably no ceramic in substantial use today whose room-temperature strength could not be improved by several hundred percent if its microstructure could be optimized.

Processing

Ceramic science has led to impressive advances in recent years, as typified by two examples. Strength of high-quality commercial alumina has been increased from a typical value of 20 to 30 kpsi into the range of 60 to 100 kpsi. The latter values are routinely produced on automated production lines for electronic substrates. Pore-free alumina and other pore-free ceramics have been achieved, making possible the development of high-temperature sodium vapor lamps and polycrystalline laser materials. In another impressive achievement, fine-grain, high-strength polycrystalline ceramic windows, lenses, and prisms have been produced by conversion of single crystals to polycrystalline form by strain-anneal techniques [12]. On the theoretical side there has been a great deal of activity on models to describe quantitatively the successive stages of the sintering process. These models are less successful when combined to apply to the overall processing of ceramics.

Impressive progress in ceramic processing has been made, but it still seems fair to say both that a general capacity to produce desired microstructures is lacking and that there is good reason to expect further progress. A brief description of the process for injection molding and reaction bonding of silicon nitride (RBSN) will illustrate an important advance in ceramic processing [13]. A polymer is filled with silicon powder (up to 70%) which is injection molded at modest temperature (about 100 to 200°C) into a die that produces a preform of the dimensions of the final part. A second heating at about 300°C removes the polymer. The preform is then heated in a nitrogen atmosphere at 1300 to 1450°C for 24 to 48 h. The final dimensions of the part are typically within 0.1% of those of the preform.

Still another emerging process is the use of sols that are gelled in desirable form, such as microspheres or fibers, and fired. Such products have unusually fine grain structure and sinter at temperatures as low as 500°C for many oxides.

Property-Molecular Structure Relationships

The solid-state chemistry of ceramics is usually complex. Typically the unit cell is large and contains a relatively large number of atoms of several species. Ideally, the goal of research in this area should be to predict the conditions under which the material forms, its structure, its stability, and

its electronic, optical, magnetic, mechanical, and chemical properties [14]. This goal is far from being reached, even though some striking successes have been achieved (e.g., calculations of band structures in relatively simple crystals). For the more complex ceramics empirical correlation of properties with structure, combined with qualitative reasoning from fundamentals (e.g., molecular orbitals) and structural rules, is characteristic of current knowledge [15]. The promise of this field seems very great, and the term "molecular engineering" has been coined to describe the process of designing and producing new materials based on chemical and structural principles. The continuing revolution in electronics, optics, communication, and computation is based on ceramic materials that did not exist before 1945. There is no reason to suppose that present materials represent the ultimate possible performance; many families of promising materials remain to be investigated [16, 17]. The field of semiconductors has moved from germanium to silicon and broadened to include compound semiconductors such as gallium arsenide, and has given rise to light-emitting diodes (gallium phosphide) and semiconductor lasers. The field of magnetic ceramics produced polycrystalline ferrites, which provided the memories for several generations of computers, and the promising family of magnetic "bubble" materials. The field of optical ceramics has developed rapidly in the last decade to produce solid-state laser hosts (ruby, yttrium-aluminum-garnet, etc.), polarizers, modulators, detectors, optical waveguides, and integrated optics.

Introduction of these materials into composites has also yielded new products. For example, ferrites plus elastomers make "plastic" magnets. Packages for microcircuits are combinations of ceramics, glass, metals, semiconductors, and plastics.

The central point is that the highly variable chemistry and generally low electromagnetic losses in ceramics provides great opportunities for tailoring materials to provide complex combinations of electromagnetic properties and that both the science of the chemistry-property relationships and the technical opportunities appear to offer great promise for future development.

Physical properties are sometimes divided into those that depend strongly on defect structure (such as diffusion) and those discussed above, which derive primarily from the perfect molecular or crystal structure. This is a useful distinction and will be followed here, but it is well to remember that there may be a close relationship as illustrated by ionic conductivity in β-alumina. This material is the basis for extensive work as the electrolyte in the very promising sodium-sulfur battery. The system Na_2O-Al_2O_3 contains several compounds, one of which is termed β-alumina, and possesses very high sodium ion conductivity at most temperatures (about 300°C). This conductivity is a consequence of the fact that the perfect structure has sodium ions filling only a fraction of available interstitial sites in an orderly pattern. It is also a consequence of the fact that a defect structure with some of the sodium ions out of their ideal, orderly positions, is easily formed thermally. Systematic molecular engineering to find crystal structures that lend themselves to the formation of defect structures is also a very promising field.

Property-Microstructure Relationships

To simplify discussion it is useful to consider microstructure in terms of fine microstructure (point defects, dislocations, etc.) and gross microstructure (grain size, porosity, microcracks, etc.), even though there is no sharp dividing line and grain boundaries must be considered to belong to both families.

Point defects are generally of interest in relation to mass transport (diffusion and creep), to localized electronic energy levels providing donors, acceptors, and traps involved in conductivity or optical behavior. This field has received considerable attention, but it has generally proven very difficult to define the rate-controlling species in high-temperature processes and to work out the thermodynamics and kinetics of their formation. This remains a very promising field both because most of the science of defect chemistry remains to be worked out in detail and because the resulting ability to improve control of solid state processes should be important.

Grain boundaries play an important role during the processing of ceramics and affect their properties after processing. Their chemistry depends

largely on the behavior of solutes in the bulk phase (i.e., point defects), about which, as we have stated, little is known. We do know that grain boundaries contribute to substantial deformation at high temperature and that diffusional processes can occur rapidly at high temperatures, influencing to a great extent processing behavior and properties [10, 18].

The larger aspects of microstructure, including grain size, porosity, and microcracks, are especially important to mechanical properties. Great progress has been made in correlating average behavior, such as average strength or steady-state creep rate, with average grain size and porosity. Deformation at high temperatures can take place by a variety of processes, including bulk diffusion, grain-boundary diffusion, and dislocation motion. Usually two or more are acting simultaneously. A considerable body of theoretical models exists. Progress seems to depend on good (and difficult) experimental work with sufficient range of materials and imposed conditions (such as stress and temperature) combined with characterization adequate for sorting out the dominant processes and for testing and improving the models. There seems good reason to believe that much better understanding of deformation processes will be achieved. This may permit maximization of deformation when desired (i.e., for hot pressing), minimization when it is desired (i.e., load carrying at high temperatures), and the ability to choose the optimum feasible combination when both are desired.

Fracture Behavior

Although average strength, as mentioned above, frequently correlates well with average microstructure, this correlation has serious deficiencies as a basis for controlling or understanding strength. Control is difficult because strengths of individual specimens can deviate widely from the average and because strength decreases with time in a variable manner, depending on a variety of circumstances, including load and chemical environment. Understanding strength in terms of average features is difficult because strength is clearly determined by fracture beginning at extreme rather than average features.

Important progress has been made through the study of artificially induced cracks. It has proven possible to study crack propagation and to develop quantitative laws relating crack propagation rate to stress and chemical enviroment. Based on these laws it is possible to remove the weak specimens from a group and to calculate the minimum strength of the remaining specimens. It is also possible to calculate the minimum long-time strength from short-time tests [19]. Based partly on this new knowledge, a procedure for high-performance structural ceramics is beginning to emerge [20]. It requires detailed stress analysis, appropriate design (avoidance of stress concentrations and sometimes redundancy for tolerance of failure of an individual part), careful quality control (sometimes requiring a combination of nondestructive evaluation and proof testing), and care not to exceed design loads. Much detail remains to be developed to reduce this to routine engineering practice, but the promise is there. On the fundamental side, better understanding of the origin and behavior of very small flaws should lead to further improvement in the short-term and long-term strength.

Heterogeneous Reactions

Ceramic surfaces are important both as the site of unwanted reactions (deterioration) and of desired reactions, including catalysis. Ceramics are frequently used as high-surface-area carriers for catalysts, with increasing recognition that the carrier sometimes plays an important role in the catalytic process [11]. In addition, ceramics are sometimes used directly as catalysts rather than as carriers. Knowledge is largely proprietary and empirical. Progress will require improved understanding of the interplay of surface reactions with local detailed atomic and electronic structure.

CATEGORIES OF PRACTICAL NEED

The above brief and incomplete discussion of the scientific and technical promise of ceramics allows us to suggest a general framework for considering possible future practical applications. Progress on processing, combined with a practical means for using brittle structural materials,

will open many new applications to ceramics. These include not only primary load-carrying applications but also applications in which strength is only a secondary requirement. This upgrading of ceramics need not be confined to high-performance applications; substantial improvements in bulk, relatively low cost ceramics with attendant savings in weight appear possible. Along with this mechanical improvement, significant improvements in electromagnetic, transport, and chemical properties are to be expected. Fundamental advances alone will not bring new materials into practical use, however. Progress in practical application is likely to be incremental, occurring when a match between practical need and fundamental opportunity attainable at competitive cost is recognized. We turn now to discussion of a series of practical technologies where such advances are being made or appear likely in the future.

Energy Systems

Maintaining a stable economy requires adequate energy supplies. Rapid consumption of our finite worldwide fossil fuel resources has led to curtailments of fuel supplies, to economic pressures, and to the need to develop sources of energy alternate to the oil and gas currently used. Because energy systems are huge and complex, the time required to effect significant changes is very long. Various analyses indicate the primary importance of conservation measures. Table 4 shows estimated energy impacts for development of various alternate energy systems [21]. Nearly all of these systems and measures can benefit from advanced ceramic technology. Thus, the energy field presents a major high-priority challenge to the ceramic scientist and engineer [22-34].

Conservation can be practiced in all parts of the overall energy system. In the domestic sector, insulation of homes has been encouraged. Glass and mineral fibers are among the safest and most cost-effective products used for this purpose. Composite boards containing ceramic fibers or bubbles, combining structural with insulation properties; new forms of insulating concrete; and vapor-coated heat-conserving glazing are among the new applications for ceramics. Opportunities may exist for storage of heat in ceramic materials. In Europe, electric heaters in homes have large heat storage capacity in the form of dense ceramic bricks. Industry offers still more opportunities. In addition to kinds of building insulation similar to that described above, there are heat recovery systems, which recently have become cost effective as energy costs have escalated. Many of these must operate in high-temperature, corrosive environments where ceramics have the required resistance to give long life. Ceramic heat exchangers made of lithium aluminum silicates, cordierite, and other refractory materials are under evaluation.

Electric load management by the utilities as well as use of electric vehicles can offer still another form of conservation. Several systems under study use ceramic materials; for example, β-alumina is used as a solid electrolyte in the sodium-sulfur battery, and various ceramic fabrics (e.g., boron nitride) are under test as separators in the lithium-aluminum-sulfur battery. Graphite electrodes, carbide and oxide electrocatalysts, ceramic separators, and solid ceramic electrolytes may be used in fuel cells. These advanced batteries and fuel cells require additional technology from the ceramic industry.

In automobiles, it is believed that ceramic-fiber-reinforced aluminum composites or plastic composites will save fuel by making possible lighter weight vehicles.

Petroleum processing is a highly developed industry that uses large quantities of ceramic carriers for noble-metal and base-metal catalysts. The petrochemical industry likewise depends on ceramics. More efficient, smaller plants may result from the emerging new generation of high-surface-area structural ceramics.

Drilling ot wells for gas and oil in more difficult sites requires new tough, hard materials and provides a challenge to the development of bonded diamond aggregates. In offshore deep well sites, ceramic flotation materials, from glass and ceramic bubbles to large glass balls, are under investigation.

Considerable effort has been given to developing higher temperature energy-conversion machinery. Experimental high-temperature turbines using silicon nitride or silicon carbide blades and stators will allow improvement in the efficiency of electric power generation if successful

Table 4

Estimated Energy Impacts of Various Technological Options

Energy Program	*Potential Total Effect on U.S. Energy Use %*	*Time Period For Its Implementation At Level Shown*
Industrial Conservation Programs Phase I Good Mgt. Small Cost 15% Phase II Longer Range Higher Cost (Plant & Process)	4.5% 4.5%	1- 3 Years 3- 5
Electric Load Mgt. (10%)	2.0%	1-10
Public Conservation Prog.	3.0%	1-10
Transportation (30%)	6.0%	3-10
Improved Oil Recovery	4.0%	1-10
Oil from Difficult Sites	20.0%	5-?
Improved Reliability of Electric Plants	2.0%	3-10
Direct Coal Combustion—Medium to Large Boilers	10.0%	5-15
Coal to Low BTU Gas	4.0%	5-15
Increased Development of Nuclear Converters	10.0%	5-20
Coal to High BTU Gas	10.0-20.0%	10-25
Coal to Liquid Fuels	10.0-20.0%	10-25
Indirect Solar Energy Hydro Tidal Wind	5.0%	1-25
Advanced Geothermal	1.0- 5.0%	5-25
Direct Solar Heat	2.0- 5.0%	5-25
Bioconversion	5.0%	1-25
Nuclear Breeder	10.0-20.0%	20-30
Solar Electric	1.0- 5.0%	20-30
Nuclear Fusion	?	25-?

materials and designs are developed. This is discussed in more detail below.

Coal is the most abundant fossil fuel and will surely return as a major contributor to the total energy supply. Direct combustion of coal offers the most immediate application. Increasing use of ceramics in boiler insulation and high-temperature filtration (for example, removal of fly

ash and adsorption-reaction-fixation of various sulfur oxides or other noxious gases) are new opportunities for ceramic technology. Successful application of these new materials and processes will make the burning of large amounts of coal much more acceptable than it was in the early 20th century.

It may be desirable to convert coal into a more convenient form such as a gas, liquid, or packaged solid. Such conversion will also remove noxious impurities at the converting facility. Many coal gasification and liquefaction schemes are under development. For example, lumps of coal are first granulated and treated to make a feedstock. The granules enter a gasifier, where they react with various combinations of oxygen, air, and steam. The products are CO, hydrogen, various hydrocarbons, char, and many impurities. High temperatures and pressures and very corrosive atmospheres require ceramics for liners, throats, valves, conveyors, and other applications. Refractories high in alumina have so far been the most promising, but there are pressing needs for improved ceramic materials and products for many components of these systems. Examples include ceramic liners for high-temperature plumbing, heat recovery, and, as in the case of direct combustion of coal, filtration of particulates and removal of sulfur-containing gases. New technology for methanation will likely require high-surface-area ceramics and catalyst composites.

For the longer range, ceramics are under development as components of magnetohydrodynamic (MHD) direct coal-conversion systems. Included are lanthanum chromite electrodes, various heat-exchange refractories, insulation, and ceramic filtration devices.

Most of the world's nuclear reactors use ceramic fuel elements, generally uranium dioxide, in pellet or particulate form, encased in metal jackets. A high-temperature reactor design has used uranium carbide microspheres jacketed in refractory pyrocarbon and encased in graphite. These products have required very high technology and many years of development.

The substantial development effort toward building fission breeders involves uranium oxides and carbides. A thermal breeder design, first burning U^{235} and then U^{233}, will require still more of ceramic technology. U^{233} will be produced by a breeding reaction of thermal neutrons with Th^{232}. The fuel material can be pyrocarbon-coated microspheres which contains solid solutions of U-Th carbides or mixtures of microspheres of each carbide. The structural materials, substantially graphite, must have very low capture cross sections for neutrons, so as to maximize the breeding reactions. The working fluid, helium, must also have minimum neutron reactivity and chemical inertness at high temperatures. In order to achieve economic electric energy, the materials technologies for thermal breeders must be further developed, particularly to achieve long life and high fuel burn-up.

The fast breeder operates on transmutation of U^{238} to Pu^{239}, using higher energy (fast) neutrons. Enriched (U^{235}) uranium oxides or carbides encased in stainless steel tubes are the fuel materials for the central reactor core. The stainless steel may be the limiting material, since the intense neutron bombardment causes swelling and failure of the metal. Surrounding the core is a "blanket" of similar fertile element tubes, containing natural or depleted UO_2. This contains mostly U^{238}. The working fluid (and coolant) in the reactors currently under development is liquid sodium. Sodium becomes very radioactive and requires an isolating heat exchanger. Helium may also be used as a working fluid. The disadvantage of its operation at high pressures is offset by the inertness of the gas and its nonreactivity with neutrons, eliminating the need for the intermediate isolating heat exchanger. The primary requirements for the ceramic fuels are stability at high temperatures, resistance to long-term radiation damage, and inertness to catastrophic reactions with the coolant in the event of fuel system failure.

In all the fission systems, fuel reprocessing technology and safe radioactive waste storage remain key problems. In the breeder reactors, it is necessary through chemical processing to separate Pu^{239} or U^{238} from the remainder of the fuel element, particularly from the fission products. The separated fuel must then be refabricated into fuel elements under partially radioactive conditions.

Storage of radioactive wastes is a very controversial subject, since "perpetual" storage safety is impossible to guarantee. At least some of the

methods involve ceramic materials. They include containers, insulation, and the admixing of glasses and clays to fix the wastes in a relatively inert form. This is discussed in more detail in the section on waste management.

Other applications of ceramics in nuclear fission systems include refractories, thermal and electrical insulation, chemical plumbing, heat exchangers, electronic components, and various construction materials.

Fusion (nuclear reactions of deuterium and tritium or boron 11 and protons) presents both one of the greatest future energy potentials and the most difficult technical challenges. An enormous fuel potential exists in all water, which contains deuterium (1 part in 5000). While fusion machines will not be free of radioactive waste problems, they will reduce this problem one or two orders of magnitude over fission systems. Two general approaches are under development: (a) a field-confined plasma, usually in a toroidal vessel with direct electrical conversion, and (b) laser-induced microexplosions, which most likely will heat a working fluid for an electric turbine/generator system. Neither of these approaches has operated in a "break-even" experiment. The materials problems and the "solutions" discussed below are only conjectured.

In the Tokamak concept of a confined plasma machine, a plasma (10 to 100 KeV temperatures) is confined by superconducting magnet fields and maintained for burn times on the order of 1 h per run. Energy is extracted by direction conversion. The most serious materials problems will probably occur in the first wall in the reactor. Radiation damage (neutrons and hot ions—up to 100 KeV) will cause erosion, bubbles, blistering, growth, and loss of strength in most metals. Molten pressurized lithium, which may be used as a coolant and tritium breeder, will present corrosion problems. Ceramics proposed for the wall include SiC and Al_2O_3. Measured erosion rates for SiC are on the order of 1.5 atoms eroded per 100 KeV He ion (about half the rate for stainless steel). An alternative approach is use of sacrificial walls, such as SAP—an aluminum/alumina cermet. Solid lithium compounds have been considered for the breeder blanket. Reflector and neutron-shield materials under study include graphite and boron carbide, either solid or dispersed in metals. Since plasma confinement is by virtue of intense magnetic fields and its stability dependent on field uniformity, current flows, and the absence of spurious plasma-deforming effects, construction materials must operate reliably in the presence of these fields. In addition, they must resist the effects of erosion, corrosion, and radiation damage.

In the laser fusion concept, a multimegajoule, nanosecond-duration, multibeam laser is focused on a micropellet target containing deuterium or tritium (or their compounds), or combinations of these. Pellets of the order of 1-mm diameter may release up to 100 MJ of energy. If the reaction is deuterium/tritium, about 75% of this energy will be released as neutrons; pellet debris will account for about 15%, and α particles and soft X-rays will make up the remainder. These microexplosions are proposed to take place about 10 per second in a machine that will, in various schemes, heat a lithium blanket. The blanket may serve as a working fluid through a heat exchanger in addition to serving as a fertile material for breeding tritium.* Materials problems will be similar to those described above in reference to the confined plasma machines. Ceramics also may be used for the pellet structure and in the lasers. For example, early experiments have used amplified Nd-glass lasers. Although the fusion machines are still speculative, some materials problems are receiving early serious study. Some of the ceramic technologies and evaluations of properties revealed in the evolution of nuclear fission systems may be applicable.

Ceramic materials may also be used in other alternate energy systems. Corrosion is a major problem in geothermal systems. High stresses are found in the rotating equipment in wind generators; possibly this is an application for high-strength, high-modulus ceramic oxide, glass, or boron and graphite fiber composites. Bioconversion may incorporate the advanced ceramic substrates in various parts of the processes.

The incident radiation intercepted by the Earth is about 1.7×10^{17} W. Of all the alternate energy systems, solar energy offers some of the most

*One such scheme, proposed by A. P. Frass of ORNL, uses a liquid lithium vortex with the microexplosions occurring at the vortex bottom. Alternative power schemes use direct conversion.

attractive features and, along with fusion, bears substantial energy potential. Unfortunately, there are also major technical and economic problems. In temperate climates at the surface, the maximum available power (practical) is between 100 and 900 W/m². Solar energy is dispersed in the form of radiation and is intermittent, presenting three primary technical problems: (a) it must be collected, (b) it must be absorbed and converted, and (c) it must be accumulated. In accomplishing these objectives, considerable improvement is needed in efficiency and cost effectiveness.

The collection of solar energy using ceramics has been accomplished by lenses of glass, reflectors or mirrors, and flat-plate systems. (The purist may cite clay soils for agricultural solar use.) Problems include dirt; damage by wind, hail, sand, and rocks; cost; and undesired solar absorption. Limitations of absorbers have led to flat-plate collectors with glass covers, coated glass, and the like, which operate at relatively low temperatures (~100-200°C). As systems that can operate at sustained high temperatures are developed, technology will move to focusing of solar radiation from collectors, as in solar furnaces.

Absorber/converter technologies using ceramics have included direct conversion involving semiconductor photovoltaic materials such as silicon and cadmium sulfide, as well as various thermoelectric materials. In direct-conversion systems as well as indirect heat systems, ceramic absorber coating and antireflection layers have been used. Efficiencies of costly solar converters in aerospace vehicles have been greater than 10%. Solar electric terrestrial systems typically have had much lower efficiencies and, in some cases, problems of deterioration. The need exists for developing lower cost semiconductors in sheets or coatings. For indirect or heat-absorbing systems, the stability of coatings that absorb visible light and internally reflect infrared are limited by temperature and corrosion. Silicon and multilayer metal-metal oxide coatings are currently under investigation. Solar energy thus may be used to heat a working fluid in a tube or other container. Ceramic materials may be used in such tubes or as reflector jackets. Ceramic insulation will also play a part in these systems.

Underlying the promise of new ceramic technology for advanced energy systems is the need for considerable basic and pioneering materials research. Early interaction with systems engineers and designers will be necessary.

Transportation Systems

Ceramics in vehicles powered by internal combustion engines are not limited to sparkplugs and windows, but include about 15 distinct types of use in today's automobile. New applications are being developed [20]. Igniters made of silicon nitride or silicon carbide are being used in large aircraft jet engines; these igniters must endure thermal shock by ignition or water quenching and must endure for hundreds of hours under severe high-temperature, high-velocity corrosion conditions. Both silicon nitride and silicon carbide are being evaluated for automotive catalytic converters and thermal reactors; magnesium aluminum silicate ceramics are in use today. There is increasing use of ceramic magnets and flexible ceramic-plastic magnets in small electric motors in autos. On-board computers are expected to enhance performance by real-time optimization of operation conditions; these computers will be largely the products of ceramic technology.

Perhaps one of the most exciting future applications for ceramics in transportation is the ceramic turbine. Metal turbines are limited in efficiency by the maximum temperature their hot parts can withstand in an oxidizing temperature. Also the alloys involved cost up to $15 a pound. Ceramics such as silicon nitride or silicon carbide offer promise of operation at 2500°F (1370°C) and perhaps 3000°F (1650°C) might be obtainable. A study by NASA indicated that a 3000°F (1650°) turbine with a performance equal to a typical eight-cylinder engine could give around 51 mi per gallon of gasoline [35]. Such a turbine is a long way off, but good progress is being made on an automotive gas turbine with ceramic parts (and on a stationary turbine for power generation). Although an all-ceramic turbine is the ultimate goal, initial efforts have concentrated on using ceramic parts at the hottest places. The principal results to date, according to Katz [13], include the following for the vehicular engine:

- All stationary hot flow path components (inlet nose cone with integral transition duct, stators,

and shrouds all of reaction bonded silicon nitride) have demonstrated at least 100 h durability in engine testing to 1930°F (1054°C), using metal turbine wheels.
- A reaction-sintered silicon nitride combustor has been rig tested for 200 h over a representative duty cycle, including 35 h at 2500°F.
- Aerodynamically functional ceramic turbine wheels have been fabricated and cold spun with encouraging results.

Equally important is the ceramic heat exchanger, a vital component of an efficient gas turbine

Work is proceeding on the automotive and stationary ceramic gas turbines, and a second generation of development efforts aimed at producing turbines for military use is now in progress. A ceramic turbine for pilotless aircraft is being developed as the first stage of a possible man-rated engine. A ceramic turbine being developed in a Navy program appears closer to field operation. A 3-year contract beginning in the spring of 1976 calls for the construction of a modifiied gas turbine (with ceramic parts in the critical places) that is to develop 100 shaft horsepower. It will be tested in a high-speed patrol boat. Potential payoffs include improved performance, 50% reduction in the use of strategic, imported, superalloy materials, lower specific fuel consumption at full and partial power, improved resistance to corrosion and erosion, and ability to use lower grades of fuel [36].

Another exciting prospect is the use of ceramic bearings. Precision-quality roller bearings have been fabricated with silicon nitride as the rolling element and steel races. Entire bearings have also been fabricated from silicon nitride. Sixteen tests of the latter conducted at 600,000 psi maximum Hertzian stress were without failure; the longest running was over 93 million stress cycles. At 800,000 psi, the life was equal to that of M-50 CVM steel (the standard of high-speed bearings) at 700,000 psi [37]. These promising results have stimulated a Navy program on ceramic bearings for either room-temperature use (e.g., helicopter rotor pitch linkage) or high temperature use (e.g., jet engine bleed air valve bearing at 1100°F, or 593°C). Improved design is expected to extend life [38].

Although glass has long been used in automobiles, its mode of use is changing [39]. There has been a transition from clear glass to heat-absorbing glass. New auto glass includes such features as rear window defogging capability, built-in radio antenna, and improved windshield safety characteristics made possible by new interlayer materials and thinner glass that weighs up to 17% less. The critical need to reduce weight in automobiles and to fight corrosion is more and more often causing the replacement of metal stampings by fiberglass-reinforced parts. Some 307 million lb of fiberglass-reinforced plastic were used by the United States automobile industry in 1974. Also, fiberglass tire cord is competing for a larger share of the radial tire market.

Major fuel savings will result if lighter weight reinforced plastics and metals can be used in automobiles. Ceramics will likely be used for such reinforcement.

Environmental Systems

As the density of population and associated industry increases, there is a corresponding increase in pollutants or waste products that must be absorbed by the environment. Solid materials and nuclear wastes are discussed in another section under waste management. This section will mainly pertain to applications of ceramic technology, to removal of small particulates and gaseous pollutants and to water treatment.

Coal-burning powerplants generate large quantities of fly ash. Some of this is used in concrete aggregates. Electrostatic precipitators and bag filters have been used to remove this fine ash from the powerplant emissions. These systems require various amounts of cooling of the emissions prior to its entry into the cleaning system. The cooled gases then must be blown up the stack, which no longer serves as a draft generator but only as a means of piping the cleaned "smoke" high enough into the atmosphere to be carried away by winds. The development of high-temperature ceramic filters for fly ash will offer the possibility of saving energy otherwise wasted in this process.

Additionally, large amounts of oxides of sulfur and nitrogen result from combustion of coal and oil in boilers. There is no practical system for removing NO_x from coal-burning powerplants.

Various means for removal of SO_x are now being developed, some of which will use ceramic materials. High-surface-area ceramic substrates and structural shapes, such as honeycombs, fiber panels, and saddles, may be used to carry catalysts and adsorbents for removal of the noxious gases. Such removal may be advantageously combined with heat exchange at relatively high temperatures.

The treatment of pollution from internal combustion engines has already resulted in a major new application of ceramics [40]. Catalytic automotive devices use either honeycomb ceramics made of magnesium aluminum silicates overcoated with gamma alumina or ceramic pellets made of gamma alumina. Oxidation catalysts for control of carbon monoxide and unburned hydrocarbons are usually mixtures of platinum and palladium deposited on the gamma alumina surfaces.

The catalytic treatment of automotive NO_x emissions has not been implemented. This requires a reduction reaction using reducing gases present in the emissions (primarily CO). Methods under investigation include noble- and base-metal catalysts on ceramic honeycombs in devices combining oxidation and reduction—"three-way catalysts"—or sequential reaction chamber devices. It is possible that the knowledge gained in this work may be applied to the industrial NO_x problems discussed above.

The benefits of cleaner air are often cited in terms of a better quality of life and health for the population, but these are general and subjective terms. Technical evaluations of air quality have been made in certain regions of the United States, such as Los Angeles County, where cost-benefit ratios are expressed as dollars per daily ton of pollutant. The studies indicate a nonlinear relationship, in which the cost rises sharply after an initial reduction is made. In general terms, the reduction of pollutants to half their present level can be accomplished relatively easily, but another twofold reduction may increase costs as much as tenfold. This particular relationship is probably not appropriate to other regions, but the point is that standards for any region must be carefully considered, so that maximum pollutant limits are set no stricter than health and safety require, since costs will rise sharply above a critical emission control level.

Another cost-benefit factor involves minimizing the energy cost of meeting the desired clean air standards. With the automotive catalytic converter, many of the mechanical modifications were eliminated or changed, so that 1975 automobile fuel consumption for a given size vehicle was reduced 10% to 20%. Offsetting this are the energy costs of modifying refineries and making the catalytic devices.

It is important that continuing cost-benefit analyses be made, particularly for industrial pollution systems where very large investments must be made.

Water treatment is likely to develop rapidly with the development of alternate energy systems that consume or at least use large amounts of water for cooling. Additionally, dispersed small water treatment devices may evolve as local pollution situations demand. Zeolites, clays, and other natural inorganic materials have been used in such treatment. The availability of high-surface structural ceramic shapes made of materials useful in ion change, filtration, and adsorption offers opportunity for improved designs of water treatment systems.

Ozone has been used in Europe for killing organisms in drinking water as well as for deodorizing purposes. Ozonators are usually made with glass tubes, and more recently with dielectric ceramic plates, (e.g., barium titanate). If chlorination of drinking water is replaced as some environmentalists suggest, there will be need for advanced ceramic technology in ozone generation.

Substantial efforts have already been made to reduce industrial pollution. Many of these efforts have been cost- and energy-effective. It is particularly effective to have processes that prevent or minimize pollution in the first place. Here, too, many opportunities exist for use of ceramic materials. It is likely that combinations of pollution prevention processes, after-treatment devices, heat recovery systems, and use of waste products will lead to industrial plants having the least adverse effect on the environment.

Communication Systems

Traditional communications systems (telephone, radio, and television) make extensive use

of ceramic materials ranging from porcelain enamel insulators to individual transistors to chips containing thousands of individual transistors combined in circuits for switching, amplification, modulation, digitizing, etc. Progress will undoubtedly continue in these areas of communication technology, but even more spectacular developments involving ceramics appear likely in the new field of light-wave communication [41]. The invention of the laser was quickly followed by the realization that it offered the possibility of very high information-carrying capacity. An ordinary telephone line requires about 5.6×10^4 bits/s for a voice conversation. Microwave systems today typically carry 1 megabit/s, or 17 simultaneous voice channels, appears possible with pulse modulation of a laser beam using a fiber optical wave guide [43]. Achieving a practical systems depends on materials, largely ceramics. The essential elements of such a system are a transmitter, a transmission medium, and a receiver.

The open atmosphere is a very unreliable medium for the transmission of light. The requirements of a maximum attenuation of 20 dB/km for a practical transmission medium combined with attenuations of several thousand dB/km for the best optical glass in 1966 posed a challenge. Meeting that challenge is one of the great success stories of ceramic processing [44]. The degree of challenge will be better understood when it is realized that not only must the loss be extremely low but the dispersion should also be low and the index of refraction must be graded toward the surface to contain the light wave and prevent loss by partial escape through the surface. Extremely low loss fibers were first produced by using silicon tetrachloride reacting with oxygen to form a fine, loosely bonded powder that produces a porous polycrystalline cylinder, which can then be densified and pulled into a fiber. Later silicon tetrachloride and volatile chlorides or fluorides of boron and germanium were used to cause chemical vapor deposition of very pure glasses on the inside surface of a silica tube, which has collapsed and then drawn down into a fiber. These processes permit steps or a continuous gradient in index of refraction. It appears that glass fibers with the required optical properties can be routinely produced. These fibers are subject to loss of strength with time due to mechanical and atmospheric attack of their surfaces. (This was discussed above as a property common in some degree to most ceramics.) Application of the basic knowledge of fracture in brittle materials has suggested practical engineering solutions to this problem, including polymeric coatings as protective sleeves.

Both light-emitting diodes and heterojunction lasers have been developed as light sources for optical communication. Lasers are preferable for single-mode operation, which permits very high communication rates. Present heterojunction lasers are based on the ternary alloy system $Al_xCa_{1-x}As$. The wavelength can be varied over the range 0.8-0.9 mm by varying composition; this is a useful range, although the lowest attentuation and dispersion in the glass fibers occurs in the range of 1.0-1.1 mm. Attempts to produce a heterojunction laser in this range have led to experiments in quaternary alloys such as $In\,Ga_{1-x}As_yP_{1-y}$ and $Ga_xLa_{1-x}As_ySb1\text{-}y$ [45]. Here is a real challenge to both molecular engineering and ceramic processing.

Perhaps an even greater challenge to molecular engineering and ceramic processing is presented by the field of integrated optics. A typical optical telephone repeater includes a laser, a modulator, a detector, a waveguide, prisms or lenses, etc. [46]. There are evidently great advantages to miniaturizing the components and packaging them as a single unit in a similar fashion to microelectronic circuits. This is a highly active (and proprietary) field of research. There seems no doubt that a technology with enormous implications for military as well as civilian communication will result.

Metallurgical and Other Processing Systems

Refractories for the metallurgical industry underwent considerable change over the past half century, as more was understood about the complex reactions between the various liquid, solid, and gas phases in the furnaces. Already cited above, for example, is the basic oxygen furnace, which has become the primary steelmaking device. With loadings greater than 100 tons (~ 100,000 kg), oxygen is blown into the molten metal, burning out carbon very rapidly with charge-discharge cycles of about 2 h. The extremely corrosive and

turbulent conditions required the development of pitch-bonded basic magnesite brick to make the operation practical.

With an expanding economy, there will be a growth in the metal industry, and the increased demand for metals will require greater reliance on lower grade and more difficult-to-obtain ores as the present high-grade ores are depleted. Combined with more stringent requirements for air pollution, water pollution, and safety will be the need to minimize energy consumption. These challenges to the metals industry will require much of ceramic technology.

Certainly in terms of economics, basic oxygen steelmaking is the most revolutionary and important new process in metallurgy [47-49]. Unlike older processes, in which air was used to burn carbon, silicon, and other impurities from molten iron, in this process pure oxygen is used, making it possible to produce very high quality steel more quickly and with less manpower and capital investment. This technology is now being extended to nonferrous metallurgy, which will again require considerable research and development not only in the process and equipment but for the ceramic refractories as well. There is increasing use of fluidized beds for various processes, and included in these are the roasting of sulfides, various forms of calcination, heat exchange, chlorination, and eventually reduction of metals.

Still another emerging technology, in both the metals and glass industries, involves vertical shaft furnaces. In the metals area, such a furnace uses counter-current flow of metal and gases to give high mass and heat transfer rates with correspondingly large throughput. The liquid-gas reaction confers considerable advantage in control of the process.

Many of these new processes will use electric energy in place of fossil fuels. These processes lead to less pollution and less waste of raw materials at the manufacturing site, which may be partly offset by increased pollution at the electric generating plant but may yield a net decrease in pollution. Currently, there is considerable activity in continuous steelmaking. Many of the major processes have been or are being developed outside the United States. In one such process, a stream of impure liquid iron falls vertically into a region containing powdered flux consisting of iron oxide and lime. Multiple jets of oxygen under high pressure are directed on this stream, breaking it into a fine spray of molten metal. The oxygen oxidizes the impurities in the liquid iron, and these eventually become a slag. The refined metal and slag are caught in a ceramic refractory vessel. At this point, some scrap may be added to the product.

In still another process, combined melting and reduction coupled with heat exchange between reducing gases and incoming raw materials provide an interesting new approach. In the most severe corrosion area, the vessel lining is made up of chilled slag or metal, because the container is a water-cooled steel structure. Continuous systems must be reliable, and the refractory problems are very real indeed. They present an enormous opportunity to the ceramic industry.

There have been many improvements in the original Hall process for making aluminum, but there have been no basic changes since the process was invented more than 100 years ago. Some new processes are under development. In one of these, recently announced, the aluminum ore is treated with chlorine to make aluminum chloride in the initial part of the process. If new forms of fused salt electrolysis are necessary for these processes, they will require new or improved refractories. Such processes and improved materials are also likely to be used for other reactive metals, such as magnesium, titanium, and sodium.

In many of the metallurgical processing systems, there is a need to prevent the emission of very fine particulates that enter the atmosphere as "smoke." Filtration of these "smokes" at high temperature, perhaps combined with heat exchange, represents a need that the newer ceramic technologies may be able to satisfy. Still another problem in metallurgical processing is the conservation and reuse of scrap metal. It is often necessary to separate one metal from the other, for example, copper from steel, so that the scrap may be useful. The recovery and separation processes may eventually be chemical in nature. On the other hand, there may be applications of high-temperature processes, which again will require special ceramics.

Refractories exist in preshaped forms, such as bricks or blocks, nozzles, gates, troughs, tiles,

and the like, or in unformed mixtures which may be gunned, cast, molded, tamped, rammed, or injected in place. Almost all refractories are used in combinations of shapes and compositions for metallurgical processing equipment. For example, the roof of a furnace may be quite different in composition from the hearth. Special compositions may be used for high erosion components such as gates, sleeves, and nozzles. It is important to note that many new fabrication processes, developed in recent times, should extend the forms of refractory materials available to the metallurgical processing industry. Ceramic fabrics, loose fibers, blankets, and matts are examples of such new materials. Additionally, composites such as cermets, laminates, and the like are useful in specialized parts such as thermocouple shields for control systems, injection nozzles, and other parts where more precision-made ceramic materials of a very special nature are required. Ceramics likewise are being used in the foundry industry in riser sleeves and other specialized parts of the molds.

Structural and Composite Systems

We have already mentioned the development of monolithic structural ceramic materials for gas turbines and bearings. Here we shall concentrate on composite structural materials that include a ceramic phase. Such materials consist of a matrix (polymer, metal, or ceramic) containing ceramic fibers.

The fibers themselves may be vitreous, single crystals, or polycrystalline aggregates. Glass fibers are, of course, already extensively used to reinforce plastic and tires as previously mentioned. Another promising use is to reinforce concrete made with portland cement. Glass fibers up to this time are not believed to have been competitive with steel for prestressing or conventional reinforcement, but hold promise for used in the form of short, chopped fibers to produce an inexpensive cement-based material having high structural and impact strength [50]. Unfortunately, the usual fibers of E, A, or Pyrex glass undergo a reduction of strength caused by chemical attack by alkalis in the cement. The attack was reduced (perhaps eliminated) by using fibers made from glasses in the system Na_2O-SiO_2-ZrO_2 or the system CaO-Al_2O_3-SiO_2MgO. Adding 20 to 30 % of pozzolanic fly ash to portland cement greatly minimizes the alkali attack on E-glass, according to R. E. Harmon (private communication).

Metals can also be reinforced, typically with continuous fibers; large increases in longitudinal strength are possible, although transverse strength remains relatively low. An approach giving a 30% increase in transverse strength of fiber-reinforced aluminum has recently been reported [51]. Continuous fibers 5.7 mil (0.15 mm) in diameter made of B coated with SiC to reduce reactivity were used as the longitudinal reinforcement. Mats of β-SiC whiskers with diameters of 1 to 3 μm were used to obtain reinforcement in one perpendicular dimension. Such two- or three-dimensional reinforcement is technically promising when the criticality of performance justifies the cost.

Another very promising ceramic fiber and some of the associated problems can be illustrated by reference to coated carbon fibers. Strengths above 500 x 10^3 psi and high elastic moduli can be achieved in carbon fibers, and multifilament yarns can be produced at reasonable cost. Although some success has been achieved with composites of carbon yarn and aluminum, there have been difficulties in impregnating the yarn with metal matrixes, and the small individual fibers are caused to deteriorate by reactions with the matrix materials at high temperatures. The properties of carbon fibers are so outstanding that real effort to overcome these problems seems justified. One promising approach involves producing a boron-coated carbon fiber by chemical vapor deposition [52].

Another interesting development involving carbon fibers is the production of SiC fiber-reinforced Si. Molten silicon is infiltrated into alined carbon fibers. A reaction occurs, forming SiC fibers with the same alinment as the original fibers. The resulting materials show no loss of strength with heating until 1300 °C is exceeded [53].

Very recently, ceramic oxide fibers have become available in short-staple and continuous form. The latter can be woven into yarns and fabrics. In addition to their applications per se as very high temperature insulation (up to

1600°C)—blankets, sleeves, belts, flame shields, and various other forms—these fibers hold good promise for reinforcement of plastics and metals. Their strengths are in the range of 200 000 to 600 000 psi, and moduli of elasticity range from 10 million to 40 million psi. Compositions include pure oxides, (for example, alumina and zirconia) and many refractory materials, such as alumina-boria-silica and mullite-like bodies. These materials appear to have mechanical properties intermediate between those of the glasses and boron and graphite fibers; however, they may have very good composite values and related cost effectiveness. Still other advantages of some of these fibers include transparency and the ability to be intrinsically colored (with potential for use in decorative applications where resistance to fire and smoke is required).

Waste Management Systems

Ceramics have three potential categories of use with regard to waste management: as components of the waste management process, as useful products made from portions of the waste, or as part of the ultimate disposal system for dangerous waste. Attention is focused on the latter two in this section.

As the need to conserve our natural mineral resources becomes more acute we must consider using the large volume of ceramic wastes being generated by many of the mining, processing, and industrial operations as well as the glass recovered from urban refuse. In recent symposia on the use of mineral wastes [54-57], many applications for using these nonmetallic materials have been cited. Table 5 gives examples of industrial byproducts, produced in large volumes, that might ultimately find much greater use as raw materials for ceramics. Some are already in use. For example, 35 million tons of iron and steel slags were sold in the United States, mainly for use as construction aggregates [58]. As another example, 5 million tons (16% of total production) of coal fly ash were used in portland cement, as a filler in bituminous or asphaltic pavements, and for such other uses as construction fill and soil stabilization [59]. The potential for increased use of fly ash in cements seems good because of such beneficial effects as rapid curing, low heat generation on curing, resistance to saltwater, and extended shelf life. Such uses are facilitated by the relatively wide range of compositions acceptable in cements and pavements. Fired ceramic materials generally require more narrow composition variations, ranging down to almost zero for high technology uses. However, it may be possible to develop high-technology products based on mineral wastes of relatively consistent composition. For example, preliminary results indicate a good possibility of using millscale from steel plants as raw material for permanent ceramic magnets for small motors [60].

Research by the Bureau of Mines [61] has shown that the energy required to produce clay brick can be significantly reduced with only small additions of urban refuse glass. For example, using 10% glass reduced the maturing temperature by 10%, a significant amount.

Other wastes such as bauxite tailing have been foamed and sintered to produce lightweight ceramic insulating panels [62]. Even agricultural wastes like rice hulls can be converted to silicon-base ceramics, and claylike materials can be recovered from papermill effluents. A current project is demonstrating the conversion of oil shale residues, after removal of the oil, to glass and glass-ceramic products having both high strength and aesthetic appeal [63]. Such industrial wastes as furnace slags have long been used for railroad beds and for conversion into rock-wool insulation.

The most important example of the potential of ceramics for use in disposal of dangerous wastes is in the disposal of radioactive waste, but as other forms of long-term health hazard are recognized other possibilities are likely to develop. A byproduct of the operation of nuclear reactors is highly radioactive fuel elements, which are dissolved to make a liquid consisting of fission products and other wastes left after most of the U and Pu have been removed for reuse. Some of the isotopes could endanger human life for tens of thousands of years. Proposals for dealing with this problem usually involve conversion to a solid and storage either in a retrievable surface storage facility or in a geological formation. Since protection from weathering action over many thousands of years cannot be absolutely guaranteed, the

Table 5

Some Examples of Mineral Wastes with Potential as Raw Materials for Ceramics [36–39]

Source	*Suggested Uses and Remarks*
Aluminum Processing (Red Muds Containing Al, Fe, Si, Ca, Na, Ti, P. H. O)	Portland Cement, Pigments, Slag Wool, Binders
Copper Processing (Si, Fe, Mg, Al, Ca, O, etc.)	Building Materials
Taconite Iron Ore Processing (Si, Al, Fe, Mg, K, etc.)	Building Materials
Phosphate Rock Process (Phosphate Slimes Containing Si, Ca, P, F, C, FC, Al, Mg, Na, etc.)	Aggregate, Building Materials (Serious Dewatering Problem)
Iron and Steel Slags	Cement, Pavement Filler, Mineral Wool
Fly Ash	Portland Cement, Pavement Filler, Aggregate, Brick
Papermill Waste (Si, Al, Ca, O, etc.)	Clay Substitute, Carbon Fibers
Recycled Glass	Asphalt Filler, Glass Beads, Fiberglass, Tiles, Bricks, Aggregate
Furnace Dusts	Aggregates, pigments, soil conditioner
Anthracite Refuse	Paving Material
Cement Kiln Dusts	Fertilizer

solid should be as stable and impervious to environmental chemical attack as possible [64].

Much work has been done on incorporating radioactive wastes into glass, primarily borosilicate type. This can be done, but glass is not thermodynamically stable. The combination of high temperature due to self heating (300 to 900°C) and long duration of storage may lead to partial crystallization and cracking. The resultant permeable mixture of glass and crystalline phases may be more leachable by ground water over long times than monolithic glass. An alternate approach is to develop crystalline ceramic compositions that can incorporate the radioactive wastes; this approach is being pursued [64, 65]. Another approach is incorporating the wastes in a glass that can then be deliberately devitrified to form an uncracked, fine-grained glass-ceramic and to design the composition so that the latter will have very low leachability.[66].

Electrical and Electronic Systems and Information Systems

Ceramics have played a major role in electrical systems, mainly as insulators. As electrical transmission systems became larger and more

sophisticated, and additional safety requirements were added, the ceramic materials used became correspondingly more complex in composition and design. Ceramics have been able to provide the outdoor weatherability and stability required of high-voltage transmission lines, including resistance to various manmade environmental problems.

In less sophisticated and lower voltage applications, ceramics have been the standard insulating material in many appliances, particularly where high temperatures develop—heaters, irons, toasters, and the like.

One of the highest volume applications of ceramic insulators has been in automotive spark plugs. Early plugs used special porcelains, but these later were replaced by high-alumina compositions. In some aircraft spark plugs, beryllium oxide has been used. Several modern ceramic fabrication processes were developed as a result of the need for advanced spark plug resistor technology (e.g., spray drying of alumina bodies; use of fast, high-fire tunnel kilns; and, perhaps most important, the automatic isostatic molding process).

Metallized ceramics have been developed for use with hermetically sealed electrical components, transformers, capacitors, relays, controls, and motors. Many of these ceramic materials must also have unusual mechanical properties and resistance to thermal shock for use in high-stress electrical components.

Ceramic chips, carriers, and packages have become the foundation of the modern electronics industry, with applications in computers, calculators, and other information-processing systems. Ceramics have been used because of their high strength in very thin sections, their excellent electrical characteristics, and, in many cases, their good thermal conductivity and stability over a wide range of temperatures. For many years, aluminas similar in composition to that used in spark plug bodies were used. Typical materials could be made with a 25-μin. (0.625 μm) finish on their surfaces. More recently, improved alumina technology with materials that are very pure have produced as-fired ceramics with a surface finish better than 1 μin. (0.025μm). Such materials are useful in advanced thin-film circuitry devices. These advanced electronic ceramics can be made in complex packages with many layers intimately bonded together, including layers of complex metal conductors and interactive terminations on edges or in holes through the ceramics, with provisions to make totally encapsulated ceramic packages of extremely complex design.

Ceramic materials, however, are applicable in many other ways than as insulators. (We include semiconductors as ceramic materials.) Their processing represents an extremely important recent technological development. In addition to their widespread use in transistors, semiconductive materials have become useful to the information-processing area, including the copying industry. These materials include selenium, arsenic selenide, cadmium sulfide, zinc oxide, and many others. Such materials also provide the basis for TV cameras, and still another class of inorganic ceramic materials are used as phosphors in the viewing tubes or monitors.

Magnetic materials, notably iron and iron cobalt oxides, have become the primary media in which information is recorded and stored. The lowest cost storage material is a composite made of plastic tape coated with fine magnetic oxides dispersed in a binder. More recent magnetic oxide storage systems use the same or similar materials coated on discs which spin at high speed and are accessed by sophisticated tracking heads. Ceramic ferrites made in the form of tiny doughnuts are the basis of the main core memories used in computer systems. Many of the new information systems also involve ceramic materials, such as magnetic bubble devices, charge-coupled devices, and materials that change resistivity by very local electric or thermal activation. Ceramic materials provide the basis for many electro-optic devices, and there may be considerable future development in electroluminescent materials, light-emitting diodes, and various sophisticated lasers. Advanced information systems combine the electrical, magnetic, and optical properties of these ceramic materials. While unusual advances have been made in the past decade with such materials, there remains a very large opportunity for further developing and exploiting ceramic technology for information systems.

Medical Systems

Ceramics can play at least two important roles in medical systems. In both cases their potential has been realized, some success has been demonstrated, and very significant future applications seem assured. These roles are as implants for bone and tooth replacement and as substrates for chemical reactions occurring on their surfaces; these reactions are highly specific and pertinent to diagnosis and/or therapy.

Work on ceramic implants, including animal experiments, was pioneered in the United States. The driving force was the recognition that the customary metal and metal-polymer implants, especially for joints, had a limited life. In fact, the limited durability of hip joint implants (about 15 years) generally restricts their use to patients over 60 years of age. Ceramics appear promising because of their similarity to natural bone, their compatibility with body fluids, and their low friction. Success with ceramics will require drawing upon basic work on processing to produce a fine-grained, high-strength product with high reliability, upon basic knowledge of fracture for design, inspection, and lifetime assurance in service, and upon knowledge of heterogeneous chemistry to optimize bone growth and attachment. Work on practical implants has moved ahead in France and even more in Germany, where several manufacturers now offer hip joint prostheses for sale to surgeons. Initial results with several hundred hip joint replacements in humans have been good. The potential application is large; some 1500 hip joint prostheses are implanted daily on a worldwide basis [67].

The use of ceramics as substrates for medically important heterogeneous reactions depends on basic advances in processing (production of high-surface-area material) and surface chemistry (attachment of enzymes). An important step was the development of understanding and control of phase separation in alkali borosilicate glasses [68]. Subsequent leaching produces a porous glass with a narrow and closely controllable distribution of pore sizes. These permit tailoring the pore size to the particular enzyme. Such controlled-pore glass has been successful; for example, immobilized glucoamylase has been used for the conversion of starch to glucose [69]. Three basic limitations of controlled-pore glass as a carrier for continuous reactor technology are: high material cost, poor durability in alkaline environments, and negative charges on the glass surface. A competing family of controlled-pore ceramics has been developed with surface areas ranging from 10 to several hundred m^2/g and average pore diameters ranging from 8 to 86 μm. These are available in SiO_2, Al_2O_3, $TiO\text{-}Al_2O_3$, and TiO_2.

Enzymes can be covalently attached to the inorganic substrate by means such as using the organic functional group of the silanized carrier and an organic group of the enzymes [70]. This is the basis of a promising technology for industrial processing, analytical use, and therapy. An example of the latter is the use of immobilized enzymes in a shunt on a human volunteer undergoing kidney dialysis.

CONCLUSIONS

The wide range of special chemical and electromagnetic properties of ceramics has led to their use in an enormous number of special applications. Many further developments in these areas are technically possible and can fill practical needs. Progress in controlling fracture and improvements in processing to produce higher strength ceramics have opened the way to much greater structural use of ceramics, both in composites and as monolithic parts.

There is a strong relationship between better understanding of fundamental properties of ceramics and their applications. Some research is useful at both ends of the spectrum, i.e., applied research designed to directly support a particular application or basic research designed simply to pursue a particular phenomena, without regard to any possible application. The former may be too specific, and be stopped too soon, to provide support for long-range technical developments. The latter may tend to produce an elaborate understanding of properties in simple materials and omit development of understanding of the basic properties in complex, technically important materials. There is, therefore, a need for a third type of research that combines some of the features of the other two. This is focused fundamental research—fundamental in the sense that it seeks to develop an understanding of behavior

and is carried to a reasonable degree of completion, at least in stages; focused in the sense that the behavior to be understood is chosen for pertinence to practical need. We have tried to illustrate this concept with our matrix framework of Figure 1 and the subsequent discussions. The concept seems to fit the field of ceramics very well. The areas of fundamental research identified as central to ceramic science are each usually pertinent to several, and sometimes to many, applications. Selections of more specifically defined themes for focused fundamental research can be developed in this context.

In its second 30 years, the Office of Naval Research could play a very important role in the correlated development of ceramics for practical needs and the associated understanding of their fundamental behavior.

REFERENCES

1. D. A. Brobst and W. P. Pratt, eds., "United States Mineral Resources," U.S. Geological Survey, Prof. Pap. 820, 1973. (See especially Introduction, pp. 1 and 7.)
2. J. Boyd, "Ceramics—Man's Assurance of Abundant Materials," *Ceram. Bull.* **53**, 655 (1974).
3. E. T. Hayes, "Energy Implications of Materials Processing," *Science*, **191**, 661 (1976).
4. J. B. Wachtman, Jr., and M. A. Schwartz, "Ceramics from Plentiful Materials as Alternates for Scarce Materials," paper presented at Atlantic City meeting of the American Institute of Chemical Engineers, Aug. 1976.
5. T. Lee and C. Yao, "Abundance of Chemical Elements in the Earth's Crust and Its Major Tectonic Units," *Int. Geol. Rev.*, **12**, 778–786 (1970).
6. H. Solwang and M. Francis, *Ceramics: Physical and Chemical Fundamentals*, Butterworths, London, 1961.
7. F. Singer and S. S. Singer, *Industrial Ceramics*, Chemical Publishing Company, New York, 1963.
8. R. L. Erickson, "Crustal Abundance of Elements and Mineral Reserves and Resources," in "United States Mineral Resources," U.S. Geological Survey, Prof. Pap. 820, pp. 21–25, 1973.
9. D. K. Samples, "Energy in the Automobile," paper presented at the Energy Seminar conducted under the auspices of the Institute of Science and Technology, University of Michigan, Traverse City, Mich., Aug. 23, 1974.
10. W. D. Kingery, "The Nature of Ceramic Materials: Needs and Opportunities for Ceramic Science and Technology," paper presented at the American Chemical Society Symposium on "Ceramics in the Service of Man," Wash. D.C., Juen 8-10, 1976.
11. L. C. Ianniello, W. D. Kingery, and D. W. Readey, eds., "Critical Needs and Opportunities in Fundamental Ceramics Research," Summary of a meeting held at the Massachusetts Institute of Technology, January, 1975. U.S. Energy Research and Development Administration, Publ. ERDA-9, Apr. 1975.
12. R. J. Stokes, "Mechanical Effects in Optical Ceramics," Sosman Memorial Lecture, American Ceramic Society, Cincinnati, Ohio, May 4, 1976.
13. R. N. Katz, "Recent Developments in High Performance Ceramics," paper presented at the Conference on the "The Physics of Materials Technology," Feb. 4, 1976.
14. "Materials and Man's Needs," Summary Report and Supplementary Report of the Committee on the Survey of Materials Science and Engineering, National Academy of Sciences, 1974.
15. R. Roy, "Rational Molecular Engineering of Ceramic Materials, Retrospect and Prospect," Sosman Memorial Lecture, American Ceramic Society, Washington, D.C., May 5, 1975.
16. R. A. Laudise and K. Nassau, "Electronic Materials of the Future: Predicting the Unpredictable," *Technol. Rev.* 77 (Oct./Nov. 1974).
17. R. A. Laudise, "Future Needs and Opportunities in Crystal Growth—Crystal Growth Toward the Year 2000," *J. Cryst. Growth* **24/25**, 32–42 (1974).
18. W. D. Kingery, "Plausible Concepts Necessary and Sufficient for Interpretation of Ceramic Grain-Boundary Phenomena: I. Grain-Boundary Characteristics, Structure, and Electrostatic Potential," J. Am. Ceram. Soc., *57*, 1, 1974. "II. Solute Segregation on Grain-Boundary Diffusion, and General Discussion," *J. Amer. Ceram. Soc.* **57**, *74 (1974)*.

19. J. B. Wachtman, Jr., "Highlights of Progress in the Science of Fracture of Ceramics and Glass," *J. Amer. Ceram. Soc.* **57**, 509 (1974).
20. "Structural Ceramics," Report of the Committee on Structural Ceramics, National Materials Advisory Board, Publ. NMAB-320, National Academy of Sciences, Washington, D.C., 1975.
21. J. R. Johnson, "An Engineer's Perspective of Our Energy Dilemma," *Amer. Ceram. Soc. Bull.* **55** (Feb. 1976).
22. *U.S. Energy Outlook,* National Petroleum Council, Washington, D.C., 1972, 1973.
23. *U.S. Energy Prospects, National Academy of Engineering, Washington, D.C., 1974.*
24. J. J. McKetta, "Energy Crisis, Today & Tomorrow," *Chem. Engr. Prog.*, **68** *(1972).*
25. "Materials and Man's Needs," COSMAT Report, National Academy of Sciences, Washington, D.C., 1974.
26. F. C. Schora, Jr., "Clean Fuels from Coal," Institute of Gas Technology Symposium, 1973.
27. D. B. Meadowcraft *et al.*, "Hot Ceramic Electrodes for Open Cycle MHD Power Generation," *Energy Conversion* **12**, 145–147 (1972).
28. *An Evaluation of Advanced Converter Reactors,* U.S. Atomic Energy Commission, WASH 1087, 1969.
29. G. R. Hopkins, editor, Summary, Topical Meeting on Controlled Nuclear Fusion, San Diego, Apr. 1974, San Diego Section of the Technical Group for Controlled Nuclear Fusion and Power Division, American Nuclear Society and U.S. Atomic Energy Commission.
30. K. Boyer, "Power from Laser Fusion," *Astron. Aeron.* **11**, 44–49 (1973).
31. A. P. Fraas, "The Blascon—An Exploding Pellet Fusion Reactor," ORNL TM-3231, 1971.
32. J. L Emmet *et al.*, "Fusion Power by Laser Implosion," *Sci. Amer.*, (June 1974).
33. H. J. Davis, "Materials Considerations for High Energy Density Batteries," Peport given Canadian Ceramic Society, Feb. 1974.
34. D. W. Rabenhorst, "Potential Applications for the Super Flywheel," Reprinted from *1971 Intersociety Energy Conversion Engineering Conference Proceedings*, p. 38, Aug. 1971.
35. T. Alexander, "Hot Prospects for the New Ceramics," *Fortune*, p. 153, (Apr. 1976).
36. *Southern California Industrial News,* Apr. 2, 1976.
37. R. A. Alliegro, "The 'New Breed' of Ceramics," *Ceram. Ind.* (Mar. 1975).
38. J. W. Van Wyk, "Ceramic Airframe Bearings," Final Report on Contract N00019-75-0170, Boeing Aerospace Company, Feb. 1, 1976.
39. R. F. Sperring, Vice President, Supply, PPG Industries, Remarks to The Automotive Engineering Congress and Exposition, The Society of Automotive Engineers, Cobo Hall, Detroit, Mich., Feb. 27, 1975.
40. J. R. Johnson, "Auto Exhaust Control," *Encyclopedia of Chemical Processing and Design,* 1976.
41. S. J. Buchsbaum, "Lightware Communications—An Overview," *Phys. Today* **29** (May 1976).
42. W. J. French, "Materials for Fiber Optical Communications," to appear in *Educational Modules for Materials Science and Engineering,* Pennsylvania State University.
43. T. Li, "Optical Transmission Research Moves Ahead," *Bell Lab. Rec.*, p. 333, September 1975.
44. A. G. Chynoweth, "The Fiber Lightguide," *Phys. Today* **29**, 28 (May 1976).
45. H. Kressel, I. Ladany, M. Ettenberg, and H. Lockwood, "Light Sources," *Phys. Today* **29**, 38 (May 1976).
46. Esther M. Conwell, "Integrated Optics," *Phys. Today* **29**, 48 (May 1976).
47. *Extractive Metallurgy,* National Academy of Sciences, Washington, D.C., 1969.
48. *Refractories, Uses and Industrial Importance,* The Refractories Institute, Pittsburgh, Pa., 1975.
49. F. H. Norton, *Refractories,* 4th ed., McGraw-Hill, New York, 1968.
50. D. R. Lankard, "Fiber Reinforced Cement-based Composites," *Ceram. Bull.* **54**, 272 (1975).
51. F. E. Swindells and Paul J. Lare, "Improved Transverse Strength of Continuous-Filament-Reinforced 6061 Aluminum Alloy," *Ceram. Bull.* **54**, 1075 (1975).
52. R. D. Veltri, B. A. Jacob, and F. S. Galasso, "Large Diameter Carbon-Boron Fiber," *Ceram. Bull.* **54**, 1077 (1975).
53. W. B. Hillig, et al., "Silicon/Silicon Carbide Composites," *Ceram. Bull.* **54** 1054 (1975).
54. M. A. Schwartz, Chairman, *Proceedings of the First Symposium on Mineral Waste Utilization,* IIT Research Inst., Chicago, Ill., 1968.
55. M. A. Schwartz, Chairman, Proceedings of the Second Mineral Waste Utilization Symposium, IIT Research Inst., Chicago, Ill., 1970.
56. M. A. Schwartz, Chairman *Proceedings of the Third Mineral Waste Utilization Symposium,* IIT Research Inst., Chicago, Ill., 1972.
57. E. Aleshin, Chairman, *Proceedings of the Fourth Mineral Waste Utilization Symposium,* IIT Research Inst., Chicago, Ill., 1974.
58. M. A. Schwartz, Chairman, *Proceedings of the Second Mineral Waste Utilization Symposium,* p. 17, IIT Research Inst., Chicago, Ill., 1970.

59. M. A. Schwartz, Chairman, *Proceedings of the First Symposium on Mineral Waste Utilization*, p. 25, IIT Research Inst., Chicago, Ill., 1968.
60. M. A. Schwartz, Chairman, *Proceedings of the Second Mineral Waste Utilization Symposium*, p. 150, ITT Research Inst., Chicago, Ill., 1970.
61. M. E. Tyrrell and A. H. Goode, "Waste Glass as a Flux for Brick Clay," U.S. Bureau of Mines, Washington, D.C., RI 7701, 1972.
62. H. H. Nakamura, S. A. Bortz, and M. A. Schwartz, "Use of Bauxite Wastes for Lightweight Building Products," *Amer. Ceram. Soc. Bull*, **50**, 248 (1971).
63. B. S. Dunn and J. D. Mackenzie, "Preparation and Properties of Glasses Made from Shale Wastes," paper presented at 78th Annual Meeting, American Ceramic Society, Cincinnati, Ohio, May 1-6, 1976.
64. G. J. McCarthy and M. T. Davidson, "Ceramic Nuclear Waste Forms: I. Crystal Chemistry and Phase Formation," *Ceram. Bull.* **54**, 782 (1975).
65. G. J. McCarthy and M. T. Davidson, "Ceramic Nuclear Waste Forms: II. A Ceramic-Waste Composition Prepared by Hot Pressing," *Ceram. Bull.* **55**, 190 (1976).
66. A. D. De, G. Luckscheiter, W. Lutze, G. Malow, and E. Schiewer, "Development of Glass Ceramics for the Incorporation of Fission Products," *Ceram. Bull.* **55**, 500 (1976).
67. "Ceramic Materials for Surgical Implants," unpublished analysis, National Bureau of Standards, Inorganic Materials Division, Washington, D.C., June 1976.
68. W. K. Haller, "Rearrangement Kinetics of the Liquid-Liquid Immiscible Microphases in Alkali Borosilicate Melts," *J. Chem. Phys.* **42**, 696 (1965).
69. R. A. Messing, "Controlled-Pore Ceramics," *Research/Development* **25**, 32 (July 1974).
70. H. H. Weetall, "Preparation, Characterization, and Applications of Enzymes Immobilized on Inorganic Supports," in *Immobilized Biochemicals and Affinity Chromatography*," R. Bruce Dunlap, ed., Plenum Press, New York; reprinted in *Corning Research 1974*, Corning Glass Works, Corning, N.Y., 1974.



John T. Yates, Jr., joined the National Bureau of Standards in 1963 as a National Research Council Postdoctoral Research Associate. Dr. Yates has done research in surface chemical physics using thermal desorption spectroscopy, electron impact desorption, infrared spectroscopy, work function measurements, and X-ray photoelectron spectroscopy. In 1977-1978 he will be a Sherman Fairchild Distinguished Scholar at the California Institute of Technology. Dr. Yates is a graduate of Juniata College in Huntingdon, Pa., and of MIT. He has served on a number of committees involved with surface science in the American Vacuum Society and the American Chemical Society.



Theodore E. Madey joined the National Bureau of Standards as a National Research Council Postdoctoral Research Associate in 1963. He has worked in the fields of surface physics and surface chemistry using the techniques of thermal desorption spectroscopy, electron impact desorption, field emission microscopy, work function measurements, and X-ray and ultraviolet photoelectron spectroscopy. Dr. Madey is a graduate of Loyola College in Baltimore, Md., and of Notre Dame University. He has served on a number of American Vacuum Society and American Physical Society committees concerned with surface science.

PROSPECTIVES FOR SURFACE CHEMISTRY

John T. Yates, Jr., and Theodore E. Madey

National Bureau of Standards
Washington, D.C.

An entirely new area of structural chemistry is just beginning to unfold, in much the same manner as organic and inorganic structural chemistries evolved in times past. This is the area of structural chemistry at surfaces. Intense scientific activity is beginning to reveal the structural and electronic details of surface layers of the order of one atomic diameter in thickness. In the last 15-20 years, our appreciation of the nature of adsorbed monolayers on solid surfaces has evolved from a position of essentially *no understanding* at the atomic level to our current position, in which atomic structure, orbital configuration, and bond energetics for surface species may be measured and calculated.

Since surface chemistry is so pervasive in its importance to broad areas of technology (heterogeneous catalysis, corrosion prevention, energy transfer at surfaces, strengths of materials, semiconductors, adhesion, lubrication, etc.), we believe that the evolution of a structural chemistry of surfaces will eventually result in major consequences in our technological age. The case for this assertion is made by analogy to the historical fact that the single most important event in the development of organic and inorganic chemistry has been the placing of these fields on a sound structural basis. Thus, the principles for tailor-making complex *molecules* with specific chemical reactivities at specific sites are fairly well understood theoretically; these principles rest on a firm knowledge of atomic and electronic structure. We believe that our ability to tailor-make the chemical properties of *surfaces* also will begin to evolve in the near future, and that this ability will be founded on an understanding of structural and electronic properties of surfaces and of adsorbed layers on surfaces.

Recent insights into the details of chemical bonding at surfaces are based on very effective cooperation between chemists and physicists who are jointly involved in a field termed "surface science." One result of this work has been a new arsenal of surface measurement techniques. In addition, new concepts and theories of surface behavior have evolved from both disciplines. Chemists have long been interested in the reactivity of the surfaces of many materials, dating back to the pioneering work on chemisorption by Irving Langmuir. An atomically clean surface is often an extremely reactive entity, in many cases exhibiting an adsorption-reaction probability of *unity* due to the presence of reactive unsaturated surface orbitals ("dangling bonds"). Physicists have tended historically to regard the surface as a window to the bulk solid, but in the last 15 years surface physics has also focused on the physics of the surface atoms or molecules themselves. This mutual concern for the physical and chemical properties of surface species has

substantially strengthened the liaison between surface chemistry and surface physics in recent years.

In broad outline, this chapter will first discuss a number of examples illustrating where we are now in our fundamental understanding of the behavior of surfaces. This is of importance in forming the basis for the next section, which deals with the possibilities for direct extension of our present knowledge and measurement ability to new areas. A final section is concerned with a more long-range view of the research directions in which surface science may possibly evolve, as well as a discussion of certain needs of the field that are not attainable by current knowledge or experimental methods. The emphasis of the chapter is on the gas-solid interface, but it should be emphasized that many of the concepts and methods of surface science are also applicable to liquid-solid and solid-solid interfaces.

THE PRESENT: SURFACE CHEMISTRY TODAY

An explosion of experimental and theoretical developments in the past 10 years has led to a new understanding of the atomistics of surface processes. The characterization of gas-solid interactions is based on developments in a variety of diverse, yet related, areas. Essential elements in the description of surface processes include such factors as

1. *The Chemical Characterization of the Surfaces*—What elements are present, and in what concentration?
2. *The Geometry of the Surface*—Is the surface structure simply an extension of the bulk, or does rearrangement of surface atoms occur? Where are the sites at which chemisorbed atoms are bound? Under what conditions are surface compounds formed?
3. *The Electronic Character of the Surface*—What is the distribution in energy and in space of the surface valence (bonding) electrons? What is the relationship between chemical reactivity at surfaces and the electronic structure of surfaces?
4. *The Dynamics of Surface Processes*—What factors control the rates of adsorption and desorption of atoms and molecules? What factors determine the rates of surface reactions and catalytic processes?

In the present section, each of these topics will be broadly treated in turn. The emphasis here is on a description of where we have been, and where we are now. To this end, it is important to define the geometrical limits of this treatment. The surface region of a solid, as considered here, is defined as the interfacial layers between solid and gas; for metals, the surface region extends no more than a few atom layers into the bulk.

Chemical Characterization of the Surface: Elemental Analysis

Fundamental to a description of the surface layer is characterization of its chemical composition. Ten years ago there were no reliable, widely-used techniques for chemical analysis of surfaces of unknown composition. Today, every modern surface laboratory should have at its disposal a variety of techniques that can detect and chemically characterize fractional monolayer quantities in the surface region (i.e., less than 1% of a single atomic layer, where typical monolayer surface densities are 10^{15} atoms/cm^2). These methods, described in more detail below, have several common features. Most of them involve exposing the surface to ionizing radiation (X-rays, electrons, or ions) and analyzing charged particles scattered from or emitted from the surface. The use of charged particles dictates that the measurements be performed in a vacuum environment, usually 10^{-6} to 10^{-10} Torr. These methods are qualitative; accurate quantitative analysis is not easily attainable. Using these methods, typical qualitative analysis times for the surface of an unknown sample are of the order of an hour. The more widely used techniques [1] include

Auger Electron Spectroscopy (AES)—In AES, a surface is bombarded with a focused beam of high-energy (3-10 kV) electrons. Electrons emitted from the surface have different characteristic energies, depending on the chemical identity of the surface atoms, and are detected using an electron energy analyzer. Advantages are high surface sensitivity (one can detect ~ 1% of an atomic

layer in a probe depth of 5 to 10 atomic layers), and lateral resolutions in the range 1 to 100 mm. Scanning Auger microscopy (SAM) provides an elemental two-dimensional map of surface chemical composition.

X-ray Photoelectron Spectroscopy (XPS)—This technique is also known as Electron Spectroscopy for Chemical Analysis (ESCA). A surface is bombarded by a flux of nearly monochromatic X-rays, and electrons photoejected from core levels of surface atoms are detected with an electron energy analyzer. XPS can discriminate both elemental composition and charge state (oxidation state) of surface atoms. Surface sensitivity is ≈ 5% of an atomic layer in a probe depth of 5 to 10 atomic layers; spatial resolution is limited to several mm^2 at present.

Ion Bombardment Methods—In ion-scattering spectrometry (ISS), the surface is bombarded by a beam of He^+ ions having several keV of energy. The inelastically scattered He^+ ions are energy analyzed, and the energy loss is related to the atomic mass of the atoms in the surface (i.e., its chemical composition). The probe depth is one atomic layer, with a lateral resolution of ≈ 1mm. In secondary ion mass spectrometry (SIMS), a sample is bombarded with energetic ions (5 to 30 keV). These primary ions impart energy to surface atoms of the sample, causing sputtering (removal) of atoms at or near the surface. A fraction of the sputtered atoms escape as ions and are mass analyzed. Sampling depths are a few atomic layers.

The implication of these and other surface analytical techniques is that qualitative elemental analysis with fractional monolayer sensitivity is routine. The surface chemist can initiate experiments on a surface of known cleanliness, monitor surface concentration during adsorption and reaction, and study the influence of known quantities of impurities on the courses of reactions (e.g., catalytic poisons and promoters). Chemical analysis of surfaces is of practical importance in microelectronics, in catalysis and corrosion, and in metallurgy. AES combined with ion sputtering provides a depth profile of the surface region of a sample; continuous AES analysis as the sample surface is eroded by sputtering enables one to sample composition as a function of depth over thicknesses of thousands of angstroms.

It is in the area of surface analysis that we anticipate some of the most far-reaching developments in the understanding of practical surface processes. Many of these developments are already underway, and future prospects will be discussed in the following sections.

The Geometry of Surface Layers

A picture of the geometrical arrangement of atoms in the outermost surface layer is basic to an understanding of surface processes. Long before tools for determining surface structures were available, surface chemists often formulated structural models to visualize and rationalize surface kinetic processes. At present, surface structural techniques are available and are widely exploited. Studies of surface geometrical structures can be broadly classified in two ways. Firstly, there are surface structures classified by long-range, periodic order (as on the surface of a single crystal). Secondly, there are surface structures in which a short-range local bonding configuration is maintained, but in which long-range order is absent. Such may be the case for surfaces of polycrystalline or amorphous materials, or even for single crystals containing disordered (in the long-range sense) overlayers. Each of these will be considered below. The broad area of surface topographical measurements using high-resolution optical and electron microscopy will not be treated. Such techniques at present are not capable of providing details of atomic arrangements and bonding geometry at surfaces.

Today, surface studies performed under ultrahigh vacuum conditions are concerned primarily with surfaces possessing long-range order, i.e., close packed or nearly close packed faces of single crystals of metals, semiconductors, and insulators. In the case of clean surfaces, the primary question is whether the surface atoms have the same geometrical arrangement as the atoms in the bulk, or whether they assume relaxed or rearranged positions. In the case of adsorbed layers, the questions involve the position of adsorbed atoms, penetration of the surface species into the bulk, and surface compound formation. Examples of long-range order in surface structures on single crystals are shown in Figure 1.

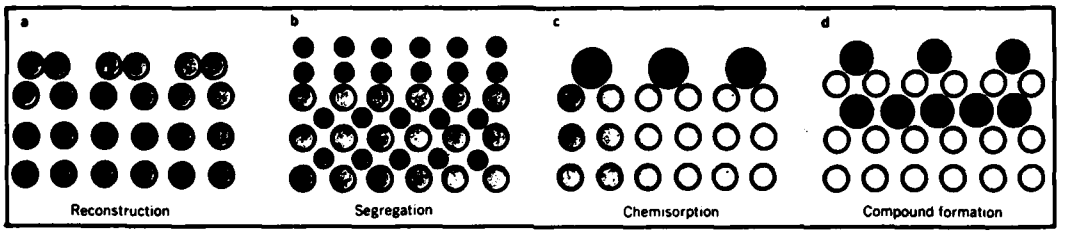


Figure 1—Four ways in which the surface of a crystal may differ from the bulk: (a) reconstruction by contraction of the interlayer spacing; (b) alloy in which one component has segregated to the surface; (c) chemisorption of foreign atoms on the surface; (d) two-dimensional compound formation at a surface following chemisorption (Courtesy Prof. P. J. Estrup)

The most widely used method for studying long-range order on crystal surfaces is low-energy electron diffraction (LEED) [2]. The basis of this method is the wave nature of the electron. A monoenergetic beam of electrons having wavelengths comparable to a crystal lattice spacing is directed at a crystal surface. If the surface atoms are arranged in a periodic array, they act as a grating for the electron waves and diffract them. Discrete electron beams are then scattered back from the surface at angles that depend on electron energy, incident angle, and surface two-dimensional periodicity. This results in a diffraction pattern that can be visually displayed on a fluorescent screen as an array of symmetrically arranged bright spots (Figure 2). Adsorption of a layer of foreign atoms frequently results in changes in surface periodicity that cause changes in the diffraction pattern. The details of atom arrangement in the adsorbed layer are contained in the intensity of the diffracted beams as a function of electron energy, and much experimental and theoretical effort is devoted to extracting this information.

LEED studies have revealed that in general, the atomic geometries of clean metal surfaces are close (within 5%) to those of the corresponding planes of atoms in the bulk solid. Purely covalent group IV semiconductors exhibit substantial surface atom rearrangements, and more ionic binary semiconductors undergo ionic reconstructions in which the smaller cations shift positions more than the anions.

A major triumph of LEED has been the revelation that mobile adsorbed atoms on single crystal surfaces frequently form ordered layers having periodicities different from the substrate crystal. Some of the adsorbate structures have lattice parameters several times greater than the substrate lattice spacing. This suggests the existence of long-range lateral interactions between adsorbed atoms, involving lateral forces much greater than those existing between free atoms at the same spacing. Thomas Grimley, Robert Schrieffer, and Theodore Einstein have demonstrated theoretically that such periodicities are the

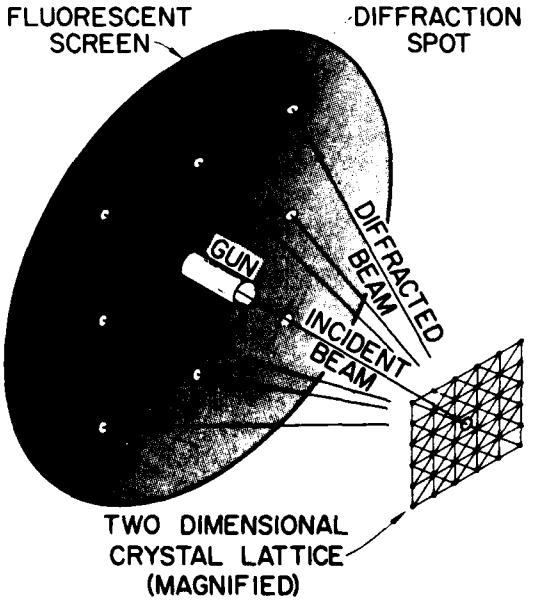


Figure 2—Schematic low-energy electron diffraction. Monoenergetic electrons incident on a two-dimensional crystal lattice yield diffracted beams. The diffracted electrons are accelerated to a phosphor screen, giving a pattern related to the two-dimensional crystal lattice spacings. (Courtesy Varian Associates)

result of long-range interaction *through* the substrate producing oscillatory attractive and repulsive interactions as interatomic distances change. These adatom-adatom interactions are also responsible for the growth of many adsorbed layers in "island" structures rather than random adsorption (i.e., islands of adsorbed atoms in a regular array surrounded by clean surface). In the limit of high (almost monolayer) coverage of an adsorbed molecule such as CO on Ni or Cu and Pd, the overlayer structure sometimes shows a periodicity that is uncorrelated to the substrate symmetry. The adsorbed species do not occupy crystallographic sites, and the geometry of the layer is determined almost completely by lateral interactions between them [3].

In another related example, Gert Ehrlich and T. Tsong have observed dimer and trimer structures of adsorbed atoms on atomically perfect surfaces studied with field ion microscopy. The adsorbate structures are bound by lateral forces that extend from one atomic "furrow" to its neighbor furrow in the substrate surface crystal structure. The adsorbate atom composing the dimer and trimer species are often observed to migrate together back and forth in the separate furrows.

Although the LEED diffraction patterns contain information related to the *distances* between adsorbed species, the *location* of the binding sites on the substrate (i.e., where the adsorbed species sit with respect to the substrate atoms) can only be extracted from the intensities of the different diffracted beams, as a function of electron energy. This is a difficult problem, requiring precise measurements of many diffraction beams and highly detailed theoretical calculations. From such analyses, Joseph Demuth, Donald Jepsen, and Paul Marcus have located binding sites for O and S atoms adsorbed on close-packed Ni surfaces. We anticipate that developments in LEED calculations will make such determinations more reliable and widespread in the future.

As will be discussed in the section "Dynamics of Surface Processes," there are a number of kinetic phenomena whose rates are sensitive to the geometry of the surface layer. The rates of adsorption and desorption of different molecules can vary greatly from one crystal plane to another. Michel Boudart has shown that there are different kinds of catalytic reactions, which can be classified by their dependence on surface structure. Gabor Somorjai has proposed that steps and kinks on single crystal surfaces are essential elements in the catalysis of certain hydrocarbon reactions.

A major deficiency of present surface structure analysis occurs in our knowledge about surface defects, despite their apparently important role in catalysis, crystal growth, and electronic properties of semiconductors [4]. A substantial question arises concerning the transferability of results obtained for flat single crystals to the case of rough, polycrystalline, or amorphous surfaces of technological interest in catalysis and metallurgy. Although this question is far from resolved, it appears that bonding structures deduced on single crystals are valuable inputs in developing theories of surface chemical bonding—theories that are presently in their infancy. Having formed an understanding of bonding on idealized surfaces, it may be anticipated that theoreticians will be better able to treat bonding on more complex surfaces of technological importance.

Experimental methods are evolving which appear to be capable of yielding information on the short-range bonding order at steps, defects, and amorphous surfaces. One such method is extended X-ray absorption fine structure (EXAFS)[5]. Richard Stern, Farrel Lytle, and Dale Sayers have shown that tiny wiggles in the characteristic X-ray absorption "signature" of an atom embedded in a solid can now be interpreted to provide clues to the exact spatial arrangement of the neighboring atoms. Analysis of the EXAFS of certain catalysts has suggested that it may be possible to determine the active site of catalysis and its local chemical environment.

Another method with potential for studies of short-range order at surfaces is electron-stimulated desorption (ESD). In ESD, a surface containing an adsorbed monolayer is bombarded by low-energy electrons (10-1000 e V), and ions and neutral fragments are desorbed from the surface due to electronic excitation of the adsorbed species. Recently, Theodore Madey and John Yates at the National Bureau of Standards have shown that O^+ ions liberated from an adsorbed layer of oxygen on W(110) by electron-stimulated desorption (ESD) are characteristic of adsorption at atom steps. The angular distribution of ESD

ions is sensitive to local bonding geometry rather than long-range order, and the direction of ion emission appears to be related to the direction of the surface chemical bond (Figure 3). The potential of this method for studies of the role of steps in catalysis and the structure of catalytic intermediates is intriguing.

In summary, it is clear that surface structural analysis is moving toward an understanding of the surfaces of technologically important materials. The broad jump from single-crystal to amorphous surfaces is underway.

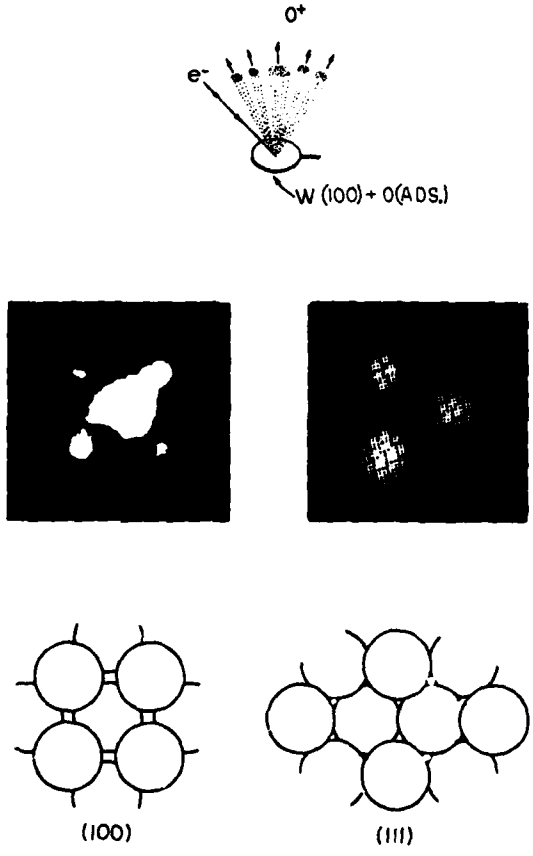


Figure 3—Electron Impact Desorption–Ion Angular Distribution. An electron beam bombarding a chemisorbed oxygen layer on tungsten single crystals causes O^+ ion ejection in specific directions as shown schematically at the top. Representative ejection patterns from W(100) and W(111) single crystals are shown in the photographs. The O^+ beams are ejected in directions corresponding to bonding directions and site symmetries of the chemisorbed oxygen species.

Electronic Properties of Surfaces and Surface Species

In the last few years there has been a major change in our physical thinking about the properties of surface atoms, or adsorbed species on surfaces. The original belief that the *collective* electronic properties of the bulk substrate were directly related to its surface behavior has, in some cases, been displaced by a more *local* view of the electronic factors of importance at surfaces. Both theory and experiment are currently focused on molecular orbital descriptions of surfaces and chemisorption bonds at surfaces. The formation of directed chemical bonds with adsorbates at the surface appears to be analogous in many cases to bonding in molecules. It has been known for a long time (through studies of the infrared spectra of adsorbed species) that chemisorption often produces surface ligands similar in their vibrational spectra to analogous ligands* in molecules [6]. The most well known example of this is the relation between the chemisorbed state of CO on transition metals and the CO ligands that exist in transition metal carbonyls. Both sp-hybridized linear CO species M-C≡O and sp^2-hybridized bridged CO species $\begin{matrix} M \\ M \end{matrix}\!\!>\!C{=}O$ are believed to exist on surfaces as they do in certain metal carbonyls (M is a surface metal atom).

Recently, ultraviolet photoelectron spectroscopy (UPS) has been used effectively in observing the involvement of specific molecular orbitals in forming chemisorption bonds. In UPS, one photoejects valence-level electrons from a surface layer using monochromatic ultraviolet radiation. An energy distribution of emitted electrons related to the density of electronic states near the surface is obtained. This permits observation of the energy and density of both adsorbate and adsorbent electrons. When the covalent adsorption bond is formed, it is then possible from the spectrum to determine which electrons are involved. A recent study by Joseph Demuth and Dean Eastman [7] of the adsorption of ethylene (C_2H_4)

*A ligand is a molecule or ion coordinated to a central atom in a chemical complex. For example, ammonia molecules (NH_3) are ligands in the complex ion $Cu(NH_3)_4^{++}$.

by a Ni (111) single crystal surface (Figure 4) has shown that the π-electrons in the C_2H_4 double bond overlap with d-electrons, possibly from single Ni atoms. Both the π- and d-electrons are shifted to higher binding energy in the molecular complex to form the chemisorption bond. Only small energy shifts of the other bonding C-H and C-C σ-electrons in the C_2H_4 molecule are observed, indicating only slight distortion in the geometry of the planar C_2H_4 molecule upon chemisorption. This form of π-bonding of olefins to transition metals has been recognized for about 15 years in metallo-organic compounds. Similarly, principal involvement of π-electrons in acetylene (C_2H_2) and benzene (C_6H_6) chemisorption by Ni has also been observed. In the case of CO chemisorption by a number of transition metals, it appears that bonding occurs via the 5σ lone pair electrons on the carbon atom, and that this is accompanied by shifts of the CO-π-electrons to lower binding energy [8].

The molecular orbital picture of chemisorption leads naturally to attempts to calculate the energy levels of an aggregate of adsorbate atoms interacting with a chemisorbed molecule. One of the more successful methods used by Keith Johnson and Richard Messmer is the SCF-Xα-SW (Self-Consistent Field—involving exchange correlation parameter α—Scattered Wave formalism).* The Xα method has been used to calculate the energy level diagram of clusters of from 2 to 13 metal atoms, and for model chemisorption systems such as S and CO on a cluster of 5-Ni atoms. The electronic state density and position of the electronic energy levels in these systems are in reasonable agreement with UPS experimental measurements for S and CO adsorbed on macroscopic Ni crystals.

With respect to semiconductor surfaces, Joel Applebaum and Donald Hamman at Bell Telephone Laboratories have devised methods for calculating the complete electronic structure of a realistic model of a solid surface. The ion cores of

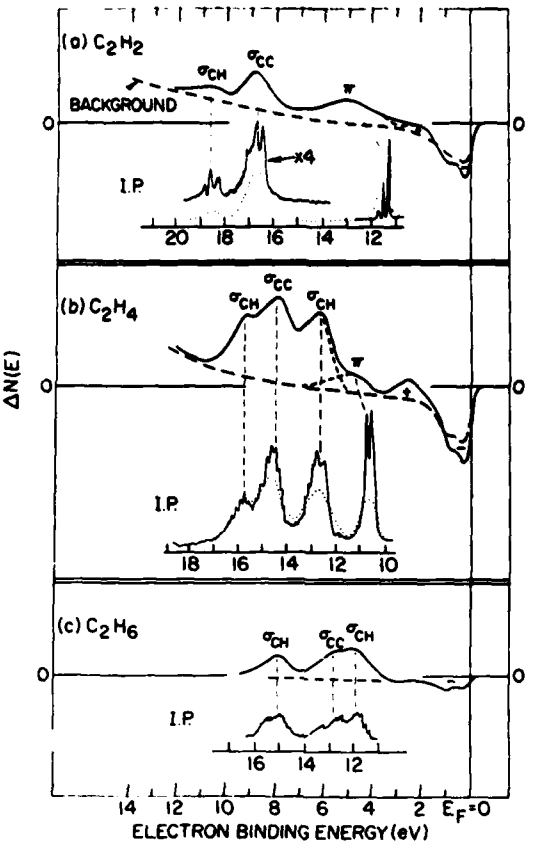


Figure 4—Ultraviolet photoelectron difference spectra for hydrocarbon adsorption on Ni(111). (Ionization potential is denoted by I. P.):
(a) Comparison of chemisorbed acetylene with gas phase acetylene. The UPS spectrum indicates that the σ orbitals are uniformly shifted upon chemisorption, whereas the π orbitals have undergone a preferential increase in binding energy due to their extensive involvement in formation of the chemisorption bond.
(b) Same comparison for chemisorbed ethylene
(d) Same comparison for physically adsorbed ethane. Note that all σ orbitals agree in relative energy to each other.
(Courtesy Dr. Joseph Demuth)

*The method involves partitioning the aggregate of adsorbate atoms into Wigner-Seitz spheres centered on each of the atoms of the cluster. Each sphere encloses a spherically s, - metric electronic potential. These spheres are surrounded by a spherical region of constant potential, and outside this larger sphere the external region again is given a spherically symmetrical potential. Schroedinger's equation for the valence electrons is solved in each of these regions, and the wave functions and their first derivatives are joined at the potential boundaries. The charge density throughout the cluster is then computed and is used together with Poisson's equation and the Xα-exchange correlation to generate a new electronic potential covering the cluster, for which Schroedinger's equation may again be solved. The iteration is repeated until self-consistent results are obtained.

a semi-infinite solid are represented by their pseudopotentials, and the Hartree and exchange potentials are treated self-consistently. Both the potential and the charge density obtained from the calculation are displayed visually in contour projections, and resemble molecular orbital charge-density representations in molecules. The chemisorption of hydrogen atoms by Si(111) has also been theoretically studied to yield bond distances and Si-H force constants that are in agreement with experiments. The calculations also yield an electronic spectrum in good agreement with ultraviolet photoemission measurement.

One of the fundamental problems involving the electronic character of surfaces is the question of the influence of d-electrons in causing many transition metals to be good heterogeneous catalysts for certain classes of chemical reactions (hydrogenation of carbon monoxide, oxidation of CO and H_2, reduction of NO, hydrogenation of alkenes, dehydrocyclization of n-heptane to produce toluene, hydrocarbon hydrogenolysis where C-C bonds are broken and converted to C-H bonds, etc.). John Sinfelt has carefully studied the specific catalytic activity (activity per *surface* metal atom) of transition elements in the first, second, and third transition series in the periodic table for a model reaction—the hydrogenolysis of ethane, C_2H_6, to produce CH_4 [9]. In Figure 5, one sees that there is enormous variation in catalytic activity (a factor of $\approx 10^7$) observed in moving from Ru → Rh → Pd (second series) or from Re → Os → Ir → Pt (third series). In both cases, maximum activity of the group VIII_1 element (Ru, Os) is observed. A similar trend does not occur as one moves across the first series elements, Fe, Co, Ni. Sinfelt has concluded that the percentage d-character of dsp hybridized orbitals forming the metallic bonds in these metals cannot be directly correlated with catalytic activity for this model reaction, although there are suggestive similarities in trends of activity and percentage d-character. At present, a simple way of correlating catalytic activity with a bulk "electronic factor" in the metal catalyst does not seem to exist.

It is also important to note that in general the transition metals are more active in chemisorption than the nontransition metals. This specificity may be related to the presence of partially filled d-orbitals that are localized in space; these d-orbitals may, therefore, overlap electronically with orbitals in ligands, leading to covalent bond formation. For the same reason, the transition metals also exhibit a rich coordination chemistry with π-electron molecules. Thus the catalytic activity of the transition metals may be related to the stabilization of transient ligands, or, in other words, to a lowering of activation energies for chemical transformations.

In the case of insulator catalysts such as MgO or WS_2, recent work by Michel Boudart and Rudie Voorhoeve has demonstrated a linear correlation of the catalytic activity of these materials with the measured surface concentration of electronic defect sites. This correlation of reaction rate has been extended over 3-4 orders of magnitude of defect site concentration, as measured

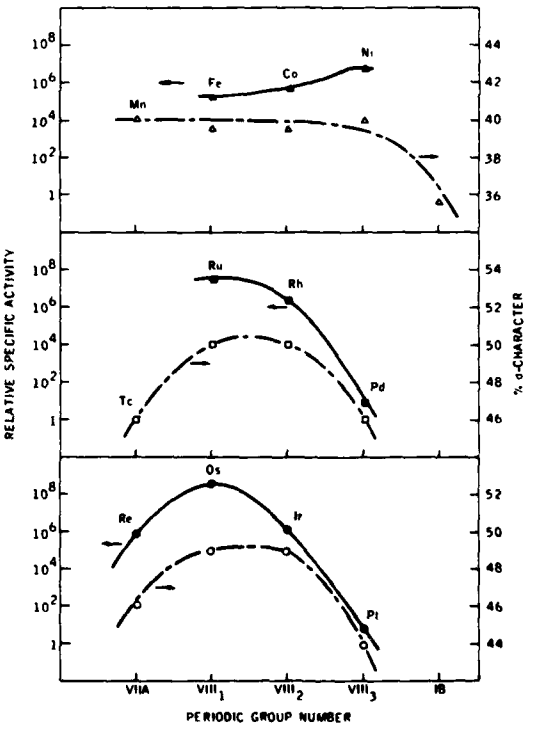


Figure 5—Relationship between catalytic activity for ethane hydrogenolysis ($H_2 + C_2H_6 \rightarrow 2\ CH_4$) and the percentage d-character of the metallic bond of the bulk solid catalyst. The three panels are shown to distinguish the metals in the different long rows of the periodic table. (Courtesy Dr. John Sinfelt)

by electron spin resonance (ESR). In both cases, it was possible to modify the catalyst to enhance catalytic activity by increasing the defect concentration—the first step in tailor-making a catalyst by controlled and understood processes. Unfortunately, for heterogeneous catalysis by metals, little information about the geometrical or electronic nature of active sites is currently available, except for the observations by Somorjai that stepped and kinked surfaces of Pt seem to be superior for certain catalytic reactions involving hydrocarbons.

It is appropriate to close this section on the electronic nature of surfaces by mentioning some current technological areas significantly affected by the electronic properties of surfaces. James Murday has pointed to the wide range of electronics devices made possible by control of electronic properties at interfaces [10]. These range from vacuum tubes using low-work-function thermionic emitters to tunnel diodes and transistors, in which control of solid-solid interfaces is of major importance. In addition, in electrochemistry, the electronic properties of electrode materials must be fundamental to their operation in electrochemical cells. Some modern methods of surface characterization (LEED, AES, XPS) are currently being applied under ultrahigh vacuum conditions to electrode surfaces in several laboratories. Finally, the tailor making of optically selective filters for efficient collection of solar energy depends on knowledge of electronic properties of thin film materials, and on the use of modern methods of surface analysis and ion sputtering depth profiling techniques for control of the properties of the thin films.

Dynamics of Surface Processes

While knowledge of the composition and the geometrical and electronic character of surfaces is of great importance, it is most often control of the pathways and rates of chemical processes at surfaces that is ultimately sought in the technological uses of surface chemistry. Consider heterogeneous catalysis, corrosion prevention, electrochemistry and electrocatalysis, failure of metals and alloys due to impurity segregation at grain boundaries, plasma hardening of metal surfaces, etc. All of these processes depend upon the *rate* of surface processes.

Control of the rates and product distributions of heterogeneous catalytic processes is one of the basic themes of surface chemistry research. A number of significant developments have recently occurred in this field. To measure the specific rate of a catalytic reaction over a powdered catalyst, or a supported metallic catalyst involving metal catalyst particles that may be smaller than 100 Å in diameter, it is first necessary to determine the number of surface metal atoms that can serve as catalytic sites. This can now be done with fair accuracy by measuring the chemisorptive uptake of "standard" molecules such as O_2, H_2, or CO. This simple "normalizing" procedure permits the specific catalytic rate *per atomic surface site* to be measured as a function of the average particle size of the catalyst particles. By studying *specific rates*, Michel Boudart has shown that catalytic reactions may be roughly divided into two classes. The first class of reactions is called "facile," or structure insensitive, and the reactions in this class exhibit specific rates *independent* of average particle size. Since it is probable that the catalyst particle size is a controlling factor in determining the atomic structure of the surface regions of the particle, the facile reactions are thought to be insensitive to surface crystallography, at least to a first approximation. The second class of reactions is termed "demanding" and reactions in this class are thought to be structure sensitive due to the marked dependence of specific catalytic rate on catalyst particle size.

In the case of one model facile reaction, the hydrogenolysis of cyclopropane to propane over Pt, it has been found by careful kinetic studies that the specific rate of the reaction is invariant over about 7 orders of magnitude in particle size, ranging from approximately 10-Å supported Pt particles to macroscopic Pt single crystals. This landmark kinetic study, performed jointly in the Berkeley laboratories (by D. Kahn, E. E. Peterson, and G. Somorjai) and the Stanford laboratories (by M. Boudart) is of major importance because it illustrates clearly that in at least some cases one may make studies on single crystals (using the newer surface measurement tools) that can be related directly to catalytic processes on actual supported catalyst particles.

The effects of crystal structure on certain catalytic reactions have been studied using different single-crystal catalysts. Robert Rye and K. Lu studied the $H_{27}D_2$ exchange over various Pt single-crystal planes. Specific rates differed by about a factor of 2. Robert Hansen and Jerome McAllister studied the decomposition of NH_3 over three single crystals of W and again found about a factor-of-10 difference in specific rate for different crystal planes. The oxidation of CO on Pd was studied by Gerhard Ertl and J. Koch and was found to be insensitive to crystal structure even though the initial heats of chemisorption of CO were found to vary from 34 to 40 kcal/mol on the different planes. Thus, to date, it must be concluded that a major effect of surface geometry causing orders of magnitude change in catalytic rates has not been detected.

Studies of the effect of crystal structure on the rate of chemisorption have detected differences in adsorption rates of many orders of magnitude from plane to plane. Thus, from the field emission work of Gert Ehrlich and his students, it has been shown that the smooth, close packed planes of tungsten and rhenium are inactive for the chemisorption of molecular N_2 or molecular H_2. However, chemisorption with dissociation occurs on rougher planes of these metals and migration of adsorbate *atoms* from rough planes to smooth, close packed neighboring planes can then occur.

While the rate of a catalyzed reaction is of practical importance, it is also important from a conceptual point of view to know the mechanism of the catalytic surface process. What are the structures of the catalytic intermediates, and how does the catalyst lower the activation energy of the rate-determining step? At the present time, it must be said that very little is known about these matters. This gap in our knowledge is partially related to the low surface concentration of many transient intermediate species, making spectroscopic detection difficult or impossible. It is also related to the present lack of application of the techniques of physical organic chemistry to catalytic processes on well-defined surfaces. More studies on well defined catalytic surfaces using isotopic labeling, reactive intermediate injection, stereochemical design, and spectroscopies of high sensitivity are needed.

As an example of the influence of molecular excitations on catalytic reaction rates, Gert Ehrlich and Charles Stewart have been able to demonstrate that vibrational excitation of the CH_4 molecule can activate it sufficiently to cause it to chemisorb on a Rh surface. A significant retardation of the reaction occurs when CD_4 is used. Jerome McAllister and Robert Hansen demonstrated that NH_3 decomposition over tungsten single crystals proceeded via two separate mechanisms and that one of these processes exhibited an isotope effect, seen when ND_3 was compared to NH_3. The involvement of hydrogen in the transition state for these two catalytic reactions is indicated from these results.

Robert Merrill and Henry Weinberg have developed a method of conceptualizing the energy surfaces related to adsorption and catalytic reaction. This procedure, called the Crystal Field Surface Orbital-Bond Energy Bond Order (CFSO-BEBO) method, is a semiempirical method for visualizing electronic energy changes as chemisorption bonds are formed at the surface, accompanied by bond weakening and distortion in the adsorbing molecule.

The modern techniques of surface analysis have recently been directed to the study of diffusion of alloy components to the surface. With some alloys, it has been found that extraordinary differences in equilibrium bulk and surface alloy composition are present. In general, in a binary alloy, the component having the higher vapor pressure at the temperature of annealing is found to segregate in the surface region. Since alloy catalysts are often used in the chemical industry, effects of this type are of major importance, although it must be remembered that the surface segregation processes are likely to be different in small particles compared to processes observed with macroscopic specimens. Another interesting and related phenomenon has to do with the influence of chemisorption on alloy surface composition. It has been observed that an alloy composed of two metals, one of which is active in chemisorption of a particular gas, will often surface segregate upon chemisorption, such that the surface is enriched in the active metal.

So far, we have discussed a few examples of the influence of surface structure on catalytic and chemisorptive reaction rates, and on bulk-to-surface diffusion processes. Perhaps one of the

most common and least understood factors in determining rates of catalytic processes is the influence of foreign substances, called poisons or promoters, on the rates of catalytic reactions. The poisoning phenomenon is so significant that major costs are often incurred to reduce catalyst poisons to very low levels in feed gas streams in order to protect industrial catalysts from failure. In Figure 6, Hans Bonzel and R. Ku have shown that ≈ 0.15 monolayer of S poison will reduce the rate of the CO (ads) + O(ads) reaction to form CO_2 by a factor of 10 on a Pt (110) single crystal [11]. In addition to poisoning, the catalytic promotion factor is equally significant, and most commercial catalysts contain traces of additive that promote activity qr lead to enhanced selectivity or long life.

In some cases, the presence of surface carbon on transition metal catalysts seems to influence the course of catalytic reactions profoundly. For example, in well-controlled experiments, Robert Madix and his coworkers have examined the influence of surface carbon on the character of the classic formic acid decomposition reaction on Ni surfaces. Studies combining AES and mass spectrometry have shown that the presence of a carbide layer on a Ni(110) crystal surface significantly alters the products of decomposition. The detailed role of surface carbon is not understood.

Gabor Somorjai and his coworkers have deduced that an ordered carbonaceous overlayer is necessary on stepped Pt surfaces to produce conditions necessary for the dehydrocyclization of n-heptane to toluene, an important reforming reaction in the petrochemical industry. Again, the detailed role of the carbon layer is not understood.

Thus, we see that while many studies of surface phenomena related to catalytic processes are now possible with modern instrumentation, we are not yet at a stage where detailed structural or mechanistic models with predictive power have evolved.

SURFACE CHEMISTRY: WHERE DO WE GO FROM HERE?

In assessing the future of surface chemistry (or of any other field of scientific endeavor) the crystal ball is necessarily clouded. On the one hand, we can point to gaps in our understanding of surface processes. We can, with some degree of certainty, predict that there will be transfer of the concepts and models gleaned from studies of idealized model systems to systems of practical and technological importance. With more certainty, we can predict that many of the experimental and theoretical concepts developed for studies of model systems will see wide and far-reaching application in such diverse areas as catalysis, corrosion, electronics, and adhesion. Indeed, substantial efforts are already underway in many of these areas. On the other hand, it is significantly more difficult to anticipate the major scientific and technological breakthroughs that can result in a great leap forward in knowledge. Twenty years ago, few if any could have envisioned ordered surface overlayers, or measurement of the molecular orbital structure of adsorbed molecules, or the ability to detect and characterize less than 1% of a monolayer of adsorbed impurities. To a large extent, these developments have been tied inexorably to technological advances in areas such as ultrahigh vacuum technology, the development of the Auger spectrometer, and the commercial availability of high-purity single crystals.

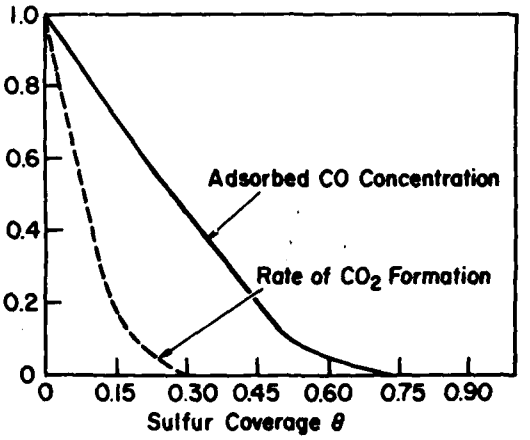


Figure 6—Effect of sulfur as a poison in the oxidation of CO to form CO_2 on a (110) Pt surface. Composite plot of the normalized adsorbate CO concentration and the normalized rate of CO_2 formation as a function of the calibrated sulfur coverage θ. (Courtesy Prof. H. Bonzel)

Despite reservations about our predictive powers, we will proceed with both general and specific suggestions concerning future developments in the science and technology of surfaces. Many practical problems of concern to the technological

community in general, and to the Navy in particular, are controlled by surface or interfacial processes in environments seemingly incompatible with the high-vacuum surface analysis tools developed to date. We simply cannot use Auger spectroscopy, ESCA, LEED, etc. for in-situ studies of corrosion in aqueous or high-pressure gaseous environments, of catalytic processes at high temperatures and pressures, of lubrication, friction, and wear. However, the development of methods that would allow rapid transfer of a sample from its operational environment directly to the high-vacuum measurement chamber *without exposure to intervening, contaminating atmospheres* would provide a unique "snapshot" of the chemical and physical state of the surface under conditions simulating the real thing. Techniques of this sort are currently under development for use in model studies of both electrochemical and catalytic processes on single crystals. Whereas such methods will allow "before-and-after" examination of the sample surface, it may be difficult to avoid changes in surface composition due to evaporation of reactants as the atmosphere above the surface is reduced to high vacuum.

A real need exists for the development and exploitation of new experimental methods for the study of both model and practical surfaces in-situ in high-pressure gaseous and liquid environments. Methods based on charged-particle analysis (electrons and ions) have limited utility under such conditions. Optical and acoustic spectroscopies, including X-ray spectroscopies (such as the EXAFS method described previously), Mossbauer spectroscopy, magnetic resonance spectroscopies, and photoacoustical spectroscopy appear to offer promise for in-situ application. Just as it will be important for surface scientists to extend their measurements on model (usually single-crystal) surfaces from the ultrahigh vacuum range up to practical high-pressure conditions, it will be equally important to apply both existing and new methods to the study of model processes on polycrystalline and amorphous surfaces and on small metallic particles supported on insulators (i.e., practical catalysts). Basic studies to understand surface chemistry in the absence of long range order are essential to bridge the gap between the world of the "clean surface" chemist and the real world.

In the following section, we will give specific examples illustrating how extensions of existing surface measurement technology can increase our understanding in practical areas. We will also summarize some of the gaps in our understanding of surface processes and suggest possible future research areas. In the spirit of the cloudy crystal ball, we shall be completely unencumbered by present-day practical experimental and theoretical limitations.

Application of Existing Surface Measurement Technology to New Areas

Frequently, the factor that limits technological progress is the rate at which new measurement techniques are transferred from the laboratory of the basic scientist to the research-and-development laboratory. Auger spectroscopy was, for its first 5 years, primarily a basic research tool in surface physics and metallurgy. During the last 5 years, however, Auger spectrometers have found their way into the laboratories of technologists of all description and are being used for a variety of practical problems. Some potential application areas as well as limitations of surface measurement techniques are listed below; the discussion is representative, although not exhaustive. Some are new suggestions; some are probably already underway in research and development laboratories.

Catalysis provides a particularly fertile field for applications of surface-sensitive measurement techniques [12]. One of the most important developments in recent years in petroleum catalysis has been in the area of alloy and multicomponent catalysts. John Sinfelt used kinetic methods and was guided by a remarkable chemical intuition in developing a new reforming catalyst having higher activity and longer life than previously used platinum catalysts. The catalyst composition is proprietary, but it is known to consist of clusters of several (not normally alloying) metals supported on an insulating substrate. Knowledge of the atomic structure and chemical composition of the individual multimetallic clusters would be highly desirable in guiding future developments. Is there surface segregation of one component, or are the atoms intermingled? AES, ESCA, and

EXAFS are particularly suited for such studies, and by pushing the sensitivity, can be readily applied now to such studies of supported catalysts.

Both supported and unsupported catalysts are frequently doped with trace constituents (promoters) that enhance catalytic activity. The action of these promoters is not widely understood. Some are largely textural (i.e., they inhibit sintering and agglomeration). Others alter the electronic character of the catalysts by modifying either its bulk or surface properties. The surface-sensitive analytical methods, when combined with sputtering and depth profiling, have real potential for studying the role of promoters.

Catalyst poisoning, either by impurities in the feed stock or by self-poisoning due to reactant or product decomposition, is a vexing problem. Even a fraction of an adsorbed monolayer can be effective in "killing" a catalyst, as shown in Fig. 6. Frequently such traces of poisons defy detection by the most sensitive bulk analytical techniques. However, qualitative analysis of catalysis before and after poisoning using specific *surface-sensitive* methods can provide engineers with new techniques to guide them in reducing or eliminating the problem.

In the area of semiconductor device development and processing, it is axiomatic that modern devices depend for their operation on the properties of microscopically thin layers of silicon, oxides, and their interfaces [13]. Demands on the performance of such devices, including long-term reliability and radiation hardness (resistance to damage by ionizing radiation), require knowledge and control of the chemical and physical nature of the compound layers and their interfaces. Surface analytical techniques having sensitivity and spatial resolution far exceeding those of traditional analytical techniques are required for such characterization.

The principal method used for fabricating semiconductor devices and integrated circuits (IC's) is planar silicon technology, first developed to the stage where 10,000 MOS (metal-oxide-semiconductor) components can be manufactured on chip areas that only 15 years before could hold no more than a dozen components. The continuing trend toward larger scales of integration and microminiaturization has consequently increased the need for high-resolution quantitative measurements in extremely shallow multilayer device structures, resulting in a growing interest in surface analysis for silicon devices. It is in this area of technology that modern surface analysis methods have found the most immediate and enthusiastic applications. Areas of process control for IC devices in which such methods as AES, XPS, SIMS, ISS are finding increased utility include determination of dopant and/or impurity profiles, surface contamination, and interface characteristics, as well as IC failure analysis, an area intimately related to the above. Generally speaking, the electronics industry needs no prodding by authors like ourselves to respond rapidly to the latest developments in surface characterization techniques.

Another area that has been little explored using modern surface spectroscopic tools concerns the environmental stability of materials. It has been known for about 100 years that the physical and chemical state of the surface layers of a component can markedly affect its strength and reliability. For example, a KCl crystal is normally brittle and fractures readily when exposed to a bending stress in air. In contrast, it can be bent easily into a "U" shape under water. Normally ductile zinc becomes quite brittle when coated with mercuric nitrate solution. Much effort has been devoted to minimizing the detrimental effects of corrosive environments on mechanical properties. The goal of this work has been *prevention* of premature mechanical failure by controlling thermal treatment or operating environment so as to reduce a solid's ability to fracture. As noted by A. R. C. Westwood and John Mills [14], however, relatively little scientific attention has been devoted to improving the efficiency of industrially important processes dependent on fracturing (e.g., machining, grinding, drilling) by developing means of *facilitating* the fracture processes involved. Pursuing this line of endeavor, it has been found, for example, that the drilling rate of a diamond bit through gray granite can be more than doubled by using certain n-alcohols rather than water as cutting fluids.

The detailed physical mechanisms of these chemomechanical processes are not understood, and such studies provide exciting fields for both surface chemists and solid-state scientists. Inves-

tigations of the influence of adsorption on mechanical properties must necessarily be predicted by the questions: what is the chemical composition of the unperturbed surface, and what is the nature of the adsorbed species? XPS can provide elemental analysis of the surface region, and can (through studies of chemical shifts) indicate the valence state of those species. Moreover, it is eminently suitable for the examination of insulating substrates. AES and SIMS, when coupled with depth profiling techniques, can reveal the variation of composition as a function of depth in the solid, indicating the range over which chemomechanical surface processes act. An understanding of the atomistics of these processes can have wide-ranging effects in improving the efficiency of a host of practical mining and machining operations for both metals and nonmetals.

A major limitation of all surface measurement technology is the fact that the newly developed techniques need to be placed on a firmer quantitative basis [15]. For AES and XPS, questions of electron escape depth, the effect of surface roughness, cross sections for electron and photon excitation of surface atoms, and the influence of electron energy analyzer design should be investigated with the goal of establishing quantitative measurement capability. The inevitable concentration gradients at surfaces containing adsorbed layers must be adequately characterized if the objective of quantitative surface analysis is ever to be achieved. (Indeed, there are skeptics who doubt that this is an attainable goal for AES and XPS.)

Ion scattering spectroscopy (ISS) and secondary ion mass spectroscopy (SIMS) are, in principle, sensitive to only the topmost atomic layer. However, quantitative surface composition analysis using these methods may be hindered by sputtering damage to the surface. In addition, ISS is not able to resolve high atomic number species. SIMS is one of the most sensitive of the depth profiling methods, but it suffers from orders of magnitude variation in sensitivity from one element to the next. In addition, the sensitivity for a single element may vary by orders of magnitude depending on its chemical bonding and on the matrix in which it exists. The factors influencing ion yield and neutralization rates have not yet been adequately characterized. A method developed by Eric Kay and John Coburn for studying neutral species sputtered from compound surfaces and subsequently ionized in the glow discharge has revealed that the yield of sputtered neutral *molecules* may exceed that of sputtered neutral atoms. It is clear that atom, ion, and molecule sputtering by energetic ions and neutrals is an area in which basic problems with direct relevance to surface analysis need exploration.

Another frequently overlooked factor that limits the utility of methods based on the use of electron beams for analysis of compound surfaces is the perturbing effect of the electron beam [16]. The tendency, at present, to develop more highly focused electron beams for scanning electron microscopy (SEM) and scanning Auger microscopy (SAM) creates new problems for quantitative surface analysis. (Electron beams having a 0.5-mm spot size on the sample surface are commercially available in SAM systems.) For adequate signal-to-noise ratio, electron beam current density must *increase* as beam size *decreases*. This results in a dramatically increased probability of electron beam-induced damage to small particles, oxides, and adsorbed layers. On the one hand, the beam can cause enough of an increase in local surface temperature to promote interdiffusion or even melting of the surface layer. On the other hand, electronic excitation of the surface region can result in selective desorption of surface atoms, cracking of adsorbed hydrocarbons, enhanced adsorption and reaction by gaseous impurities, and even microscopic topographical changes. A major effort should be made to minimize beam damage to surfaces by increasing detector efficiency to allow much lower beam current densities to be used in all surface chemical and topographical characterization methods using focused electron beams, including AES and high-resolution electron microscopy.

In summary, we note that the ideal quantitative surface analysis probe has not yet left the designing boards; it employs a nonperturbing beam providing single atom sensitivity and spatial resolution at the angstrom level!

New Horizons in Surface Chemistry

A number of areas in surface chemistry offer frontiers of opportunity for advancing our knowl-

edge base. In many cases, the opportunity for intellectual pursuit is enhanced by a significant technological impact made possible by a fundamental breakthrough. Consider, for example, the field of catalysis. It is estimated that catalysis is currently involved directly or indirectly in the production of approximately $100 billion of the Nation's annual gross national product [17]. The vast importance of catalysis extends far beyond the chemical and petrochemical industry to areas of environmental protection, to critical roles in our future conversion of fossil fuels to gas and liquid synthetic fuels, and to areas of electrochemical power storage and generation. It is astounding that an area of such vast economic and social importance is so little understood at the fundamental level. Listed below are selected areas in the field of surface chemistry that are thought to deserve scientific attention. Many of the objectives cited cannot now be met with existing experimental techniques or theories.

- Although active sites on insulator surfaces have been identified spectroscopically, the characterization of active sites on *metallic* catalysts remains one of the major unsolved problems in surface chemistry. We need to have experimental techniques that can measure active site densities and characterize these sites geometrically and electronically. Experiments should be able to correlate active site density with overall catalytic rates. Theoretical developments should closely relate to the experimental results and should be concerned with the influence of site geometry and site electronic character on the catalytic reaction. This will necessarily involve a knowledge of the nature of the adsorbed intermediates on the site and their mechanistic involvement in the catalytic reaction.
- A search for a major effect of surface atomic geometry on catalytic reaction rates should be initiated. We need to find a system that exhibits orders of magnitude difference in catalytic reaction rate on different single crystal planes. Studies of this system involving many different crystal planes may then lead to better understanding of the influence of geometric factors on catalytic activity. If major geometric effects are not found in a thorough search of a number of reactions on single crystals, then it may be possible to lay to rest theories that attribute to surface geometry a major role in determining catalytic activity.
- The basic question of why the d-metals are good catalysts should be examined theoretically. To do this it will be necessary for the theoretician to know the identity of the activated complex responsible for the slow step in the reaction. Theoretical calculations should be aimed at understanding the influence of the d-electrons and orbitals on the chemical reaction. A suitable ultimate objective would be to devise a *working* theoretical picture that would allow the catalytic chemist to electronically tailor-make superior catalysts by alloying techniques. This will require major refinements in the ability of electronic theory to calculate total system energies, since reaction rates and routes are often determined by energy differences of fractions of an electron volt.
- The reasons for catalytic specificity should be studied from a very fundamental viewpoint. Why does a catalytic reaction such as the $CO + H_2$ reaction choose to occur along a specific pathway to yield particular products? Why does changing the catalyst (from one transition metal to another) sometimes result in the selection of a new pathway? If the electronic and geometrical factors responsible for such catalytic selectivity were really understood at a fundamental chemical physics level, then it might be possible to tailor-make catalysts using these principles.
- New physical chemical techniques should be applied to the study of catalytic reaction mechanisms on well-characterized (free from impurities, structurally defined) surfaces. These techniques should seek to answer the questions:

1. What are the elementary steps involved in the reaction?
2. Which step(s) impede the reaction due to activation barriers?

3. What is the structural and electronic involvement of the steady-state intermediates with the catalyst?

4. Does the chemistry of the chemisorbed catalytic species resemble that of identical ligands in organometallic compounds or is bonding influenced significantly by collective properties of the substrate?

5. Can experience with the modification of the properties of organometallic compounds using various substituents be transferred to catalytic chemistry?

- The influence of catalytic promoters and poisons should be studied experimentally to determine the mode of their operation at the atomic level. Do poisons and promoters act in a local fashion, or are we dealing with effects having a longer range? Is it possible to experimentally discover "antidotes" for catalyst poisons that will enhance the life of catalysts? Once enough good experimental data has been obtained that we are able to see systematic effects, theoretical efforts should be directed toward understanding the influence of poisons and promoters at the geometric and electronic level.
- Tunable ultraviolet and infrared sources should be employed to produce specific electronic or vibrational excitation in molecules causing the onset of specific reactions with surfaces. Detailed information about the nature of the activated species in catalysis could be obtained in this manner. In addition, excitation in this manner might allow the invention of new reaction channels of importance in synthesis (photocatalysis). The use of specific electronic laser excitation coupled with surface separation processes such as surface ionization should be investigated for potential use in energy-efficient uranium isotope separation processes.
- Spectroscopic surface measurement methods involving enhanced sensitivity should be continually encouraged. At present, various analytical methods display sensitivities ranging from a fraction of a percent of a monolayer (Auger spectroscopy, XPS) to methods which are sensitive to about 10^{-2}% of a monolayer under ideal conditions (SIMS). The development of new highly sensitive surface measurement techniques such as inelastic electron energy loss and electron tunneling spectroscopy (sensitivity ≈ 0.1% of a monolayer) and ^{13}C-NMR spectroscopy should continue to be encouraged, particularly if important structural information about surface bonding is measured.
- Research on surface techniques involving angular measurements should be encouraged as new tools for surface structural determination. Present techniques known to involve anisotropic emission of charged particles from surfaces include ultraviolet and X-ray photoelectron spectroscopy, Auger electron spectroscopy, and electron stimulated desorption of positive ions. It is anticipated that other forms of surface measurement techniques (such as photodesorption and optical absorption spectroscopies) will probably also exhibit anisotropies. These techniques may offer an opportunity for the measurement of short-range order or structure. This information is of critical importance in the characterization of bonding to surface sites. In summation, it is necessary to devise new methods that can tell us directly exactly where adsorbate atoms are located on a well-defined substrate crystal lattice.
- Research directed at learning the systematics of surface chemistry is likely to supply the type of experimental data of most use in the formulation of unifying theories. We need to know in a systematic way how the energetics of adsorption vary with substrate crystal structure. We also need to determine the influence of crystal structure and surface and bulk electronic properties on the electronic and vibrational spectrum of adsorbed species. How does the change of coordination number of a surface atom affect its bonding properties in forming the chemisorption bond?
- Modern surface measurement methods should be extended to the study of electrochemical surfaces. In particular, the carbonaceous layers present on the fuel anode in fuel cells should be characterized geometrically and electronically. The oxygen elec-

trode should be similarly characterized with the object of improving electrode efficiency. Studies of electrode poisoning would be useful in increasing lifetime and improving the efficiency of electrochemical energy-conversion devices.

- A basic understanding of photocatalysis is needed. In what manner does a catalyst reduce the photon energy required to cause a photochemical reaction? Does photon interaction occur with the catalyst or by interaction with a chemical bond, weakened in its interaction with the catalyst? Further knowledge in this field may lead to unique synthetic methods for production of new compounds, as well as possibly the use of catalysts in the harnessing of sunlight as a source of power.
- The reaction of steam with carbonaceous surfaces to yield $H_2 + CO$ should be exhaustively studied, since it will be the primary step in coal gasification, a major new source of energy. What is the influence of small quantities of inorganic substances on the rate of the reaction? How does the crystalline and chemical form of the carbon influence the efficiency of the reaction? Can methods involving catalysts be devised to reduce the extreme conditions of temperature and pressure necessary for coal gasification?

EPILOGUE

The history of surface chemistry has been an exciting sequence of scientific discovery, beginning in the early days of Langmuir and Taylor and extending through the development of the Brunauer-Emmett-Teller theory of multilayer adsorption to the present application of modern spectroscopic and diffraction techniques and quantum mechanical theories. One cannot help being impressed by the array of concepts and methods evolved. The pace of events in this field is still quickening, and many opportunities exist for significant contributions. It should be emphasized that in many cases our knowledge has been generated or improved because of scientific curiosity rather than explicit technological need. It is often true, however, that a technological development such as a new measurement technique is a major factor in opening new horizons for scientific discovery.

It seems reasonable to conclude that the support of research in the field of surface chemistry should continue to be imaginative, with emphasis on both technological benefits and on improvement of our knowledge for its own sake. The selection of the potentially most significant areas of scientific research in surface chemistry remains a difficult and yet most rewarding task for workers in the field.

REFERENCES

1. R. L. Park, "Inner Shell Spectroscopy," *Phys. Today* **28** (Apr. 1975).
2. P. J. Estrup "The Geometry of Surface Layers," *Phys. Today* **28** (Apr. 1975).
3. J. C. Tracy and P. W. Palmberg, *J. Chem. Phys.* **51**, 4852 (1969).
4. C. B. Duke, "What We Do Not Know About Surface Structure and Bonding," *Mater. Sci. Engr.* (in press).
5. E. A. Stern, *Sci. Amer.* **234**, 96 (1976).
6. L. H. Little, *Infrared Spectra of Adsorbed Species*, Academic Press, London, 1966.
7. J. E. Demuth and D. E. Eastman, *Phys. Rev. Lett.* **32** 1123 (1974).
8. D. E. Eastman and J. E. Demuth, *Japan J. Appl. Phys.*, Suppl. 2, Pt. 2, p. 827 (1974).
9. J. H. Sinfelt, *Catal. Rev.* **9**, 147 (1974).
10. James S. Murday, "Review of Surface Physics," NRL Memorandum Report 3062, May 1975.
11. H. P. Bonzel and R. Ku, *Surface Sci.* **33**, 91 (1972).
12. J. T. Yates, Jr., *Chem. Engr. News*, p. 19, Aug. 26, 1974.
13. A. G. Lieberman, ed., *Semiconductor Measurement Technology: ARPA/NBS Workshop IV. Surface Analysis for Silicon Devices*, NBS Spec. Publ. 400-23, Mar. 1976.
14. A. R. C. Westwood and J. J. Mills, MML Tech. Rep. 75-39 C, Martin Marietta Corp., Baltimore, Md. 21227, Oct. 1975.
15. P. W. Palmberg, *J. Vac. Sci. Technol.* **13**, 214 (1976).
16. T. E. Madey and J. T. Yates, Jr., *J. Vac. Sci. Technol.* **8**, 525 (1974).
17. V. Haensel and R. L. Burwell, *Sci. Amer.* **225**, 46 (1971).

S.N.B. Murthy, Professor of Mechanical Engineering, is currently the Director of ONR Project SQUID. He was educated at the Indian Institute of Science and the Imperial College of Science and Technology in London. After gaining several years of experience in the gas turbine industry, Dr. Murthy has taught in India, the United Kingdom, Canada, and the United States. He is the author of more than 50 research publications and three books in the fields of gas dynamics, energy generation, and propulsion.



FUTURE OF AIRBREATHING PROPULSION

S.N.B. Murthy

School of Mechanical Engineering
Purdue University
West Lafayette, Ind.

Airbreathing propulsion has come to fulfill a vital need in human activities, in transportation, and in defense and therefore has a developing future so long as air can be used effectively in the engine without affecting the quality and existence of life on Earth and so long as the required energy can be found and used efficiently. There is considerable scope for advances in aeronautics, although the rate of growth may be determined in the future by more complicated economic and political factors than in the past.

In view of the limited resources available for the development of any one technology, the general problem in propulsion technology becomes the optimal use of our resources: energy, materials, manpower, and airspace itself. To gain national backing for the technological opportunities that are obviously available in this field, it has become increasingly necessary to prove that the technology base exists to justify the claim of well-balanced returns for resource investments in civil air transport and military needs. This means that not only should all development be based on the best scientific and engineering analysis but also that new technology should be introduced into this field with the highest national interest and public acceptance in mind.

Developments in airbreathing propulsion are both expensive and time-consuming. Historically, such developments have occurred through both evolutionary changes and "quantum breakthroughs." Basic research and the "learning" process in engineering are therefore clearly important in the ordering of priorities in this area. In any engineering product the incorporation of the results of research depend on the economics of the market. In the case of an aircraft, despite the fact that the product is designed to operate at peak performance, replacements are made more according to economic and strategic considerations than because of deterioration of the aircraft. It is therefore important to sustain a level of effort in research at the fundamental level. Progress can be made at that level in the broad context of problem areas, and such progress should be incorporated in the development and production of the product whenever the opportunity arises.

The central theme of this paper is a discussion of research and development needs in the technology of airbreathing propulsion. The outline selected for the discussion is as follows: special features of airbreathing propulsion technology; aeronautical propulsion development; and some research areas.

Considering the enormous extent of the field of aeropropulsion, it is unavoidable that one is selective in a review such as this. The selection is based here somewhat on areas in which there is personal interest and almost entirely on the developments in the United States. The airbreathing propulsion industry is well established in a number of countries, and there are important reasons to be ex-

tremely competitive in establishing superiority in this field. The U.S. research and industry communities are entirely alert to this factor and have maintained a global leadership.

SPECIAL FEATURES

The history of aeropropulsion in the past four decades has been one of continuous growth, and one can still see scope for further growth and improvement in both engineering and economic performance. The industry is showing no signs of "maturity"—a small rate of technological change of a basically frozen product. Substantial changes in technology can be foreseen on a long-term basis. However, such questions must take into account the special features of this technology and industry. Four that appear significant in the present context are as follows. (a) Airbreathing propulsion is part of the overall transportation system for civil and military use. (b) Economic and political considerations play a central role in determining the direction of development and production of civil and military vehicles. (c) Resource management determines the major impacts of research and development. (d) Defense procurement can be based in the ultimate only on overall strategic considerations.

The aeropropulsion business, second only to aerospace activity, has provided a continuous stimulus to research and development as well as made it a necessity for its own survival and progress to make a success of research and development. One would therefore think that in an industrial and military activity with such potential for advances, research and development would find assured support. In the past these resources have become available either because the market would accept any new engine or aircraft as it became available, because there was a commitment to certain goals, or because there was a generally accepted policy in the military that advances in technology provided invariably a superiority in defense capabilities.

The latter has attained an entirely new perspective in view of changing strategic considerations and concurrent developments in a number of military technologies. The advances in remotely piloted vehicles, "standoff" capabilities in different parts of the world, and the logic of balancing tactical, strategic, and defense capabilities bring in considerations that make the independent superiority of any one technology rather less significant. Nevertheless, advances in technology must not be confused with decisions of procurement. The case of the development of a bomber such as the B-1 in the United States may be pointed out in this connection. Current estimates for the development of a fleet of 244 bombers by 1985 is about $21.4 billion, not accounting for the cost of weapons (delivered and used for survival) or of the tanker fleet required to fuel the bombers on missions of ranges longer than 6000 mi (9600 km). The assessment for such a bomber will have to rest on overall defense strategy much more than on its cost or its effectiveness as a weapon. In civil aeronautics, there has been some progress in identifying certain technological goals in regard to efficiency, noise control, and scale-speed-range possibilities. However, such technological goals have not yet been translated into specific aircraft engine requirements. One therefore has again to secure resources for research from considerations of establishing a strong, rational technology base for derived, refitted, or new aircraft whenever they may come into being.

In that connection, it is significant to emphasize how developments in aeropropulsion in the past have been initiated and sustained by research-and-development projects undertaken to meet military requirements. The funding for development is largely a function of the various purposes, sophistication, and reliability that are demanded in a system. It therefore can change from year to year. In the United States, the average funding for aeronautical development on a yearly basis has been as follows during the past 10 years in terms of 1973 dollars:

Government funding on development (defense)	$2.7 billion
Industry funding on development	$750 million
Government funding on research	$750 million
Government funding on development (nondefense)	$500 million
Government funding on research (nondefense)	$150 million

The government funding on defense development has varied by about $300-500 million from year to year.

Most research undertaken for defense needs has relevance to the entire aeronautical industry but not to the same extent today as in the past. Even 20 years ago, there was almost direct exchange between military development and civil aeronautics, but this has declined today to the point that new agencies with specific missions are being suggested for advances in civil aeronautics. To some extent, this is due to the rather different emphasis in civil aeronautics in recent times; it is also due to the uncertainties in defense requirements and the enormous expenditures in time and money involved in undertaking new military development. Nevertheless, military and civilian agencies demonstrate continuously their ability to coordinate their efforts in every problem where common technological goals can be established.

While basic research has been supported from external sources in various organizations, the aeropropulsion industry has also been encouraged by sponsorship through allocations for business expenses to undertake independent research in order to permit immediate utilization of innovative talent in the industry. Such independent research in the industry also fosters interindustry competition in meeting specific procurement requests and in developing a broad-based capability in this field. The latter is especially important in developing confidence in the industry for undertaking exploratory and development activities.

The ultimate need in all progress is scientific talent. It is extremely important to see that studies in aeropropulsion attract young talent. Adequate support for research is one way of achieving that objective. On the other hand, there is considerable need in the universities to orient their study programs to instill in the students the broad methods of rational analysis and experimentation for creative design and management in the aeropropulsion industry, which certainly presents many interesting challenges.

Aeropropulsion and Transportation

An engine is an essential feature of most aeropropulsion systems. It is a system in itself consisting of a carefully matched set of components. The basis of the engine as a system is its thermodynamic cycle, and a variety of engines can be derived from each thermodynamic cycle with variations in geometry, airflow path, combustion of fuel, and heat transfer. Air, which is the natural propulsive fluid, can be used in an engine to generate energy in combination with a chemical fuel or as a medium for the transfer of energy from an energy generator to a thrust generator, as in nuclear systems.

In all cases, major components of the engine which perform different processes of a thermodynamic cycle, components which perform various mechanical functions and control of the engine are, of course, important. If one considers the overall performance of an engine, the object of all improvements is to reduce fuel consumption, weight, pollution of the atmosphere, noise, and engine life-cyle maintenance cost and to increase component life, operational simplicity, and system reliability. In the use of airbreathing engines for propulsion, the only constraint is the availability of air. However, while the nature of the exhaust products and some of the noise characteristics are controllable by engine processes, the lift, drag, vibration, and much of the noise depend critically on the engine-vehicle integration.

The engine-vehicle propulsion system itself should be looked at as part of what may be called the transport system, whether we are considering civilian transport, tactical aircraft, or weapons. One then has to take into account the nature of the mission, the conditions at the origin and end of the mission, and the integration of the mission and operation of one unit with all other units involved in the overall objective of transport or defense. A simple example is the integration of a group of flights with the overall transportation of people from door to door. Vehicles with different ranges, speeds, and payloads offer different kinds of challenges in different missions.

One can therefore summarize the prospects for airbreathing engine propulsion in terms of the following: (a) ability of the engine to accept a variety of fuels and the reduction in fuel consumption and emissions over a mission; (b) reduction in the undesirable characteristics of noise over prescribed areas; (c) improvements in the weight, performance, reliability, and overall life of components of a controlled engine; (d) integration of

the engine system with the propulsive force generator and the vehicle; (e) coordination of the vehicle system operation with the transportation or defense environment in which it is expected to perform a mission; and (f) assurance of predictable reliability and safety.

It is of interest to note the substantial improvements in safety achieved over the years in air transport. The number of fatalities per 100 million passenger miles in civil transport has declined from 2.4 in 1940 to 0.1 today. There is a corresponding improvement in military aircraft safety: the number of major accidents during the first 100 thousand flying hours is about 25 today, compared to 75 in 1952.

Economics in aeropropulsion—Such considerations, however, cannot be based entirely on the technical merits of machine efficiency, reliability, and operational simplicity. Economic factors enter deeply into development and in fact dominate the decision making process even in defense requirements. Resources are limited for any one task in a nation and the cost of any product cannot grow faster than the gross national product. Meanwhile, many military aircraft and systems take on the character of capital goods. The air combat capability is a function of unit effectiveness and number of weapons and therefore of cost. The cost of civil transport, which continues to rise, is generally dictated by scale—scale of a single vehicle and of a fleet—and is therefore subject to the considerations of the service the public desires and the cost the market will bear.

One can obtain some idea of the problems involved in determining developments in this field by examining the concept of efficiency of a propulsion system. The efficiency of a propulsion system can be expressed in several different ways. One of the more common measures of efficiency is the Breguét range for an airplane with known values of chemical, propulsive, aerodynamic, and structural efficiences. By relating the propulsive efficiency and lift-drag ratio to the flight Mach number, one can obtain a rough guide to the ranges of application for various aircraft: classical aircraft at subsonic speeds for short ranges, slender aircraft at supersonic speeds over long hauls, and hydrogen-fueled hypersonic aircraft for longer global-scale ranges. Another useful definition of efficiency is the ratio of the actual range obtained with a mass of fuel in a given aircraft over a certain mission to the ideal range that could be obtained with the same mass of fuel on the basis of its calorific value (4300 km for kerosene and 11 800 km for hydrogen). It is possible then to obtain an overall pattern of fuel usage or requirement for different aircraft on different flightpaths or missions.

Three other factors that also enter into the efficiency of the propulsion system are the noise footprint of the airplane, pollution of the atmosphere, and contrail formation, strength, and composition. These are basic considerations in the development of future civil air transport.

Other measures of efficiency can be obtained on the basis of economic considerations for example: cost of developing an aircraft to the point of flyaway; cost of production aircraft; direct operating cost; cost per ton-mile, modified in various ways for passengers, cargo, and ordnance; available seat-miles per hour; seat-miles per gallon of fuel; seat-miles per dollar of overall cost of aircraft; and life-cycle maintenance cost.

A skilled analyst can prove the superiority of almost any system by a selected combination of the foregoing criteria for economic performance. However, the impact of such criteria is very real in the growth of the aircraft propulsion industry. The ultimate significance of such analyses must be assessed on the basis of several other factors: impact of rational analysis, measurement, and testing in the evolution of a product in this technology; detail to which the desired product is specified; risk of performance failure, resource curtailment, market uncertainty and, in the case of defense needs, advances in related areas and adversary moves; and nature and extent of large-scale regulation of development and investment.

The interaction of various criteria in the final economic analysis can be seen in Figures 1 and 2.

Resources for Aeropropulsion Technology

The status of any economic effort is determined by the availability of resources. In the past 40 years the aeropropulsion industry did not have to contend with the problem of resources as much as with establishing itself as a dependable technology with continuous efforts in improving reliability,

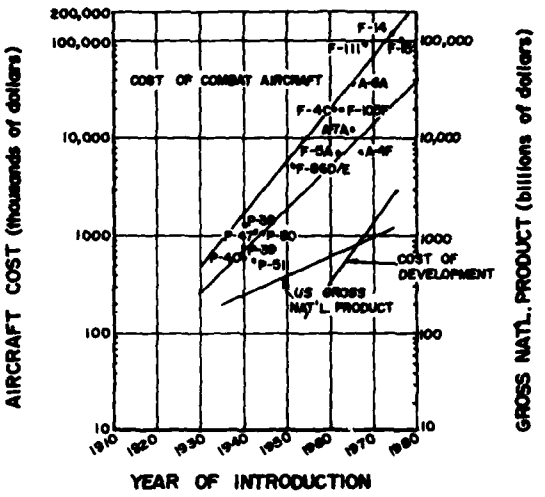


Figure 1—U.S. combat aircraft development: cost of aircraft, cost of development, and gross national product

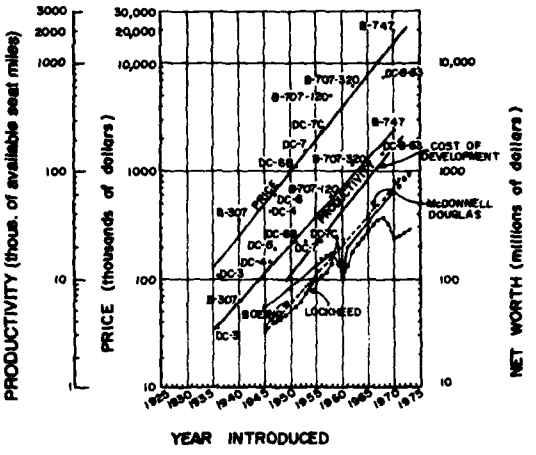


Figure 2—U.S. transport aircraft development: cost of aircraft, cost of development, and gross national product

safety, economy, speed, and range. Military requirements have demanded in the past the introduction of every foreseeable technological advance in the product from the point of view of *meeting* various effectiveness criteria and assessments of threats and challenges.

In the past few years, there has come about a changing attitude to technology, to resource management, and to needs in both civilian and military markets. There has also arisen a point of view that barter at the international level for natural and industrial products should be based on equal opportunities for all nations. The question here is not whether these are temporary *anxieties* but rather what the intensity and implications of such attitudes are towards expansion as a way of meeting demands. In that spirit, aeropropulsion technology is concerned with the availability of the following resources: air, fuel, materials, and support for research and development.

Air and Airspace—In aeropropulsion airspace must include the surface of the earth and should be considered both globally and locally from the points of view of (a) chemical pollution, (b) noise, and (c) density of traffic and overall transportation management. Each nation claims sovereignty over its airspace, and this has obvious implications in international affairs.

The principal factors in airspace utilization from the point of view of pollution are altitude, speed, range, and flightpath of aircraft, and location of airports. They should then be related to the types of fuels available in different geographical locations and the dynamics of atmospheric motions at different altitudes, including the surface of the Earth. The problem of pollutant dispersion and effects can be solved only by understanding the interaction between the engine emissions and the macroscale and microscale air motions. Such considerations also draw attention to the uncertainties of where pollution can become substantial locally, relative to the flight of aircraft and the location of airports. This problem can arise in respect to military training facilities also.

The uncertainties in modeling atmospheric motions in regard to pollution become particularly clear when one examines the recent anxiety in the United States over the depletion of ozone in the stratosphere due to water vapor and NO_x emissions from supersonic flight at those altitudes. Ozone is the primary absorber of solar ultraviolet radiation, and its depletion would increase the radiation received on Earth. In addition, the formation of clouds and large contrails can change the balance of radiation both from the Sun and the Earth. If one assumed that the emissions from a fleet of supersonic aircraft and hypersonic aircraft would remain stratified in various atmospheric

layers with residence times of the order of months and with some organized growth of the wake, one can show a substantial loss of ozone in the stratosphere. However, by including a more detailed consideration of the mass and heat transport processes, one can also show the possibility of a relatively fast subsidence of a contrail and hence the lack of the necessary residence time for substantial reductions in ozone.

The other two considerations in the use of airspace, (a) noise and (b) traffic and transport management, are intimately related to the use of community land space. One then must take into account such semiqualitative factors as psychoacoustical reactions of the community, passengers' attitudes toward transportation from door to door, and the relationship of safety and workload for airplane operators and air traffic controllers. The latter is important in introducing noise-abatement procedures and associated airplane equipment.

The DOT-NASA Office of Noise Abatement has undertaken many detailed analyses of noise impact in the vicinity of airports. The Noise Exposure Factor (NEF) 30 contour in at least six of the U.S. airports with mixed fleets encloses an area of about 200 km². The 30-NEF contour corresponds to the 90-EPNdB noise level of a typical aircraft operation of 600 flights per day. The quietest aircraft in the current jet transport fleet (the three-engine wide-bodied aircraft) has a 90-EPNdB footprint of 20 km². In the face of this, NASA has set for itself the following goals: (a) noise footprint for wide-bodied aircraft of about 2 km²; (b) noise footprint in the Advanced Transport Technology Program for high-performance commercial transport aircraft of 5 km²; (c) noise level of 95 EPNdB on a 150-m sideline and a noise footprint area of about 2 km²at 90 EPNdB for a 150-passenger powered lift aircraft. The NASA Quiet Engine Program (QEP) and the Quiet Clean Short Haul Experimental Engine (QCSEE) Program (Figure 3) are directed toward the attainment of such goals.

Noise abatement solutions have taken three directions: modifications to aircraft landing and takeoff procedures, design of parts of the engine and engine locators that would be incorporated into current aircraft, and design of new equipment and airports for future use. The two-segment approach, the decelarating approach, and the microwave landing system for a flexible and, where necessary, curved approach provide various operational procedures for changing the noise footprint. Similarly, during takeoff the power cutback procedure is a means of controlling the overall noise footprints. Ultimately, advanced avionics and active controls will have to provide the required capability for simple operation with the necessary cockpit and airport control displays. The NASA terminal configured vehicle program is expected to assist greatly in the development of advanced displays, autonavigation and guidance systems, and digital flight-control systems. The refitting of nacelles, fans, and nozzles (up to 20-EPNdB) reduction) is a more complicated and expensive solution for noise reduction, but it can be shown to be justifiable on various grounds, including possible fuel economy.

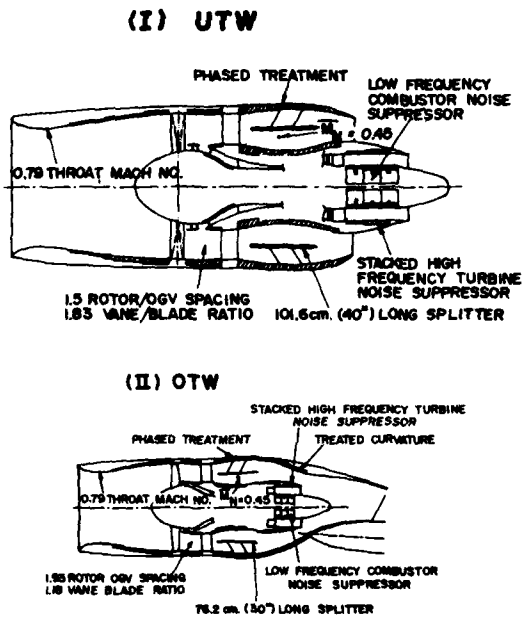


Figure 3—QCSEE configuration: (I) Under-wing; (II) over-wing.

Noise certification plans are being continuously revised (for example, the FAR 36 with the new NPRM) based on foreseeable advances in technology that can be introduced after a variety of considerations. The FAR 36 noise measuring criteria include flyover, approach, and lateral (sideline) effect in terms of EPNdB, the perceived

noise level corrected for the annoyance due to discrete pure tones and the time duration of aircraft noise signal. In meeting the existing rules and their possible modifications, one must take into account the interconnections among basic engine cycle, exhaust gas temperature, takeoff gross weight, scale of the airplane, and direct operating cost.

From the points of view both of noise abatement and of airport and airspace congestion, it may be necessary in the long range to examine the question of airport location and the separation of different kinds of aircraft in different airports of large metropolitan areas. Two areas in which developments are expected to enlarge current civil air transportation productivity are short-haul aircraft and small business aircraft. The latter must be operated on an unscheduled basis and will require both special traffic control procedures and noise reduction. The latter may not be feasible with changes in landing and takeoff procedures alone, and V/STOL aircraft engine noise itself may have to be conditioned to take advantage of the attenuation of high-frequency noise in the atmosphere. The airframe noise will probably set the limit in these (and in fact most) aircraft for the lowest achievable noise level (Figure 4).

Fuel—Aviation is based at present on the use of hydrocarbon fuels derived from natural oil, and the use of alternative chemical fuels, (for example, cryogenic fuels) is unlikely to come about without considerable advances in supply and handling. The principal factors in the use of chemical fuels are: source of fuels, heat of combustion, storability and handling, sensitivity of the engine and its installation in aircraft to fuel composition and properties, emissions in the combustion products, and safety. The fuel quality desirable in aeropropulsion depends on a combination of a variety of properties, the influence of which cannot be separated from the flight mission and the details of fuel and air management in the vehicle (for example, thermal stability, volatility and vapor pressure, freezing point, density, and flammability and explosivity).

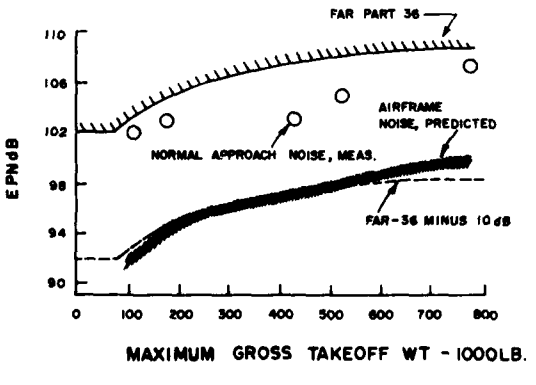


Figure 4—Significance of airframe noise

The energy content and density of fuels are directly related to the range of aircraft. Departure from conventional hydrocarbons (limiting energy around 23 000 C H U/kg) has not been easy for gas turbine use. On the other hand, liquid hydrocarbons suitable for gas turbines can vary in density over a range of 20%. Some airline specifications therefore favor kerosene over JP-4. However, the principal concerns in advanced fuels are thermal stability and heat-sink characteristics.

At present aviation accounts for 4% of the total world output of oil-based fuels and therefore for about 1.5% of the total fossil fuels. Petroleum products while supplying about 45% of the total energy demand, account for 95% of the transportation energy. Aviation probably is using up to 12.5% of the total energy requirement for transportation, which itself is of the order of 25% of total energy consumption. Both AV gas and jet fuel are included in this, although the AV gas consumption is fairly steady at about 6 million gal./day. Civil aircraft demand for kerosene has increased steadily and is nearly three times the demand a decade ago, while the military requirement has doubled in the same period. Some studies have shown that the demand for aviation-type fuels will probably double in the next 30 years, by a rather conservative outlook on the growth of this transportation market.

The grades of gas turbine fuel available for military and civil aircraft are given in Table 1. Several alternative fuels have been examined and some of them are listed in Table 2. The availability and properties of these fuels raise several questions: (a) development of production methods and availability and cost prediction from different sources and in different geographical locations; (b) techniques for supplying fuels to terminals and aircraft; and (c) design for safe, reliable, and eco-

Table 1

Aircraft Fuels

Aviation Gasoline			
100/130 Grade		1937	
114/145 Grade		1945	
Jet Fuels (Military application)			
JP-4	Wide-cut	1950	Air Force
JP-5	High Flash Kerosene	1950	Navy
JP-7	High Flash Kerosene	1965	Special Application
JF-8	Low-Volatility Kerosene	1968	Special Application
JP-9	High-Density Hydrocarbon Blend	1974	Special Application
Jet Fuels (Civil applications)			
Type A	Kerosene	1958	
Type A-1	Kerosene	1958	
Type B	Wide Cut	1958	

nomically feasible handling. Other questions arise in respect to flight equipment design. However, the logistics of production, supply, and handling are the central issue.

Liquid methane can become a useful fuel in any air transport system where its higher energy content and heat-sink capacity can be shown to mitigate the effects of its low density. If the fuel handling problems can be solved, it appears one can define missions where volume limitations can be balanced against takeoff and landing performance for a given wing loading and payload, for example, supersonic cargo aircraft for speeds above Mach 3, both in terms of payload and direct operating cost. Liquid methane (natural gas) has also been examined in various designs, including the Boeing Arctic Resources Carrier, and will probably prove successful for such missions.

Between liquid methane and liquid hydrogen, the latter can be shown to be superior from several points of view: cost, heat of reaction, combustion products, and other properties. There is also considerably greater experience with hydrogen both in handling (for example, NASA used 75 000 tons of liquid hydrogen each year in the Apollo program) and in gas turbine combustors (since the mid-1950s). Liquid hydrogen introduces new considerations in regard to mission, speed, and range. The cruising altitude and range can both be increased compared to other fuels. However, it appears that large subsonic and hypersonic aircraft (requiring high cooling capacity) are more likely candidates for hydrogen than small supersonic aircraft in view of the bulkiness of the fuel. However, considerably more knowledge is required concerning tankage, insulation, fuel management, and certain aspects of safety in handling, as well as the production of hydrogen. The cost of producing liquid hydrogen may remain substantially larger than that of JP fuels in the next two decades even if coal gasification and decomposition of water through electrolysis or thermal cracking (requiring other advances in nuclear or solar energy) become economically feasible.

Developments in such alternative basic energy sources will also permit "conventional" hydrocarbon fuels to be synthesized from organic sources, limestone, atmospheric carbon dioxide,

Table 2

Alternative Fuels

Fuel	*Heat of combustion (L)*		*Density*	*Boiling Point*	*Cost*
	(Btu/lb)	*(Btu/ft³)*	*(lb/ft³)*		*($/10⁶Btu)*
JP Synthetic (Jet A)	18 590	940 000	50.5	370°F-550°F (Liquid at Normal Temperature	1-3
Hydrogen LH_2	51 500	222 000	4.3	−423° F (Cryogenic)	2.50-8.50
Methane LCH_4	21 500	570 000	26.5	−259°F (Cryogenic)	1.50-3
Propane C_3H_8	19 940	720 000	36.1	−44°F (Low-Temperature or Compressed Gas)	0.75-2
Methanol	8 640	426 000	49.4	149°F (Liquid at Normal Temperature)	1-2
Boron (type B_5H_9)	50 000	1 188 000	39.6	137°F (Liquid at Normal Temperature)	100-300
JP from Coal	18 830	996 000	53.0	370°F-550°F (Liquid at Normal Temperature)	1.50-3

and hydrogen in water. Once again, there is no way of establishing the economics of such industries. The synthetic, conventional hydrocarbon fuels will of course have generally the same economic potential in terms of heat content as current hydrocarbon fuels.

There is substantial possibility, however, in the near term for synthetic fuels derived from coal, oil shale, and tar sands. Coal probably holds the best promise. There is of course wide variation in the composition and quality of raw material at different geographical locations. Several JP-type fuels have been synthesized and tested, for example, at the Naval Air Propulsion Test Center and the Wright Field Laboratory. It is important to recognize here that very extensive engine tests are necessary before laboratory samples can be accepted as satisfactory.

Among alternative fuels one should also note the possible use of slurries, for example, with boron, boron hydride, or possibly metallic hydrogen in the far future. The density and energy content of fuels can both be increased, but there arise other problems such as toxicity and deterioration of turbine blades. Some of these therefore may have to be looked upon as extreme concepts for further consideration. A more promising development is that of emulsified fuels—usually water dispersed in a conventional fuel—which seem to offer improvements both in performance and emissions.

One other possibility for aeropropulsion is the direct use of nuclear energy. There is probably sufficient technological data for the use of conventional nuclear energy in subsonic aircraft, including various aspects of safety such as crash-

worthiness and thermal failure. However, it appears that considerably more system-type studies are required, including consideration of advanced reactors, before one can formulate a mission for nuclear aircraft.

Materials—It seems unlikely that shortages will arise in the basic materials needed for aeropropulsion engines or vehicles. The total cost of materials is rather small in any aircraft, and a 100% rise in the cost of materials may only lead to a 15% rise in the cost of any aeropropulsion system. The introduction of composites in place of metals is based on weight and performance considerations and not on the unavailability of metals and alloys, although titanium, nickel, and copper will play a critical role. In connection with the latter, one should note (a) the impact of environmental protection considerations on the production of metals and (b) the leadtime and cost involved in the production of standard items out of those materials. Defense management is alert to the latter, but the aeropropulsion industry needs to be very strongly interested in national policy on supplies of standard items made of special materials.

The rapid progress of composites in the last decade compares favorably with progress in metallic materials introduced earlier. However, the use of composites continues to be limited because of lack of confidence and cost. In principle, NASA and the Department of Defense are solidly committed to the use of composites and support a variety of programs. Engine components such as fan blades, compressor blades, and frame sections are important potential areas of application for advanced composites. Currently composites represent some 2.5% of engine weight, and future projections indicate savings up to 30-35% in weight and 20-25% in cost. Among various requirements for increased use of composites are (a) improvements in fatigue characteristics, dominated by either the fiber or the matrix, and (b) advances in tooling technology for composites.

AEROPROPULSION DEVELOPMENT

Developments in aeropropulsion can be classified in various groups on the basis of speed of the vehicle (subsonic, supersonic, or hypersonic), range (short-haul or long-haul), lift generation (CTOL or V/STOL), and mission (military, cargo, or transportation). The aeropropulsion systems in those groups of course overlap to a considerable extent. Accordingly, we shall discuss developments under the following: subsonic aircraft, short-haul transport, supersonic transport, flight above Mach 3, air cargo systems, and some military developments. Some typical projected developments in aviation are given in Table 3.

Subsonic Airplane Propulsion

The past 30 years have seen about 30 significant passenger transport programs in the Western world. Most programs have been based on closely related or "derived" models. Considering engine types, the piston engines were replaced in 1959 by jet engines. The turboprops continued until 1962, when jet aircraft replaced them. The fan engines have since then attained supremacy. The technology of the jet engine and the swept wing has made possible a generation of narrow-body aircraft that has made a great contribution to air transportation, but they will have to be retrofitted, or replaced by the newer generation of wide-body aircraft incorporating high-bypass-ratio fan engines, in view of environmental considerations.

The wide-body fan-jets can be expected to dominate the large commercial subsonic aircraft market and possibly enter military service for airlift, as tankers, or as airborne command posts. In the future it is possible that there will be a need for a slightly smaller aircraft as well as one in the 60 thousand-lb-thrust size. The future developments in propulsion systems in this area therefore may be classified (a) component improvement in engines for production and derived aircraft to increase specific fuel consumption (SFC) and reduce noise and (b) new engines for new medium-range aircraft, narrow-body short-range aircraft, and long-range wide-body aircraft. Improved versions of the 40 000- to 50 000-lb-thrust engines, derived engines in the 25 000- to 30 000-lb-thrust class, and new engines in the 10- to 12-ton category are thus under development.

In the next 10 to 15 years the kinds of subsonic transport that may come into being in civil aviation may be grouped in various ways; for instance:

Table 3

Representative Projected Advances

Advance	*Year of Earliest Introduction*
Civil	
Derivative and growth versions of transport and general-aviation aircraft	1980
Efficient long-haul transports	1985
Large cargo transport	1995
Military	
Derivative transport/tanker aircraft	1985
Long-endurance reconnaissance and patrol aircraft	1985
Very large logistic transport	1985
Civil	
Efficient short-to-mid range R/STOL transport	1985
Medium-size utility/business rotorcraft	1990
Intercity VTOL aircraft or rotorcraft transport	1995
Military	
Long-range rotorcraft	1985
Subsonic V/STOL fighter aircraft	1985
Carrier-borne miltimission V/STOL aircraft	1990
Civil	
Derivative "Concorde II" based on near-term technology	1985
Advanced supersonic transport	1995
Military (tactical)	
Maneuvering missiles and RPVs	1985
V/STOL supersonic fighters	1990
Advanced weapons carriers	1990
Advanced fighter/bomber	1995

750-1000 passenger intercontinental transport, 300-500 passenger airbus, 150-200 passenger medium range STOL, 50-100 passenger short-range STOL, and advanced general aviation aircraft. In addition there may be some justification for developing a transonic transport for flying at Mach 1.15 over land, the highest Mach number possible without the appearance of a sonic boom. In the different classes of vehicles there is a slight shift in emphasis in the usual demands made on propulsion systems, but in all cases it is necessary to match the vehicle thrust over the full operating range (Figure 5). Thrust matching shows the speeds at which turboprop, turbofan, and an advanced "propellor fan" are useful. Advances in propulsion will come through improvements in the engine and the propulsor.

In general aviation, where initial cost is important, the piston engine has generally been preferred to the gas turbine. However, there are significant developments in small turbofans, geared and ungeared, and gas turbines may come into

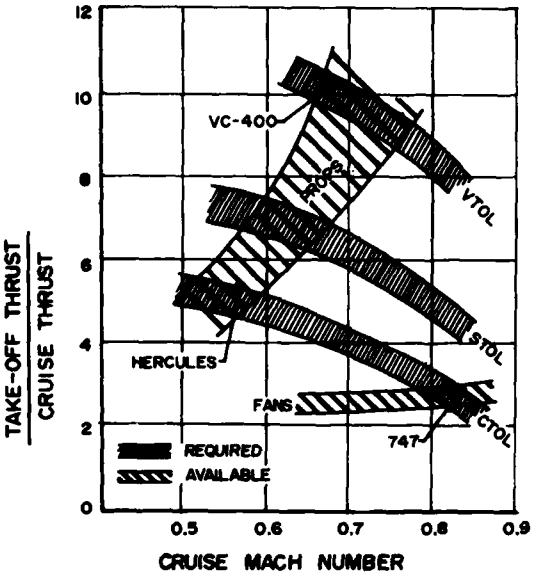


Figure 5—Thrust matching

wider use in general aviation. One example of engine/airframe interrelation that needs to be established is the possible application of the F107-WR-100 turbofan (specifically intended for the USAF/Boeing AGM-86A Air Launched Cruise Missile) in an airplane cruising at 300 n.mi./h at 6000 m with a gross weight of 1800 kg and requiring a runway of about 360 m at sea level.

The thrust-to-weight ratio, which has improved in the past 30 years by 50%, can probably be improved in Mach 0.8- 0.85 airplanes by another 75% when turbine inlet temperature is raised to 1550°C, overall pressure ratio to 30-40, and fan pressure ratio to 1-7; these are foreseeable improvements. Turbine cooling air may then have to be precooled with bypass air. If cooling air is at compressor delivery temperature, it is not easy to achieve stoichiometric turbine entry temperature. At cooling airflow of 2.5-3.0% of hot gasflow, conventional cooling effectiveness may vary between 0.4 and 0.7. In optimizing the engine cycle it is also useful to note that the value of specific thrust at which optimum SFC occurs rises with Mach number, and SFC becomes less sensitive to specific thrust variations as speed rises. Installed thrust and drag considerations become important as speed increases.

The propulsors, namely rotors, propellers, shrouded propellers, "Pro-Fans" and fans, have unique characteristics and are best suited to their own operating regimes. Turboprops and turbofans are the principal propulsors of interest in C/R/STOL systems for moderate- and long-range flights.

The modern turbofan uses the bypass concept to improve propulsive efficiency and hence overall efficiency. Currently 30% of input energy becomes available for propulsion in fan engines, and one of the principal incentives to advances in technology is to improve that figure. The bypass ratio being considered in advanced engines is 6-8:1. The large bypass ratios are possible because of reduced weight penalties in blading and thrust reversing (controllable pitch reversal). In a given propulsor, the bypass ratio is a function of fan pressure ratio and specific power. Each major increase in bypass ratio has produced improved takeoff thrust, and this has in turn yielded thrust-to-weight ratio improvements from about 4 to 6.

The turboprop, which may be said to have a bypass ratio of 35 to 70, has a propeller pressure ratio of 1.015 to 1.05. It has good noise and takeoff thrust characteristics, in addition to lowest cruise SFC (Figure 6). For short-range transport, where cruising speed is relatively unimportant, the turboprop continues to be a useful propulsor with high propulsive efficiency that could be combined with cruise speeds, altitudes, and reduced vibrational levels comparable to those of current jet-powered transports. Fuel economy and reduction in noxious emissions both can be improved by regeneratively heating the combustor inlet air using engine exhaust heat.

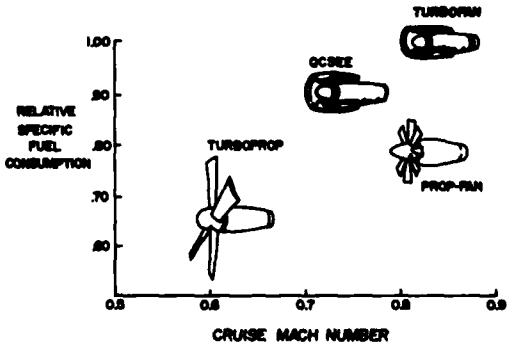


Figure 6—Specific fuel consumption of several engines

The Prop-Fan is a further advance in turboprop technology. It is a controllable pitch fan with a pressure ratio of 1.10-1.20, tip speed of 210-225 m/s, and 8-12 blades. It has a fast thrust response. At Mach 0.8, a predicted efficiency of 74% has been quoted for it, compared to 62% for the turbofan. It is currently undergoing wind-tunnel tests. Improvements in efficiency at high loading are expected to come from reduction of compressibility losses and recovery of swirl energy in the slipstream.

The improvements in thrust SFC (TSFC) arise from advances in component technology—performance and materials, the latter including cooling effectiveness. In the case of fans and ducted propellers, advances lead to reduction in airfoil and endwall losses. Shock losses can be reduced with controlled shock blading. In compressors, higher rotor speeds and loadings and adjustment of clearance between rotating and stationary parts to suit different operating conditions (active clearance control) will lead to improvements. In turbines, in addition to better sealing, lightweight blading is used to reduce the structural loading. The blading losses can be reduced by the use of laminar-flow airfoils. Material, cooling, and structural-mechanical improvements can be introduced throughout the engine. An important feature of SFC and thrust-to-weight ratio improvement is the reduction in direct operating cost.

Short-haul Transportation

Short-haul air service, generally under 500 mi (800 km), now constitutes about half of the air traffic and is expected to grow in the next two decades. High- and low-density populations present slightly different problems in the organization of short-haul transportation, but the basic need of this part of propulsion is the development of a V/STOL system that has the required technical and environmental-impact merits.

NASA and the Department of Defense have both independent and collaborative programs in V/STOL technology. The object of such programs is to combine ascent and descent capability with more efficient horizontal flight than is possible today with helicopters. The NASA propulsion studies in this program consist of the quiet, clean short-haul experimental engine (QCSEE) research and development and its incorporation into the quiet experimental STOL transport research airplane (QUESTOL). The support for engine development includes both externally blown flap and augmentor wing propulsion systems.

In the development of short-haul aircraft, it is necessary to integrate fully the propulsion aircraft and the guidance, control, and information systems. The NASA program is in many ways complementary to the Air Force experimental prototype STOL aircraft with an engine in the 20 000-25 000-lb-thrust category.

The two major applications of V/STOL systems in the military are in low-level close support and air superiority. Thus, the military requirement includes aircraft with speeds from near 0 to Mach 2 and operational altitudes from sea level to almost 18 000 ft.

Both shaft-driven (helicopters, tilt rotor, ducted fan, and tilt wing) and jet types (lift fans, thrust augmented wing, vectored thrust, and composite lift-thrust generators) are of interest. The U.S. Navy program thus consists of the Sea Control Ship and Marine Corps requirements; namely, the V/STOL Fighter Attack program and the sensor carrier or medium VTOL program.

The Fighter-Attack prototype program has considered the thrust augmented wing (TAW) XFV-12A, the lift plus lift/cruise, and the advanced Harrier (AC-16A). The medium VTOL program has considered "rubberized" engines for evaluation in various concepts. The TAW concept employs high-temperature air ducted from a gas generator to ejectors in the wings and in the forward canard surface. In the Mach 2 fighter-attack aircraft the propulsion system consists of two lift engines mounted vertically and one horizontal engine for cruise/lift, with a swiveling nozzle at the aft end. The direct engine exhaust in this case may present problems in deck-handling. Finally, experience with the Harrier engine has indicated that several advantages can be added to the aircraft with the addition of vectoring capability. However, several new considerations also arise, such as the inclusion of boost at takeoff and vectoring in forward flight. The introduction of the Plenum Chamber Burning (PCB) system boost

in a vectored turbofan can provide an increased boost ratio that is almost independent of fan pressure ratio and becomes a unique function of the ratio of core thrust to total thrust. The principal limitation arises from the temperature that can be allowed (775-875°C) in the exhaust pipe without cooling.

While the foregoing presents some long-range solutions to short-haul transport, there are also short-term needs for aircraft, capable of using runways 450-750 m long, with engines of thrust/weight ratio equal to 0.55 to 0.60. Powerplants for such missions may take the form of separate lift and propulsion engines with some form of lift augmentation or multiple-function powerplants with wing blowing. Fuel weight in such systems may be of the same order as engine weight, and that will have some influence on the engine cycle.

Supersonic Transport

The long-haul airline transportation base is continuing to rise, according to estimates, and this seems to indicate the need for increasing speeds as a means of increasing air transport productivity. The balance of cost may arise through flight offerings, especially in the intercontinental flights.

It is generally felt in the United States that the first-generation supersonic transport aircraft produced in Europe and Russia is unlikely to be fully acceptable and economically sound. The current objective in supersonic transport development is a broadly based interdisciplinary program in propulsion, structures, aerodynamics, stability, and control, to provide the technology base for a second generation of military and civil supersonic cruise aircraft. Apart from increasing technological and economic efficiency, great emphasis is being laid on meeting the environmental control requirements for noise and pollution.

It is important to recognize that a considerable body of knowledge already exists on jetliners for about Mach 2.2 flight in the military field. There may be further scope for reconsidering an increase in cruise Mach number in relation to range and capacity and associated economics of operation. The Aeronautics and Space Engineering Board of the National Academy of Engineering has expressed the view that inventions of a breakthrough nature are required in technology. Nevertheless, within the limitations of funding, systematic advances are being made. The NASA Advanced Supersonic Technology (AST) program is oriented toward attaining such advances. The spinoffs from this program have implications not only for future VTOL, RTOL subsonic transports and alternative fuel technology for aircraft but also for future military aircraft.

A number of advanced technology programs are also being supported by the Department of Defense in this area: the Air Force Advanced Propulsion System integration and the advanced turbine engine gas generator programs; the Navy V/STOL and PCT programs; and the Air Force-Navy joint technology demonstrator engine program. However, the position regarding development of advanced military aircraft is so unclear at the moment that there is little transfer of technology from military developments.

One important aspect of supersonic transport development is engine-airplane integration with a cooperative autopilot stability augmentation system-propulsion control package (Figure 7).

Another important aspect is the reduction in powerplant size and weight. Size reductions come mainly from optimum geometry, reduction in fuel consumption, and aircraft/engine matching. Weight reductions can be obtained by improved aerodynamic loading of the compressor and turbine components, increased heat release rates in the combustor and augmentor, and improved material and structural techniques in compressor and turbine blades and nozzles.

YF-12 COOP AUTOPILOT SAS PROPULSION CONTROL SYSTEM

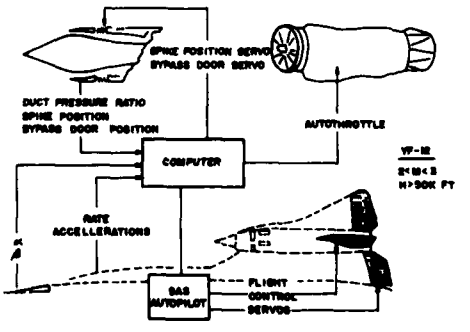


Figure 7—Propulsion control system

Current development of fan blades consists of evolving boron-aluminum blades that could reduce gross weight by several percent. The new blades, which can be processed in air, are made of a ductile aluminum-alloy matrix containing large-diameter boron filaments. Gas turbine vane temperatures can be increased to over 1000°C (the best allowable with superalloys) by using directionally solidified eutectics with almost three times the mechanical strength of standard superalloys. In the case of turbine blades, in addition to basic material strength, it is necessary that the blades should be capable of withstanding surface phenomena such as erosion and corrosion. The combustor liner and, in military applications, the augmentor liner also require attention in this regard. It is important in turbine blades that application of air cooling (350-450°C cooling air, gas stream temperatures of 1950°C, and vane skin temperatures of about 1100°C) does not reduce the directional strength of composite materials. In the case of nozzles, the possibility of using SiC-fiber-reinforced superalloy sheet has brought about weight reductions of 2 to 5% in airplane gross weight. Several engine testbed programs are planned for the next 5 years to demonstrate the required technology base in these problems.

Flight Above Mach Three

The most important factor affecting a propulsion system at a flight speed greater than Mach 3 is the relation between the allowable metal, cooling air, and fuel and lubricant temperatures. Above Mach 4, the internal structure heating must be considered in addition to skin heating. Beyond that speed, part of the energy that should have been available as thrust becomes absorbed in the molecular dissociation of exhaust products. It is therefore usual to divide flight regimes above Mach 3 into several groups: Mach 3 to 5, 5 to 7, 7 to 10, and 10 to 12. The latter are the speeds desired for future airbreathing launch vehicles.

The powerplants for high-Mach-number flight are the turbojet engine, ramjet, supersonic combustion ramjet, and composite engines. Candidate fuels are those with the required cooling capability and thermal stability—certain JP-type fuels, methane, and hydrogen. Thus, some potential candidates for high-Mach-number flights are (a) turbojet that is JP-fuelled for Mach 4.0, (b) precooled turbojet that is hydrogen-fuelled for Mach 5.0, and (c) turbo-ramjet with JP fuels for Mach 4.5, with methane for Mach 5.0 and with hydrogen for Mach 7.0. Such selection is based on the maximum allowable temperature for a critical part, such as a turbine disk in a turbojet engine, or a case in a subsonic combustion ramjet engine (Figure 8). Once air cooling is not feasible, one must resort to cooling with fuel that may have to be vaporized.

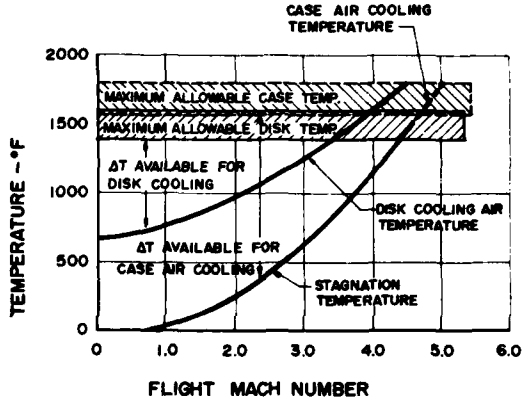


Figure 8—Disc and case cooling requirements

Many studies have been undertaken on the technical and economic performance of hypersonic transport (HST) for civilian and military applications. Estimates of cumulative international air passengers vs range indicate that in another decade 90% of the traffic would probably be in a design range of 10 000 km; therefore a hypersonic aircraft (for example, a turbo-ramjet aircraft with a cruise speed of Mach 6 at an altitude of about 30 000 m) has been examined for such missions. A number of configurations have been considered for flight at Mach numbers beyond 5; they range from an all-body to the standard wing-body.

The environmental problems—pollution, noise and sonic boom—are expected to be generally less severe in hypersonic transports than in supersonic transports. The ramjet can be turned on at Mach 3.5 and be operated with liquid hydrogen. In the development of such an aircraft, the cost of

initial development and the return on investment have to be considered carefully. It is certainly beyond the ability of any one manufacturer to undertake the development of such a vehicle unilaterally.

The development of hypersonic propulsion rests to a considerable extent on the success of the supersonic combustion ramjet (scramjet). In the Mach 3 to 5.5 range, the ramjet is more efficient than the scramjet because of the smaller pressure loss in the ramjet combustor. Beyond Mach 6.0, the scramjet is clearly superior.

The scramjet development is supported in the United States by the Navy, Air Force, and NASA. The Navy development efforts, carried out principally at the Johns Hopkins University Applied Physics Laboratory, has centered around experimental and analytical studies on various components of the engine. A heavyweight free-jet engine has been built and is the basic experimental tool for studies on the combustor, the nozzle, and various accessories.

The Air Force interest in hypersonic propulsion began with the Aerospace Plane, the single-stage Earth-to-orbit vehicle. This lead to the development of a subsonic combustion thrust chamber capable of hypersonic flight and to several scramjet engines. During the past decade the Air Force has sponsored several scramjet developments in various industries. One of these is the dual-mode scramjet, in which the combustor operates in both subsonic and supersonic modes. In the past few years the Air Force has concentrated much more on smaller missile systems with principal attention to high-density fuels with large energy content and associated problems.

The NASA scramjet development was initiated in 1965 with the Hypersonic Research Engine (HRE) Project, but the opportunity for testing the engine was lost when the X-15 program was terminated in 1968. The structural and aerothermodynamic performance of the HRE was tested in two models, the Structural Assembly Model and the Aerothermodynamic Integration Model. The success of those tests has lead to the current NASA effort in airframe-integrated scramjet research. An important factor in the development of Mach 10-12 flight vehicles is the necessity of using in the engine nearly all of the air between the underside of the vehicle and the vehicle shock wave. This can be achieved successfully with an engine built on the modular concept.

Currently, there is a NASA-Air Force effort to define a new versatile research airplane, the X-24C. The propulsion system study in this program consists of the establishment of the performance of a flight-weight, regeneratively cooled version of an integrated scramjet module with a high performance potential. Once the propulsion system is developed on the ground, it is proposed to flight test it in the X-24C, which can operate as a "flying wind tunnel" for the study of a variety of flight systems.

Finally, there is continuous interest in space launch vehicles that are fully reusable. The basic building blocks for propulsion systems in such vehicles are the rocket and airbreathing engines that can be installed either separately (combination systems) or integrally (composite systems), with the latter providing better performance. Advanced high-performance composite propulsion systems that operate over a wide range of Mach numbers can be established by combining ramjet, ejector ramjet (using a rocket to operate the ejector), supersonic scramjet, and liquid air cooled engine (LACE) technologies. In LACE engines, the cooling capacity of liquid hydrogen can be increased by using slush hydrogen with a lower boiling point. At present, the morphology of such composite engines is being established, and various component developments are under consideration.

Air Cargo Systems

The Department of Transportation has predicted that the U.S. domestic cargo demand will double in the next 20 years, but the air cargo part of this is not established at present. The Military Airlift Command (MAC) has released the "Military Concept of the C-XX," which establishes characteristics of a large all-cargo civil transport that can also be used in the Civil Reserve Air Fleet (CRAF) in a period of crisis. There is also a joint government-industry program (the DOT/Industry Intermodal Air Cargo Test, INTACT) for demonstrating the synthesis of air and surface modes of cargo transport. Some of the recent design concepts for advanced cargo aircraft are summarized in Table 4.

Table 4

Advanced Cargo Developments

Design	*Propulsion*	*Aerodynamics*	*Material (Composites) (%)*
Conventional	Advanced Turbofan	Supercritical	60
Delta Wing	Advanced Turbofan	Supercritical	50
Swept Spanloader	Advanced Turbofan	Supercritical	60
Ram Wing	Turboprop	Conventional	50
Unswept Spanloader	Advanced Turbofan	Supercritical	60

Some Military Developments

Developments to fill military needs are usually divided into strategic, defense, and tactical programs even in aeropropulsion. The basis of strategic capability is deterrence and that of tactical systems is capability in conventional war, defense, and striking back. In both cases, it is essential to develop a few systems while also creating a broad spectrum of viable options for other systems. The defensive part of strategic capability consists of air defense in providing surveillance and interceptor force with airborne radar capability. In air warfare, various systems should be developed: air superiority, deep strike/interdiction, defense suppression, tactical surveillance, command and control, and air mobility.

In view of such considerations, a number of systems with airbreathing propulsion systems are under development in the United Sates. Some of these are the B-1 bomber, air-launched and surface-launched cruise missiles, utility tactical transport aircraft (UTTAS), advanced medium STOL transport (AMST), heavy-lift helicopters (HLH), Air Force and Navy air combat fighters, F-15 and F-16 fighters, A-10 attack aircraft, and the advanced attack helicopter.

One area where there are important DOD-NASA joint programs is in V/STOL technology. Two aircraft with existing gas generators but new lift fans are being provided as test vehicles. There is also collaboration in the AMST program. Developments in V/STOL technology for military applications have been discussed in the section on short-haul transportation.

The Marine Corps continues to be interested in the thrust-augmented wing (TAW) and the growth potential of the AV-8 Harrier with a redesigned aircraft using the Pegasus 15 engine.

Department of Defense interest in hypersonic flight programs has been described earlier.

SELECTED RESEARCH AREAS

Aeropropulsion technolgoy is an excellent example of engineering activity in which systematic and sustained research has substantially helped to determine the return on investment. The technology involves advances in practically every field of engineering science, material development, and manufacturing processes. Research and development in any of those subjects has some influence on the design of the propulsion system.

It is obviously impossible to discuss the potential for research in all the areas of interest in this technology, which range from large-scale transport system studies to such small but critical items as seals in air passages. It seems more profitable—and certainly more illuminating—to

select a few areas for illustrating the nature of problems that need solution. On that basis, quite arbitrarily, the following topics have been chosen for futher discussion: turbine engine systems, fuels, combustion, turbomachinery, engine-airplane integration, and noise. In each case, the need for basic research is illustrated in connection with a few selected problems of technological interest.

In all aeropropulsion activities, the development of electronic computers and computational mathematics has played a central role in research and design. The development of computers, analog and digital, has lead not only to increased analytical applications but also, and in fact often faster and on a much larger scale, to the development of hardware systems for data processing, flight control, navigation, and weapon delivery. In a period of 5 to 10 years, computational capacity has been increased better than tenfold for a doubling of cost.

Such advances and corresponding developments in computational mathematics have led researchers and designers to apply computational techniques to a variety of problems. Broadly, the problems solved can be divided into two groups: (1) those in which an analytical approach is unavoidable because of the complexity of measurements, although not all aspects of the physical processes involved may be clear, and (2) those others in which one tries to establish a theory to compare with available experimental results.

In general the same classes of problems can be identified in design and performance estimation calculations. It is clear that computational methods can be extremely successful in complementary roles to experiments in the first class of problems, but one should proceed with considerable caution in the second class, where experimental studies are still needed primarily for observation and gaining understanding. Calculations in turbulent flows, nonsteady boundary layers in cascades and diffusers, heat transfer, aerodynamically induced vibrations, and so on are examples of the second class.

Developments in aeropropulsion will always depend on experimental test facilities. The need for such facilities has grown rather than diminished in recent years. The importance of redundancy in design verification and of obtaining as much performance data as possible in ground simulation and testing sufficiently large scale models is well established. Test facilities should be capable of incorporating such models at the required simulated flight and environmental conditions.

The NASA Langley Cryogenic high-Reynolds-number wind-tunnel program will fill a long-felt need in high-Reynolds-number transonic flow testing. Other engine testing facilities exist at the Naval Air Propulsion Test Center (NAPTC), Arnold Engineering Development Center (AEDC), and NASA test installations.

In addition, the Department of Defense has proposed the construction of an Aeropropulsion Systems Test Facility that will permit nearly full-scale ground testing. Current specifications for such a facility are as follows:

	Proposed Facility	*Best Available Capability*
Airflow (kg/s)	650	300
Air temperature range (°C)	-73 to +600	-73 to +430
Cooling system (tons/refrigerating capacity)	23 000	9510
Motor drive system (installed kW)	611 000	344 500
Test cell dimension (diam. x length) (meters)	7.5 x 20.5	6.0 x 30.0
Instrumentation channels	2170	1200
Cooling water (gal/min)	387 000	140 000

The economics of such a facility, estimated to cost about $437 million (1975), can be easily seen in terms of improvements in engine performance obtained through large-scale testing both with respect to fuel consumption and life-cycle cost: the savings over in-flight tests in only a few years will recover the capital outlay on the system.

In addition to computers and test facilities, the development of measurement techniques and instrumentation has had an important and universal impact on aeropropulsion research. In the past few years there have been recognizable advances in embedded probes, nonintrusive measurement techniques, nondestructive testing, and telemetry of data. There has also been substantial developments in data processing (for example, image processing and conditional sampling).

The major problems of measurement in propulsion systems arise in the following: fluctuating velocity and pressures in turbomachinery, temperature in cooled turbines, spray and particulate characteristics in combustors, turbulent and mean flow properties in reactive environments, shock-boundary layer interaction, transonic flow, positional changes in stationary and moving components, and flow structure interactions. In the latter two, high-energy radiation techniques and optical methods for mechanical movement detection in turbomachinery show considerable promise of becoming useful. The identification of processes such as separation movement during shock-boundary layer interaction and of unsteadiness in transonic flows continues to be difficult.

The measurement of pressure fluctuations away from boundaries is virtually impossible at present in small-scale flows. Regarding velocity measurements, recent advances in embedded probes and laser-Doppler velocimetry are quite promising. Density data can be obtained from laser interferometry and holography. The measurement of temperature and concentration in reactive environment is more complicated. When the environment is turbulent (as in gas turbine combustors) and when soot is present, it is not clear whether Raman laser spectroscopy or its variations can be adapted. In-situ measurements in two-phase flows (size and velocity characteristics) are under development in many laboratories. Imaging techniques include spark photography, laser holography, and telemicroscopy. In nonimaging methods, one obtains information from a small, continuously illuminated control volume as a function of time.

The U.S. DOD is interested in hydrocarbon exhaust plume characteristics, infrared emissions, condensation, and diffusion. The relationship between the IR scanner measurements and the flow and chemical kinetic parameters requires further investigation. The possible use of coherent anti-Stokes Raman spectroscopy (CARS) for thermometry and concentration measurement needs development.

Turbine Engine Systems

A recent survey in the United States has shown that of a total of about 141 000 registered airplanes, turbine-powered aircraft number about 2535, roughly half the total in the world. The remaining are piston engine powered. Aircraft piston engines are of course a small percentage of the piston engines in the United States, and there is continuous consideration of replacing some of the piston engines in aircraft with gas turbines. The gas turbine has established itself as the principal aircraft powerplant for major military and civilian transport in the past 30 years. At hypersonic speeds, the turbomachinery in a gas turbine may become impractical, and in any case the ramjet engine is superior to the gas turbine at such speeds.

A combustion chamber is common to all powerplants that use combustible fuels. In nuclear powerplants, there is need for a heat exchanger and also turbomachinery unless a nuclear ramjet is under consideration. A propulsor is common to all of the propulsion systems, but it can take the form of a propellor or a jet. The propulsion system may be required to provide lift in addition to thrust, as in V/STOL systems.

A question that immediately arises in regard to turbine engine propulsion systems is whether a modular approach can be adopted in the design of engines and propulsors. The answer is that the state of development in aeronautics does not permit at the moment more than a minimal modular approach. In fact, one may say that a continuing challenge in aeropropulsion is to identify a number of missions and establish a series of modular propulsion systems that can be integrated

with various aircraft to meet the mission requirements. When such a stage is reached, one can truly visualize a fully integrated engine-vehicle system. One area in which modular construction may be attempted even at this stage is the combustion chamber. As scaling laws become better established for diffusers and nozzles, the design of those parts can also be considered on a modular basis. A modular approach to turbomachinery in aircraft gas turbines can also be examined, although that will probably coincide in time with a great deal more certainty in regard to missions. The mission here is to be understood in terms of overall transportation for civilian applications. The implications of a modular approach for military aircraft missions can only be examined by taking into account logistic requirements and challenges.

In the development of future turbine engine systems, the following subjects are expected to be most significant: small gas turbines, variable-cycle engines, life-cycle performance, and integrated control systems. We shall discuss these briefly from the systems point of view.

Small Gas Turbines—A small flying engine, whether identified in terms of the small thrust or small physical size, cannot be "derived" from a large engine and yield the same performance parameters. A small engine, on the other hand, has a distinct role to play in terms of its operational capability. Cycle pressure ratios of 10:1 or higher and turbine inlet temperatures of 1200-1300°C are being considered for advanced small engines. The development of such engines presents some unique problems in air compression, cooling, dynamics, and manufacture; small engines may be said to be a generation behind the larger engines in development.

The small turbofan engine (1000-lb-thrust class) has a number of applications in the military area, for example, in remotely piloted vehicles (RPVs), cruise missiles, low-cost energy-efficient trainers, and a number of special applications such as the Subsonic Cruise Armed Decoy (SCAD). The small turbofan engine also has to be developed to enter the general aviation market.

The small gas turbine also finds extensive nonpropulsive application in both civil and military fields. The problems of development are again related here to size, fuel consumption, materials, and cost; some attention to noise will also be required.

Variable-Cycle Engines—The variable-cycle engine concept has been studied in various forms during the past 10-15 years, as a way to meet the needs of mixed mission aircraft that encounter high levels of throttle-dependent drag and are expected to meet somewhat contradictory performance and environment control goals. Thus, in its simplest form, it is expected to operate at least in two distinct modes of operation: (a) a high-airflow, low-jet-velocity mode for low-noise takeoff and/or efficient subsonic cruise and (b) a turbojetlike higher jet velocity, lower airflow mode for supersonic cruise.

The NASA Advanced Supersonic technology (AST) program has supported investigations on the applicability of several variable-cycle concepts. Two of them are shown in Figure 9. A single-valve variable-cycle engine would use a valve or diverter between dual fans so that parts of the airflow could be passed through either one or the other. The dual-valve engine can operate in

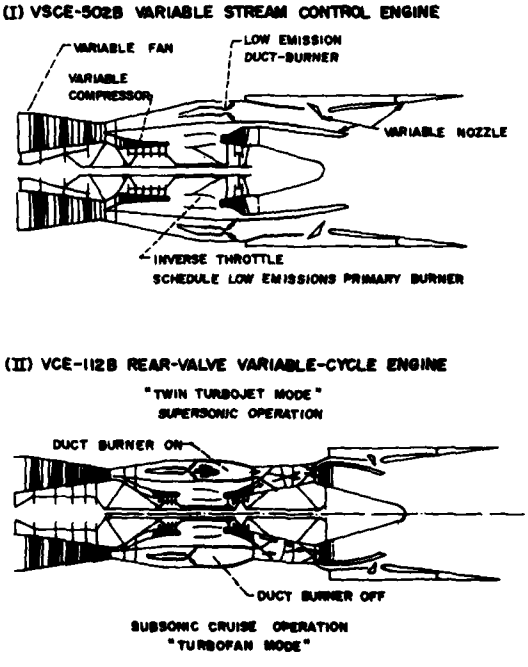


Figure 9—Typical variable-cycle engines

dual mode. The other scheme uses a dual burner and is called the Variable Stream Control Engine (VSCE); it incorporates the unique "inverted throttle schedule" and variable geometry.

Integrated Control Systems—The integration of airframe and propulsion system (engine, inlet, nozzle, and installation) has become vital in all aircraft, especially in V/STOL systems, supersonic aircraft, and military vehicles. The modern engine itself demands control of engine geometry as well as fuel flow. The number of engine variables controlled has increased in subsonic engines from 2 to 4 and in supersonic engines from 4 to 6, and in the latter it may increase to 10. The number of sensed control parameters has changed from 2 to 8 in subsonic engines and may increase to as many as 20 in the near future. It has therefore become essential to move from analogic electronic technology to digital control, especially because the engine control has become an interface in the overall system and is required to provide on-line optimization of propulsion system with a multivariate control.

In the application of such control systems, analytical methods need to be further developed for calculating the steady-state and dynamic performance and the loading of engine and airplane components. The interactions can become very complex, but they must be established in developing integrated or cooperative controls.

Life-Cycle Performance and Cost—Changes in engine performance and structural integrity are a function of engine cycle, design, production, and usage during various missions, including takeoff and landing. Engine performance deterioration can be measured in terms of changes in exhaust gas temperature and SFC (Figure 10). Structural changes are most significant in blading and turbine discs ($10 000 to $30 000 per disc), and therefore the reference parameters can be chosen as temperature and number of cycles.

Performance deteriorates because of loss of component efficiency, change in clearances, and variation in effective gasflow areas. It is accordingly possible to express changes in performance in terms of changes percent change in component efficiency or in terms of influence coefficients for clearance changes. Engine deterioration characteristics can be categorized in three specific time periods: less than 1000 hours (structural changes due to takeoff and landing procedures), 1000 to 3000 hours (erosion and other damage), and over 3000 hours (turbine blading and disc changes).

Several extremely difficult problems arise: measuring changes in performance and structure during flight or relating test cell measurement to onflight performance; analyzing the performance changes; determining a restoration program at site; establishing life-cycle cost; and taking life-cycle cost into account during design, production, and acquisition. In addition, the coupling between engine deterioration and control system is important in most tactical aircraft.

Life-cycle cost should include R & D, acquisition, and operations and support cost. While some data on past experience exists, there is no sufficiently unified and accepted methodology for estimating or taking into account life-cycle costs. A joint AF/Industry engine life-cycle-cost group has been formed to establish accounting models for life-cycle cost.

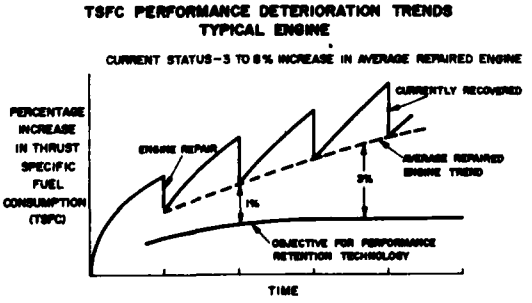


Figure 10—TSFC deterioration trends.

Fuels

It may safely be said that in the next three decades the greatest emphasis in aircraft fuel technology will be in two areas: (a) the determination of a broad spectrum of hydrocarbons high in density, energy content, and safety and low in volatility, freezing point, and deposits and undesirable emissions and (b) the production of synthetic hydrocarbons. The combustion parameters of significance are the combustor liner temperature, combustion products, and combustion efficiency. Development of hydrocarbon fuels in the near term will probably be in the direction of

determining various blends of standard fuels with additives such as xylene or pyridine, the former for increasing hydrogen content, the latter for increasing nitrogen content. The effectiveness of decreasing liner temperature and smoke with increased hydrogen content can also be demonstrated in shale-produced JP-4. The method of testing such fuels in the laboratory and in actual combustion chambers requires further investigation.

Apart from the alternative fuels mentioned earlier, consideration has been given to such apparently extreme developments as laser-beam-generated compounds. It has therefore become of utmost importance to understand the detailed chemical mechanisms involved in synthesis of fuels and combustion and to be able to calculate and verify experimentally the rates of chemical and transport processes. All of these involve experimental studies; the computational models require at least certain crucial reaction information. A large amount of chemical kinetic data has been established in the past, but much of this low-temperature data cannot be extrapolated to higher temperatures without considerably more studies on the formalism of the reaction mechanism and on the branching ratio.

The temperature range of interest is 1000-2000 °K. A case in point is the controversy over whether the formation of NO_x, is due to a super-equilibrium concentration of free radicals or to other reactions involving fuel derived species and what the influence of temperature is on the NO_x occurring in a certain location. Similar difficulties arise with the pyrolysis and oxidation reactions of hydrocarbons. It is, for instance, safe to speculate that coal-derived liquid fuels will contain the higher order aliphatic hydrocarbons and aromatics. It is then necessary to examine gas phase oxidation mechanisms and kinetic rates of higher order paraffins, olefins, and benzene.

The ultimate understanding of a reaction mechanism should probably depend on the ability to perform quantum-mechanical level calculations. Such calculations are extremely difficult to perform, and it may be almost impractical to apply quantum-mechanical calculations to complicated fuel molecules even with advances in computation and computers. Nevertheless a beginning has been made in simple reactions (for example the hydrogen atom-hydrogen molecule exchange reaction), and these calculations should at least provide a means of verifying the assumptions introduced in the study of various reactions. The most important data needed for quantum-level calculations are transition probabilities, and it appears that further advances in molecular beam and other techniques should be useful in this regard.

The study of oscillatory combustion reactions, coupling of acoustic-chemical kinetic interactions, and turbulence-reaction kinetic interactions are other directions in which advances show promise for the development of fuels.

Combustion

The object in combustion chamber design is to obtain in a chamber of the smallest volume and length the largest enthalpy release in the most orderly fashion, over a range of fuel composition, fuel/air ratio, and entry conditions, with the smallest heat transfer to the walls, the smallest quantities of undesirable emissions in the exhaust, and the least noise. Both piston and continuous combustion chambers are of interest in aeropropulsion, and the ideal is far from achieved in both types of combustion chambers. To see the implications of fundamental research in this area, it may be useful to examine some design aspects of combustion chambers, for example, from the point of view of pollutant generation and control. Various criteria of performance will in any case become intertwined. Chamber liner temperature and thermal efficiency are both affected by changes made to reduce emissions.

The Environmental Protection Agency (EPA) prescribed in 1973, following the Clean Air Act of 1970, standards and test procedures for aircraft engine emissions. The emissions of interest are undesirable gaseous constituents and smoke. There is obvious interest in military applications in radiative emissions from the engine exhaust. The chemical pollutants, gaseous and particulate, are the results of impurities in the fuel (e.g., sulfur), incomplete combustion (resulting in CO and toxic hydrocarbons, or THC), and distribution of enthalpy generation in the chamber. The particulates can be reduced to some extent in

conventional combustion chambers by fuel blending, atomization, and mixing, but considerable efforts are required for understanding the formation of soot and in producing measurement techniques for soot before this problem can be said to have been eliminated.

Aircraft piston engines are designed for the best integration with aircraft and generally operate in minimal ranges of speed and power. The fuel/air ratio in high-performance engines is generally "fuel rich" under high power (takeoff and climb) conditions, causing overheating and detonation. There is obvious scope for improvements in fuel management.

A continuous combustion chamber is governed by a number of time and space scales related to fuel vaporization, reaction rates, flow Mach number, transport processes, and frequencies of unstable processes. The influence of each of those becomes further complicated by composition of the fuel, fuel/air ratio, geometry of the combustion chamber, and flow configuration.

In continuous combustion chambers, the emission indexes (grams of pollutant per kilogram of fuel burned) for CO, THC, and NO_x depend on operating condition of the engine. Emission indexes for CO and THC are highest under idling conditions, whereas that for NO_x is highest at peak power. It is therefore necessary to consider solutions to the problem that are not entirely incompatible with one or the other requirements. Fuel and air distribution is one principal means of obtaining overall improvement, although in practice this may eventually necessitate variable-geometry combustion chambers. It is very important to recognize here the development work that will be required over many years before laboratory demonstrations of reduction in pollution can be translated into economically feasible (in terms of fuel consumption, weight, and pressure losses) engine designs.

If we divide the operation schedule of engines into peak-power and low-power operation, emissions control research may be classified as follows:

Peak-power operation:	Leaner mixtures
	Premixing
	Swirl can
Low-power operation:	Fuel scheduling
	Fuel atomization and distribution.

Fundamental research related to such development is needed in the areas of fuel injection and atomization, vaporization dynamics, reaction rate control, flame extinction, and combustion instability. Fluid flow considerations become dominant in all of those problems unless one can control reaction kinetics entirely independently.

Another area of research is related to the fact that combustion in propulsion engines is invariably turbulent. The interaction of turbulence and chemistry is therefore crucial in the design of combustion chambers. Whether the reactants are premixed or undergo mixing while in the process of reacting, the two engineering parameters of interest are the rate of flame propagation and the rate of enthalpy generation. Calculation procedures have to be evolved for these so that they can be incorporated into the combustor design. Two crucial factors in regard to turbulence are scale and intensity. A variety of time and space scales becomes of interest in combustion chambers. The influence of turbulence scale on chemical reaction processes is generally difficult to assess in universal fashion. On the other hand, the influence of turbulence intensity on combustion processes has been recognized at least on a global basis.

The exothermicity of reaction raises another basic question: what is the effect of heat release on turbulence structure? It is not clear how to separate the effects of Mach number, density, and temperature on turbulence quantities, either in measurements or in analysis. Progress is being made in the understanding of free mixing processes, but two areas of greatest interest in practical systems, the effects of Mach number and variable density on turbulent mixing, remain largely unsolved questions. Turbulence and flow parameters are also known to affect ignition energy and quenching distance. Considerable research is required in this area before general conclusions can be reached. This subject is also of obvious significance in fire research.

In subsonic combustion chambers feeding turbines, it is important that the combustor exhaust should have uniform properties; to achieve it, one

needs uniformity of dilution air entry conditions into the chamber. The diffusion of air into the chamber, for example, with a "dump" diffuser—a small angle diffuser followed by a sudden expansion—provides interesting opportunities and challenges.

Supersonic combustion is of interest in hypersonic aircraft and supersonic combustion ramjets. The inlet and the nozzle become even more fully integrated with the combustion chamber here than in subsonic combustion ramjets. Such integration further restricts the geometry of the combustion chamber, especially its inlet. The static pressure in a supersonic combustion chamber is lower than in subsonic combustion, and this provides considerable savings in structure, but the static pressure rises during heating in a supersonic stream and there can arise choking of the flow. Fuel injection schemes for supersonic combustion are yet to be developed, although some gross results are available on parallel and tangential injection with combustion.

Supersonic combustion with oblique or normal injection of fuel is of interest in external burning, that is combustion in the free stream flowing past an aerodynamically shaped body. The ignition, stabilization, and location of the flame with respect to the body then become critical parameters; they depend on the gas dynamic and mixing processes in the vicinity of fuel injection.

Finally, the problem of combustion instability is of interest both in a main combustion chamber and in an augmentor. Regarding the combustion chamber, many investigations have been conducted on the occurrence of high-frequency instability, and a "screech liner" has been developed for damping the instability. The afterburner is also faced with the problem of low-frequency instability (below 100 Hz), which can cause a blowout of the combustion process and induce stalling of the fan. The problem of blowout is acute under certain maneuvering operations and with increases in humidity such as in rain. Ignition and thrust modulation require stable operation over a wide range of equivalence ratios, 0.03 to 1.0, and therefore fuel zoning or stratification becomes unavoidable. One method of avoiding the use of flame holders is to introduce air in the form of high-velocity swirling jets into the combustion chamber. Combustion in swirling flows is a complicated subject and there is little basic understanding of the processes involved although development programs have yielded excellent results in specific cases.

Turbomachinery

It is generally recognized that developments in turbomachinery have gone through four stages during the past 30 years: (a) application of classical aerodynamic theory to turbomachinery blades and passages and correlation of experimental results on that basis; (b) development of various computational schemes for the calculation of internal flows; (c) adaptation of the more refined fluid mechanics and aeroacoustics for turbomachinery, and (d) utilization of such theoretical and computational developments in design and in performance estimation.

None of these efforts can be said to be complete or unnecessary even today. Turbomachinery has become an independent discipline with a variety of applications. In both the education of designers and the synthesis of research and development teams, it is necessary to demonstrate methods of successfully incorporating analytical and experimental results. During the past 10 years there has been an attempt at obtaining the kind of physical understanding and experimental results that are vital to the application of some of the more comprehensive computational programs in design. This effort will continue for at least another two decades, for example, in the area of viscous, nonsteady flows, where many of the earlier ideas rightly are being questioned today.

Fans in turbofan engines generally have supersonic velocity at the tips. This is based on the desirable turbine RPM and obtaining high efficiency at various values of bypass ratio. The design of transonic blading is therefore of importance in this problem. In axial compressors also, the reduction in the number of stages and increase in stage pressure ratio (up to 1.26), both introduced to reduce weight, and the increase in mass flow have increased flow velocities, and the occurrence of supersonic velocities has become common. Both the performance estimation and design of such stages are continuing challenges to the research analyst.

While three-dimensional flow programs are being developed for compressors, the complete mathematical solution of three-dimensional flow fields is beyond current expectations. Even two-dimensional viscous flow calculations cannot be carried out in adequate detail. In more elementary cases, singular perturbation techniques have yielded benchmark solutions that can be used as building-block solutions in attacking more complex problems. However, this is hampered by the appearance of various processes such as nonsteady separation, transition, and relaminarization and the lack of adequate experimental data for checking analytical solutions. It is therefore extremely important to carry out experimental studies involving flow visualization (smoke, fluorescence, etc.) and measurements employing nonintrusive optical techniques.

Unsteady Phenomena—Unsteadiness can arise in every part of a propulsion system, but in turbomachinery the flow is basically unsteady. There is design point unsteadiness, but there are no satisfactory theories or calculation procedures for design of turbomachinery as nonsteady machines, except perhaps at low loading. At higher loading one must take compressibility into account, and at supersonic speeds it is necessary to include oscillatory shock waves.

The problems of surge and rotating stall continue to be unclear. Linear theories are almost certainly inadequate, and measurements in detail on the surface of blades are required. There is both upstream influence—inlet vortex—and response to downstream influence, and they are not accounted for in current theories.

The problem of maldistribution is distinct from the onset of instability. Maldistribution can consist of radical and circumferential distortions of flow, temperature, and turbulence. The determination of the alteration of maldistribution, and in fact the characterization of maldistribution, are completely unsolved problems. Maldistribution affects surge margin, but no detailed analytical or experimental results are available. It is also known that distortions may be self-generated by interaction between blade rows. The manner in which they appear finally at the exit of the compressor should be related to the design of successive blade rows.

The two other basic questions of wide import are the calculation of unsteady boundary layers and of flutter or the aeroelastic interactions. The latter is important in both the stalled and unstalled conditions. Three-dimensional effects must be fully incorporated in analyses of such conditions. Low-pressure turbine blades are also susceptible to aeroelastic unsteadiness.

In experimental studies, there is a continuing question of whether cascade results are applicable to rotating machinery. Annular blade rows are certainly required in many problems.

In an actual engine compressor, there is the interaction between stages and the resulting change in distortion and overall stability characteristics. It is then important to establish how to get out of the stalled condition quickly, and also the postsurge condition.

Turbine Cooling—A variety of techniques have been developed for cooling turbine blades. Their application is essentially a function of changes in material and manufacturing technology. At present, considerable data has been accumulated on film cooling effectiveness as a function of blowing rate in a variety of coolant injection configurations in free stream flow conditions corresponding to various Mach and Reynolds numbers. But, experimental information is not adequately detailed for developing analytical models for calculating heat transfer reduction due to injection and for scaling such performance with respect to flow and injection parameters.

The flow field, for example, in the neighborhood of a single hole through which a coolant is injected into a cross-flowing stream can be established accurately only for selected flow conditions that are in general far removed from engine flow conditions. In the case of microsized holes drilled over the surface of a turbine blade for injection of coolant, there is little hope for analysis without detailed experimental observations. The interaction between the injectant and the boundary layer fluid (or the free stream) under turbulent mixing flow conditions must be established eventually on an analytical basis for incorporation into design.

Engine-Airplane Integration

Because engines and airplanes are produced separately, so long as there is variation in mission

characteristics the problems of engine, airplane, and control integration will be important. Integration can take several years since in-flight tests are essential in most cases and adequate facilities are yet to be established for on-ground testing. In an installed propulsion system one must take into account the interaction among engine, air inlet, exhaust nozzle, installation, and airplane. The problems of integration become acute in the case of transonic, supersonic, and V/STOL aircraft.

Two problems of common interest in all aircraft are (a) the changes in the aerodynamic and structural performance of the vehicle and engine and (b) the influence of engine exhaust on wing vortexes. The changes in performance can only be established by proper accounting of forces and moments on various parts of the system, which can in turn be established only by detailed testing at appropriate scales. One is still left with the problem of accounting for such forces and moments, and this can be controversial.

The problem of wing vortexes has become critical in large, high-speed aircraft, which must be separated by 5-10 km in flight. While good progress is being made on vortex control through dissipation, the vortexes can become stabilized with engine exhaust ingestion, depending on the location of engines. This also has strong implications for pollution of the atmosphere. A contrail from large supersonic aircraft can be several tens of kilometers in lateral scale and persist for days. The understanding and modeling of various processes connected with exhaust gas ingestion is an important problem.

The problems of integration in transonic and supersonic aircraft may be divided broadly into the following: air intakes and airframe-inlet interactions; nozzles and afterbody flow field interactions; and forebody-afterbody interactions. Considerable advances are required in fluid dynamics and structures in the detailed interaction processes. In particular one may emphasize shock-boundary layer interaction, nonsteady flows, turbulence distortion effects, base drag, spillage effects, and aerodynamically induced vibrations.

Very complex integration problems arise in V/STOL systems. Some concepts for such systems that will continue to receive attention are high-lift wing using boundary layer control and cruise thrust deflection; jet flap and augmentor wings using engine flow and ejector-generated flow; externally blown flap using engine exhaust for blowing; and direct lift with lifting or vectored thrust engines.

The technological implications of these concepts are clear, but both analytical and experimental investigations are required before the highest efficiency is attained in any concept. For example, calculation of three-dimensional duct flows, determination of ground effects and induced loads, study of exhaust gas ingestion, estimation of losses in flow deflectors, three-dimensional wing theory with and without blowing and mixing, and behavior of turbulent jets in streams oblique to the jet are some of the broad areas in which further investigations are required. Ejectors, mentioned earlier, require further analysis.

Noise

A broad attack on the problem of noise involves at least three areas of investigation: (a) the source, which deals with noise generation and suppression. (b) the path, which deals with propagation and attenuation of noise and hence is coupled with aircraft flight operation, and (c) the receiver, which encompasses individual and community response through compatible land use.

The greatest advances are yet to be made in the areas of noise generation and suppression, especially from the point of view of noise control through suppression. Broadly the required effort in research can be divided into reducing airframe noise (including the noise due to the integrated lift and propulsion systems) and engine noise (including jet and inlet noise). It may be emphasized here that progress in the calculation of noise generated from various sources, at least on a global basis, is not matched in general by progress in noise suppression, although ducted fan and jet noise have been lowered substantially in the past few years.

Further research is required on static and moving jets for high-bypass-ratio engines with lower speeds and temperatures and coaxial configurations. The ejector or augmentor nozzle is also significant in V/STOL systems. One solution to jet noise is the use of multielement nozzles consisting of multiple tubes, chutes, spokes, coaxial

elements in combination, and various combinations of all of these. One is then interested in the complex noise field produced by interaction of the various elements. The interaction leads to both a change in frequency of the radiated sound and a shielding effect due to the peripheral jets. There is clearly some influence of turbulent mixing leading to a change in structure at various interfaces but it is not a solved problem.

Further progress is also required in regard to fan, compressor, and turbine noise. Fan noise, a considerable nuisance during approach to land, is related to rotor blockage; further research is required in transonic fans to establish various means of changing the rotor blockage effect, including blade row spacing. The fan rotor profile can be changed to obtain a flow pattern in which the location of shock waves from blade tips is displaced to avoid interaction with neighboring blades. The casing wall directly above the rotor tips can be treated acoustically to absorb sound. Fan noise may be prevented from radiating outward with a variable-area inlet that can provide near-choking conditions for various airflows. The most successful method of controlling turbine noise appears to be through aerodynamic loading and acoustic treatment of casing walls. These problems require further investigation.

Both in high-bypass-ratio turbofans and in variable-cycle engines, it is necessary to establish the noise generated during duct burning and to reduce it. General investigations on turbulence and acoustics of confined flames are required. Experimental investigations on specific configurations can be of doubtful validity since it is important to identify and to measure true sound in the system.

The control of noise from ducts and air passages can be achieved through increased speed of air and incorporation of sound-absorbing materials. However, one must take into acount all the noise sources in such ducting, and it is not easy to isolate the causes clearly for investigation. The effect of some small rotation in the flow is a case that illustrates the difficulties.

Noise due to engine-airplane integration is especially critical in the case of blown flap systems, in which the entire engine flow may be exhausted directly over the wing. Nozzle development is one aspect of this research. The over-the-wing engine installation is the more favorable from the point of view of interference noise. However, further research is required in understanding noise from deflected flows. Some measurements are available on the three-dimensional noise field in installed configurations, and the identification of noise sources should eventually lead to noise suppression methods. The effect of scale in tests should be understood further.

There is a possibility of noise-induced structural fatigue in certain situations, and although the dynamic features of the problems are clear the problems of isolating the effects of noise remain.

An important question in noise-control technology is what application of the technology will cost in terms of increase in direct operating costs due to increased specific fuel consumption, weight, and thrust losses. In fact, in the case of supersonic aircraft much of the subsonic jet silencing technology may become inapplicable. This should be taken into account in all aspects of noise-control research.

CONCLUSION

The United States export in transport aircraft alone is currently on the order of $9 billion and, taking into account domestic purchases on the order of $17 billion, the amount involved in the balance of trade could be on the order of $26 billion. Currently it is clear that U.S. aeropropulsion technology as a whole is superior in most respects, although there are undoubtedly areas of advances in Europe. That superiority should be sustained with national support.

In view of the recognized implications of this technology in transport and defense, the continued backing by foreign governments of aeroindustry in their own countries can reduce the U.S. lead to a dangerous low. The U.S. method of developing various options with advances in technology base through support of basic research and development is very well suited to this technology. It should, however, be combined with bold and imaginative commitments to fulfilling opportunities and needs.

ACKNOWLEDGMENT

The bibliography is an indication of the kind of literature to which the author owes his acknowledgment in this study. Numerous other references consulted have not been mentioned. Personal discussions with James R. Patton of the ONR Power Program during preparation of this article have been most helpful. In addition, discussions with Professor Bruce A. Reese, Purdue University, have always been enlightening on the subject of airbreathing propulsion developments and related subjects.

BIBLIOGRAPHY

Adamson, T. C., and M. R. Platzer, eds. "Transonic Flow Problems in Turbomachinery." Proceedings of workshop held Feb. 11-13, 1976. To be published in 1976 by Hemisphere Publishing Co., Washington, D.C.

"Advanced Aeronautical Concepts." Hearings before the Committee on Aeronautical and Space Sciences, U.S. Senate, 93d Cong., 2d sess., July 16 and 18, 1974.

"Advanced Supersonic Technology." Hearings before the Subcommittee on Aeronautics and Space Technology of the Committee on Science and Astronautics, U.S. House of Representatives, 93d Cong., 2d sess., Feb. 22, 1974. U.S. Government Printing Office, Washington, D.C., 1974.

"Aeronautical Research and Development." Hearings before the Subcommittee on Aeronautics and Space Technology of the Committee on Science and Astronautics, U.S. House of Representatives, 93d Cong., 2d sess., Jan. 18, 19, and 20, 1972. U.S. Government Printing Office, 1972.

"Aircraft Fuel Conservation Technology." National Aeronautics and Space Administration, Task Force Rep., Sept. 1975.

"Aircraft Fuel Efficiency Program." Hearings before the Committee on Aeronautical and Space Sciences, U.S. Senate, 94th Cong., 1st sess., Sept. 10, Oct. 23, and Nov. 4, 1975. U.S. Government Printing Office, Washington, D.C., 1975.

"Airplane/Propulsion Interference." Agard Conference Proceedings, CP. No. 150 NATO Neuilly-sur-Seine, France, 1974.

Carta, F. O., ed. "Unsteady Flows in Jet Engines." Proceedings of workshop, July 11 and 12, 1974. Project Squid Report (UARL)-3-PU, Nov. 1974. ADA003853, NTIS, Springfiled, Va.

"Civil Aviation R/D Policy Study" (DOT-NASA). DOT-TST-10-4. NASA SP-265, 1971. (See also NASA SP-266, 1971.) National Aeronautics and Space Administration, Washington, D.C., 1971.

Covert, E. E., and J. L. Kerrebrock. "Coming Together on Airbreathing Propulsion Research." *Astron. Aeron.*, Sept. 1975, p. 58.

Eltis, E. M. "The Influence of Effective Research and Development on the Aero-engine Business." *Proc. Roy. Soc. A* 312:333 (1969).

Ferri, A. "Review of Scramjet Propulsion Technology." AIAA Pap. No. 66-826, 1966.

Ferri, A. "Possibilities and Goals for the Future SST" (Dryden Lecture). AIAA Pap. 75-254, 1975.

Flax, A. H. "Aeronautics—A Study in Technological and Economic Growth and Form." *Aeron. J.*, Dec. 1974, p. 537.

Fuhs, A. E., and M. Kingery, eds. "Instrumentation for Airbreathing Engines." *Progress in Astronautics and Aeronautics*, vol. 34, MIT Press, Cambridge, Mass., 1974.

Fultz, J. R. "Future Air Force Requirements for Hydrocarbon Fuels." Wright-Patterson AFB Rep. No. TR61-728, May 1962.

Glassman, I., and W. A. Sirignano. "Summary Report of the Workshop on Energy Related Basic Combustion Research." Energy Related General Research Office, Rep. No. 1177. National Science Foundation, Washington, D.C., 1974.

Goulard, R., ed. "Combustion Measurements in Jet Engines." Hemisphere Publishing Co., Washington, D.C., 1976.

Heiser, W. H. "New Perspectives for the Universities in Airbreathing Propulsion." *Astron. Aeron.*, Sept. 1975, p. 60.

Hooper, J. A., et al. "Lift Augmentation Devices and Their Effect on the Engine." Agard Lecture Series No. 43, Apr. 1970.

Kuchemann, D., and J. Weber. "An Analysis of Some Performance Aspects of Various Types of Aircraft Designed To Fly Over Different Ranges at Different Speeds." *Prog. Aero. Sci.* 8 (1968), Pergamon Press, London.

Lighthill, M. J. "Sound Generated Aerodynamically" (The Bakerian Lecture, 1961). *Proc. Roy. Soc. A* 267:147 (1962).

Murthy, S. N. B., ed. *Turbulent Mixing in Nonreactive and Reactive Flows*. Plenum Press, New York, 1975.

Murthy, S. N. B., ed. *Aerodynamics of Base Combustion*. MIT Press, Cambridge, Mass., 1976.

Muse, Thomas C. "Military Contributions to Civil Aviation." AIAA Pap. No. 73-67, 1973.

National Research Council. "Environmental Impact of Stratospheric Flight." National Academy of Sciences, Washington, D.C., Mar. 1975.

Olsen, J. H., A. Goldburg, and M. Rogers, eds. *Aircraft Wake Turbulence*. Plenum Press, New York, 1971.

"The Outlook for Aeronautics 1980-2000." National Aeronautics and Space Administration, 1976. NTIS, Springfield, Va.

Platzer, M. F., ed. "Prediction Methods for Jet-V/STOL Aerodynamics," vols. I and II. Proceedings of workshop held July 28-31, 1975, Naval Air Systems Command, 1975.

Rom, F. E. "Status of the Nuclear Powered Airplane." *J. Aircraft* 8:26 (Jan. 1971).

Sears, W. R., ed. "Unsteady Aerodynamics." Proceedings of a symposium held Mar. 18-20, 1975. University of Arizona, Tucson, 1975.

Stever, H. Guyford. "How Should Civil Aviation Develop To Serve Our Society Best?" Keynote address at the President's Forum, AIAA 5th Annual Meeting and Technical Display, Philadelphia, 1969.

Stewart, J. T. "Evolving Strategic Airpower and B-1." *Astron. Aeron.*, June 1972, p. 22.

Sumey, I. E. "Influence of Fuels and Lubricants on Turbine Engine Design and Performance." Rep. No. AFAPL-TR 73-54, vol. II, June 1974.

Taylor, E. S. "Evolution of the Jet Engine." *Astron. Aeron.*, Nov. 1974, p. 64.

Torell, B. N. "The Significance of Propulsion in Commercial Aircraft Productivity." *Aeron. J.*, Dec. 1975, p. 537.

Walthrup. P. J., G. Y. Anderson, and F. D. Stull. "Supersonic Combustion Ramjet Engine Development in the U.S." Paper presented at the 3rd International Symposium on Airbreathing Engines, Munich, Germany, 1976.

Weber, R. J. "The NASA Research Program on Propulsion for Supersonic Cruise Aircraft." SAE Pap. No. 75-629, May 1975.

Zollinger, Joe E. "Structural Integrity for Propulsion Systems." *J. Aircraft* 12:195 (Apr. 1975).

James L. Tocher is Manager for Engineering Computing of the Energy Technology organization of Boeing Computer Services, Inc. Dr. Tocher joined the Computing Department of the Boeing Company in 1964. In 15 years of dealing with engineering mechanics problems, he has been involved in the development of new finite elements and in the application of finite elements to problems in applied stress analysis; he has worked on problems of inelastic analysis, large deflections, thermal stress, and automated weight minimization; and he has directed work in finite-element technology, structural computing techniques, optimization, vehicle occupant simulation, numerical analysis, and computer-aided design. He has written more than 20 papers describing these activities and has held a part-time appointment in the Civil Engineering Department of the University of Washington. Dr. Tocher earned B.S., M.S., and Ph.D. degrees from the University of California, Berkeley, and did postdoctoral work at the Technical University of Norway.



FUTURE DESIGN AND ANALYSIS OF NAVAL STRUCTURES: THE IMPACT OF COMPUTING TECHNOLOGY

James L. Tocher

Boeing Computer Services, Inc.
Seattle, Wash.

The digital computer has had more impact on structural analysis in the past 15 years than all the structural developments of the preceding 200. The computer is changing the way the Navy designs, analyzes, tests, and operates its ships, airplanes, helicopters, and other hardware. This paper will focus on structural computing—its growth, its impact, and its future directions—and try to describe how naval structures will be influenced by, say, 1985.

The impact of computing on structural analysis can best be seen by looking back to 1959 (when things were just beginning) and comparing it to what we have in 1976. In those 17 years the changes have been enormous, and for that reason it is difficult to project ahead just half that time, in 1985. Fortunately, there is always a substantial lag between research and application. (It takes some time to separate the wheat from the chaff in the research business and to arrive at cost-effective tools and methods.) Thus, if one looks at present-day research and judiciously projects ahead a few years, a reasonable guess can be made as to what will be happening in production projects in 1985. A list of present-day awkward or unsolved problems that very likely will be resolved in the next 9 years can be drawn up. This kind of projection strategy can work well provided one is not too specific. For example, 5 years ago it might have made sense to write a forward-looking paper on the future of new space-age materials in the construction of the slide rule!

The framework for describing structural computing developments has three parts: analytical capability, computing power, and data handling (user-computer interface). The three work together like the legs of the old milking stool, although it is common for researchers to think of their particular leg as more important than the other two. This three-way dependence will appear as a recurring theme throughout this paper.

A BRIEF DESCRIPTION OF COMPUTERIZED STRUCTURAL ANALYSIS

We should first provide a little background to this discussion of the past, present, and future by describing what a structure is, why we want to analyze it, and what kinds of analyses are performed.

If you ask people to name a few kinds of structures, the usual response would be bullfrogs, bridges, dams, and towers. Few people would name such important structures as the human skull or spinal column, a ship propeller and its bearings, an energy-absorbing helicopter pilot seat, a nuclear reactor, or an airplane wing. All of these are critical naval structures to which computerized structural analysis is applied. Structures are made of many different materials, some

mathematically well behaved (such as steel) and some that are downright hard to characterize (such as the spinal column). Structures come in all sizes and shapes, but, surprisingly, with the advent of computerized structural analysis they can almost all be studied with a single analysis technique called the finite-element method. This technique, which we will describe later, probably is used for 80% of the complex structural analysis done in the United States and Europe.

Structural analysis is the computation (prediction) of the behavior of a structure as it works. (The analysis can be done either by hand or on a computer, depending on the problem and the tools available. This paper will focus entirely on computerized analysis, although hand analysis is still an important part of any structural study.) The internal stresses, deflections, tie-down forces, vibration frequencies, buckling load, ultimate strength, and energy absorption capability all can be computed. In the structural business, terms such as static loads, transient dynamics, steady-state forced response, large deflections, natural frequencies, mode shapes, plasticity, composites, flutter, and fatigue failure are seemingingly flung about with great abandon. Actually, all of these terms are rather precise engineering descriptions of common structural phenomena. A layman's guide to some of the common structural terms is given in Table 1.

Predicting Behavior

Why do structural analysis at all? Why not just build it and test it? This approach was good enough for the Wright Brothers, but they didn't get too far off the ground or go too fast. Once they found that their structure wouldn't collapse under normal maneuvers, though, they may have wondered if possibly some of the parts were too big

Table 1
Structural Definitions of Everyday Observations

Example	*Structural Term*
Atlas holding up the world	Static load
A pole vaulter's pole bending	Buckling
The shape of a vibrating violin string	Mode shape
A tuning fork vibrating at 400 Hz	Natural frequency
Inflating a rubber balloon	Nonlinear elasticity
Boy Scouts crossing a rope bridge	Large deflections
Twisting of a "frozen" bolt by mistake	Plasticity and inelastic failure
Fiberglass boat hull or snow ski	Composite material
A car front end crushing under impact	Structural crash dynamics
A golf club hitting a ball	Transient dynamics
An unbalanced washing machine spinning	Steady state forced response
A venetian blind vibrating in the wind or the Tacoma Narrows bridge (Galloping Gertie)	Flutter
Bending a paper clip back and forth until it breaks	Low cycle fatigue failure
An engine mount that breaks after 50,000 miles	High cycle fatigue failure
Nylon tires which thump every morning until you drive a few miles	Creep

and the airplane was carrying around excess weight.

So we immediately see the reason to do structural analysis—to check that a structure is safe under service loads, not too heavy, and not too costly. Preferably this checking is done while the structure is "on the drawing boards," where design and modifications are cheap. Analysis will never completely replace flight tests, shakedown cruises, and track tests, but with the computing capability now available, the role of analysis will become much more important in the cycle of engineering design, fabrication, and test, and these other functions will produce far fewer "surprises."

Computer programs, given the appropriate input data characterizing the structure, can predict the phenomena listed in Table 1 with varying degrees of success. General-purpose programs such as NASTRAN, ANSYS, SAP IV, STARDYNE, MARC, ASKA, and STRUDL can handle many of these problems. Each program has its specialties as well as capabilities common to all the other programs. There are hundreds of other general-purpose programs with features similar to the few listed above and thousands of specialty programs that range from research tools to specialized production programs for particular types of structures. Almost all of them have grown from the classic works of Argyris and Kelsey (1) in Europe and Turner, Clough, Martin, and Topp (2) in the United States.

IN THE BEGINNING THE COMPUTER ARRIVED

In this bicentennial year we all have spent a good deal of time reviewing history. If you live in Philadelphia or Boston, you readily reach back to the origins of the United States of America well over 200 years ago. A similar search for the origins of electronic computing takes you back only 30 years, to 1946. This is an interesting date, but hardly in the same league as 1776. In fact, the period from 1946 to 1956 was spent getting computers to the point where they could produce almost as much useful work as was required by the user to make them do that work.

Finite Elements the Hard Way

As a first-year graduate student at Berkeley in 1959, I remember watching the birth and growth of a computerized structural analysis scheme, which was soon named the "finite-element method." The first continuum mechanics problem (other than a frame or a truss) that I saw solved by computer was the computation of the stress distribution in a cross section of a concrete gravity dam subjected to upstream water pressure and its own deadweight (Figure 1). The cross

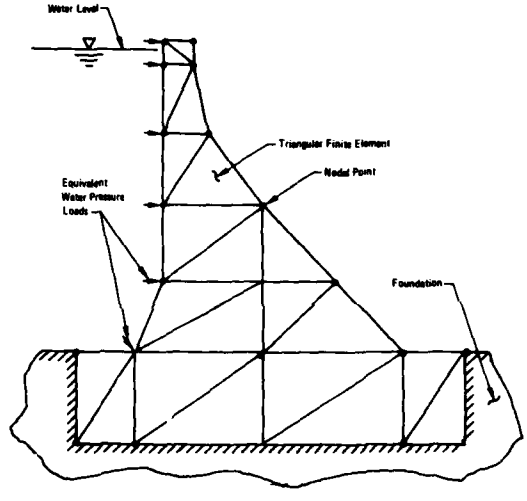


Figure 1—A concrete dam with a finite-element mesh laid upon it

section was divided into imaginary triangular regions (finite elements). The deflections of the corners of the triangles (the nodal points) were computed by solving a system of linear equations. Once the deflections of the nodal points were known, the stresses within the triangular regions were computed and plotted. The computer (an IBM 701 with a matrix operations of software package) was used to solve the linear equations and to multiply various matrixes together. The graduate student who was supporting this research project had his summer's work cut out for him. He first evaluated all the coefficients of the 6 x 6-element submatrixes for all 19 elements. Then he generated a large sparse matrix representing the element connectivity information (probably of order 114 x 42) and punched all of these thousands

of coefficients onto cards. An additional column matrix was generated to represent the loads. Boundary condition information (deflection constraints) was then incorporated in the matrixes. After about a month's work in preparation and data debugging, a successful computer run was made. The stress information (3 number for each of the 19 elements) was printed out along with all of the intermediate results. The graduate student spent the next few days attempting to draw contour plots of what turned out to be very approximate results. (It was later learned that a good representation of this problem would require about 100 triangular elements, far beyond human endurance for data preparation.) The results of all this work were published by Clough in Ref. 3. (Aficionados of the finite-element art will note that this paper contained the incompatible modes isoparametric element, although it wasn't "officially" discovered until 10 years later.

Data Generation Emerges

Something very important was revealed by this study: the computer should be programed to generate the matrixes so that the analyst didn't have to do it. The third leg of the milking stool, the activity of data preparation, data checking, and data display, was going to have to be done by the computer if this new analysis method was going to be successful. Within 1 or 2 years, a computer program* had been written (by E. L. Wilson and I. P. King at Berkeley) that did a substantial part of the data preparation, and plane-stress and plane-strain problems like the dam could be solved in a week. (Now we can do it in a morning.) The analyst had only to enter the coordinates of the nodal points, the nodal point numbers of the elements, and some information on loads, boundary conditions, and materials properties. The powerful new IBM 704 took care of all the rest. After putting the cards for the computer program and the structural data into the card reader, the analyst simply pressed the start button and watched the blinking lights until the results were printed out.

*This classic program, somewhat revised, is described in Ref. 4. The program contains about 300 FORTRAN statements and no subroutines.

The Finite Element Explosion

During the period from 1959 to 1969, computerized structural analysis grew throughout the world from a concept to a widely used production tool. Computer power expanded enormously. By 1969 we had IBM 360/65s, CDC 6600s, and UNIVAC 1108s instead of IBM 701s and 704s. Computer operating systems had become increasingly complex, and one only saw the computer on a tour of the data center—pushing the buttons yourself was forbidden. The mathematical methods for solving equations, extracting eigenvalues, and solving differential equations (the analytical core of the structural analysis process) became reliable and efficient. New, more accurate finite elements were incorporated in the thousands of programs that were written, but our ability to prepare data, check it, and display results hardly changed from the first concepts developed in 1959-1960. The reason for this slow development was that research concentrated on the other two legs of the stool, leaving the production users the nontrivial chore of working with data sets for complex structures that might have over a thousand finite elements. But still, we had a magnificent tool, and computerized structural analysis was having a major impact on the way analysis was done in a production environment.

By 1969, structural analysis in some industries had shifted from using solely hand methods to using predominantly computerized methods. At Boeing, the 707 jet liner (which entered passenger service in 1958) was analyzed entirely with hand methods, using either slide rules or desk calculators. The 747 (which entered passenger service in early 1970) was extensively analyzed with computers.

All of this computerized structural analysis (probably half of a CDC 6600 kept busy around the clock just doing 747 structures work at Boeing) had its benefits. Fatigue and ultimate strength testing on the full-scale airplane uncovered fewer problems, more efficient (but more complex) design concepts were used because they could now be analyzed accurately, and weight was reduced, all while keeping good safety margins. Computerized analysis had eliminated literally thousands of pounds of excess weight from the initial design, which meant greater payload

and range at lower operating costs for the airlines.

The finite-element method was used on the 747 to predict ultimate wing strength, to compute natural frequencies and mode shapes, to design the wing to avoid flutter problems, and to locate areas of potential fatigue cracks. Of course the computer had also been used to study the automatic control system, the landing gear geometry, the propulsion system, the aerodynamics, and a host of other things. Computing had become an essential part of engineering analysis.

The whole aerospace industry, NASA, and the Department of Defense had converted to these methods. Civil engineers had always pioneered with these analysis methods, and they kept right in step with the aerospace activity. About 1970 the automotive industry discovered the finite-element method, and within a few years was developing tools and methods that would make an aerospace structures engineer green with envy. In 1976 American industry will probably spend several hundred million dollars on computerized structural analysis. We've come a long way in 20 years!

THE WORKING ENVIRONMENT OF THE 1976 STRUCTURAL ENGINEER

So far we have described structures and structural analysis, computers and computerized structural analysis, and the growth of the whole process. It now seems appropriate to take stock of where we are in terms of the capabilities available to a structural analyst in a reasonably progressive company—not the ivory tower or research laboratory and not the buggy whip industry.

Standard Analysis Capabilities

Let us start by listing the computerized analysis types which, although requiring a reasonable amount of engineering expertise for use, could be considered off-the-shelf capability available to the structural engineer. These "standard" analysis types are linear static analysis, computation of modes shapes and frequencies, linear spectral (earthquake) analysis, linear transient dynamic analysis, linear buckling, steady-state frequency response studies, and thermal conductivity studies.

This assumes our engineer uses only the following "standard" finite elements: beam and truss elements; plane stress, plane strain, and three-dimensional solids; axisymmetric solids; and plate-bending elements.

Most of our standard analytical capabilities in finite elements are covered by Zienkiewicz (5).

Problems for the Specialist

The next class of analytical problems is presently beyond the reach of our average structural analyst. However, these problems can be handled by specialists with specialized programs. The analytical problems in this class involve:

Large deflections
Nonlinear buckling
Nonlinear elasticity
Inelastic behavior (to a certain extent)
Composite materials
Flutter
Random processes
Structural and control system interactions
Crack propagation and fatigue
Nonlinear dynamics (to a certain extent)
Complex thermal problems

This assumes the expert uses the finite elements listed previously, as well as shell elements, crack elements, and most higher order finite elements.

A fine collection of state-of-the-art papers is presented in Ref. 6.

Actually the analytical methods that the experts now can handle have numerous pitfalls. Problems constantly arise that even the expert must simplify grossly to obtain a solution. Some of these will be described in a subsequent section, to relieve the concerns of the graduate student who fears that finite-element methods have solved all the structural mechanics problems and that there will be no topics for a thesis.

Computing Power

Present-day computer power is substantial, if not particularly easy to use. First off, operating

systems (the software that keeps track of your job and the dozens of other jobs in the machine at the same time) are usually extremely complicated. The control language required for telling the computer what to do to your particular job seems to the novice to be written in Sanskrit and is completely unforgiving of even minor errors. Second, the FORTRAN language, which is used for probably 90% of all scientific programing in the U.S., is not particularly easy to use. Unfortunately, alternate languages are either worse or very limiting. Reference 7 provides a perspective on computing growth and a status report on where we are. Many of the items mentioned in the article will filter down to structural computing in the next few years.

Raw computing power is impressive, however. Large computers can do more than a million multiplications a second and provide you a working space of a million words of high-speed memory (access time of 1 ms or less). All you have to do is figure what you will do with a million 16 digit numbers. Transfer of data to backing storage, in which you have tens of millions of words of additional space, can move at 100,000 16-digit words per second. This raw computing power can be translated into substantial structural problem solution capability. An hour on a big computer (which might cost from $1000 to $2000) can produce a static analysis of a complex structure with 2000 finite elements. A problem this size will require the computer to generate and solve some 6000 to 12 000 sparse linear equations. Linear dynamic analyses are more expensive—typically two to ten times more computations are used than are required for static analyses. For this reason problem sizes are tempered by budgets, and the analyst who can do a good job with half the number of elements is worth his or her weight in gold.

Data Handling by the Engineer

The handling of engineering data—the translation of information from drawings to the computer and then from the computer back to graphs and tables—is the area that occupies most of the time of the engineer doing structural computing. Presently, well over half the total cost of the analysis is in data handling. Unit computing costs keep going down but labor costs continue to go up, so our present data handling methods will need improvement.

For irregular structures, each finite element, each nodal point, and each load still must be defined on separate input cards. Some available computerized input data generation works well for regular (uniform) structural idealizations; automated data generation for regular structures can cut data preparation time by a factor of two to five. (Conversely, with present automated data preparation, there is a tendency to use extra nodal points and finite elements, which means the computer must solve larger problems.) Interactive graphics is used to some extent for data generation, but in my opinion the payoff at present is not great enough to justify the additional cost. In the automobile industry, digitizing tables have been used successfully for transferring data for drawings directly to the computer. The operator traces the finite-element mesh with a special sensing device, pressing a button whenever an (X, Y) coordinate should be entered. For irregular parts and irregular meshes, this technique has considerable merit.

Input data checking has advanced considerably with the advent of low-cost interactive graphics. (Contrary to the previous comments on interactive graphics for data generation, interactive graphics for data checking are of significant benefit.) The average engineer may have access to a Tektronix scope, which displays results on a screen called a storage tube. The device is coupled through an ordinary telephone, using standard voice-grade lines, to a large time-sharing computer and communicates to the computer as though it were a low-speed typewriter-type terminal. The system concept is shown in Figure 2. Figure 3 shows a picture of an actual display of two shells intersecting. Figure 4 shows a blowup of Figure 3 with nodal point numbers added by the computer. The development of low-cost computer graphics is the most significant and cost-effective advance in data handling since we made the computer generate the matrixes instead of doing it ourselves. Because such a large volume of data is required for computerized structural analysis, numerous data errors are made. If they are not caught, the computer either analyzes the

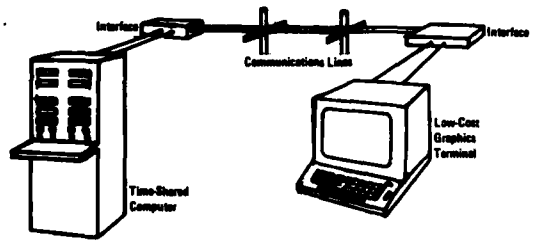


Figure 2—Hardware elements of a low-cost graphics system

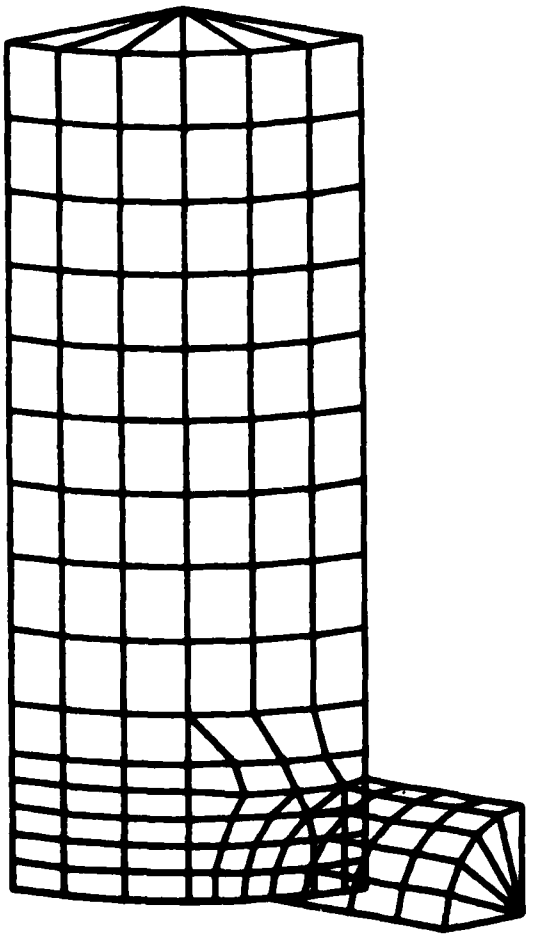


Figure 3—Intersection of two shells (hard-copy plot)

"wrong" structure or "bombs off" after running up a big bill. Anyone who has used interactive computer graphics for structural data checking

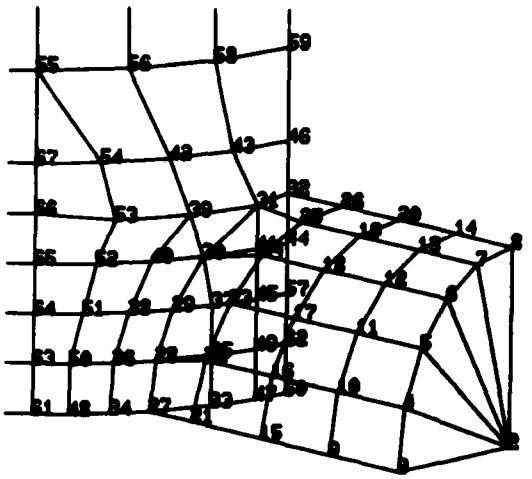


Figure 4—Blowup of shell intersection with node numbers shown

would never go back to scanning unending columns of numbers in search of errors. Flow times and overall costs are cut by a factor of two to five with interactive graphics data checking, regardless of whether the structure is regular or irregular.

At the end of the analysis process the computer spews out great stacks of numbers, and our engineer is left with the task of interpreting them and displaying the significant results. Plotting is used to some extent, but an extensive capability is not generally available. The deflected structure, with the deflections greatly enlarged, can be easily drawn by the computer. This strategy also works well for vibration mode shapes. Contours of stress levels are drawn for planar structures (our concrete dam, for example), but stress contours on surfaces or inside solids are seldom computer-generated. Graphically representing the stresses in beam elements is seldom done. At present we spend far too much time translating information into numbers for the computer and then translating numbers from the computer into usable form.

HOW GREAT IT'S GOING TO BE

The title of this section is flippant, but it reminds me of the promises that are always made about the next computer or display device or operating system or analytical method or compu-

ter program. We are always somewhat disappointed with the reality, as compared to what was advertised. But in spite of our disappointment, as we look back we can see that substantial progress has occurred over the years. Keeping that in mind, let's try to look forward to 1985 and see what our average structural engineer will be using to do his or her analysis.

Analytical Capabilities in 1985

In the area of analytical capability, I do not foresee any dramatic breakthroughs. Rather, it seems reasonable to expect development emphasis on present poorly solved problems, on cleaning up problem areas that could be addressed now but have been ignored, and on improving numerical methods. The category of poorly solved problems includes many structural material characterization questions. Analytically, one can invent all kinds of material behavior. The trick is to generate analytical descriptions that accurately characterize real materials. The problem is difficult because materials data bases seldom contain all the parameters necessary for the mathematical characterization. I expect materials technology people and theoreticians to work together seriously to resolve these problems. Otherwise, by 1985 the materials people will still spend most of their time sticking strain gages on specimens and breaking them, and the structural theoreticians will continue to solve complex academic problems. Steel, aluminum, titanium, etc., need better inelastic dynamic characterization. Materials such as graphite fibers embedded in epoxy, steel fibers embedded in concrete, honeycombs, rubber-like foams, crushable foams, and laminated sheets (and many other materials that are coming into use) are so different from, say, steels that our analytical models (which grew from isotropic assumptions) are frequently inappropriate. Substantial characterization of these new materials is needed, and our computational methods will need considerable work in this area. This work of material characterization will still be active in 1985, but many usable models will be generally available.

Elastic large-deflection work should become routine, as will all areas of linear dynamics. Thermal problems of most types should be straightforward and a temperature-distribution computation will be available as an integrated part of structural software. The analytical methods for nonlinear problems will be either simplified or better packaged. Static inelastic behavior, including unloading or cyclic behavior, should be well in hand. Dynamic inelastic analysis should be reliable for metals, and it is hoped that computation times will not be as exorbitantly expensive as they are now. Crash dynamics work (dynamic inelasticity) should produce energy absorption computations of reasonable accuracy. Many areas of dynamic inelasticity will still be under active investigation, however.

Finite-element libraries (in the computer programs) will contain a somewhat larger collection of elements than is available in most programs today. The hundreds (or thousands) of elements in the literature today will have been sorted into those that are usable, (reliable, simple, accurate, etc.) and those that should die quietly in the archives. The standard elements will produce reasonably accurate stresses and will be simple to use. The element definition data will default to the most common case, with additional features called up by the user as required and specified. One of the elements in the basic set will be a two- and three-dimensional crack element that analytically represents the infinite elastic stress state that the fracture mechanics people enjoy.

The users of these programs will be better educated in setting up analyses and interpreting results. At present, finite-element modeling is an art form, the appreciation of which is only obtained by hard experience. We should be able to teach this black art to students by 1985.

Spanning Engineering Technologies

We will be able to handle the interdisciplinary technologies much better. Already computerized structural analysis has eliminated the left wing tip stress specialist, and it is rapidly eating into the barrier between stress people and dynamics people. We older structural engineers already wish we had paid more attention to those "boring" courses on thermodynamics, electronics,

and hydraulics. The interdisciplinary analysis capability that is developing will require program users to have at least a broad knowledge of these disciplines.

Computer programs from other engineering technologies will communicate with one another inside the computer. This thrust is already being developed at Boeing in the ATLAS program, which combines the disciplines of stress, loads, weights, dynamics, flutter, stability and control, and aerodynamics (Figure 5). A far more ambitious concept is being funded by NASA Langley in its IPAD (Integrated Preliminary Airplane Design) project. The whole spectrum of aerospace technologies will communicate via integrated computer programs and a controlling operating system. By 1985, we should be using integrated programs regularly, because structures will no longer be considered separate entities isolated from their operating environment.

Design (as differentiated from analysis) will be assisted by the computer. Our present approach is to choose a configuration and then analyze it to see if it is satisfactory. If it is satisfactory, the configuration is usually accepted as is; if it is not, changes are made and the analysis performed again. One direction this process will take is to improve our present fledging automated (mathemtaical) design processes and make them more economical and reliable. The other direction is in user-controlled optimization. The analyst/designer will be able to use his computer terminal (work station) to make design changes much more readily than can be done now.

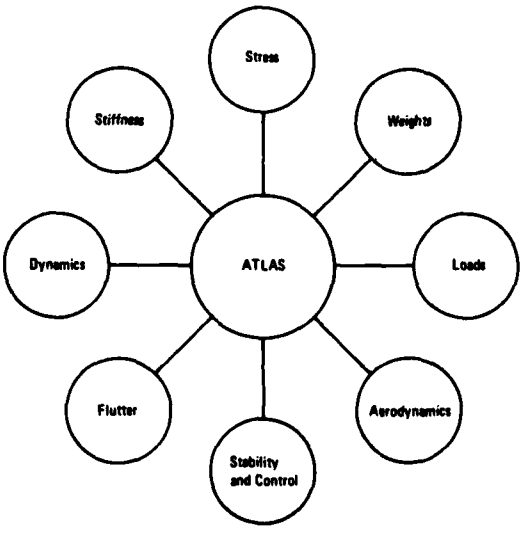


Figure 5—Boeing's integrated aircraft structural analysis system

Numerical Analysis and Software Design

Numerical analysis will continue to have its impact on the internal workings of the programs. We will achieve more reliable results with less computing effort as time goes on. It is to be hoped that by 1985 users will be interested in whether the great quantity of numbers spewed out by the computer program are valid. We will have incorporated the numerical error measuring techniques that are now available. New sparse-matrix techniques will be used, more efficient differential equation solvers for structural problems will be in use, and efficient eigenvalue/eigenvector extraction techniques will be incorporated in most programs.

The design of the structural analysis programs on large computers will change, but the user will not see the differences. Internally there will be software design changes to improve maintainability. Structured programing and top-down design techniques (new programing methods that are now coming into use) will be used in developing new programs. Logical data base design techniques will be used for the input/output handling. Data base managers will simplify communications with other computers and make preprocessing and postprocessing of data easier to program and more readily expandable. Programs will have a good checkpoint/restart capability and good substructuring analysis capability. The user will not have to nurse substructure runs through the computer, as he must do, because these operations will be controlled by a set of preprogramed internal procedures that activate modules and manage file allocations.

Computing Hardware in 1985

The computer hardware side of things could go several ways. Setting aside for the moment the user terminal and its development, let's look at

the number crunching. First off, in the area of "supercomputers," we are already bumping up against the fact that light (and electrical signals) travels at a rate of only 30 cm/ns. This means that a signal from the central processor to a memory cell 3 m away will take 10 ns to arrive and 10 ns to return. Some present-day computers have cycle times of 10 ns. Operations can be speeded up only by making the computer memory and central processing unit more compact. Alternatively, the computer can do several operations at one time, giving the effect of higher speed. Both of these strategies are being implemented, the former requiring new concepts in arithmetic units and memory devices (e.g., bubble memories) and the latter requiring some very clever circuitry and associated programing techniques. In this latter case we have the STAR, ASC, ILLIAC IV, and CRAY, today's supercomputers.

The Super Computers

The STAR, ASC, ILLIAC IV, and CRAY have certain similarities in that they do their best while doing many operations simultaneously. These so-called array processors have internal cycle times about equal to those of standard "maxicomputers," but because they can execute many operations simultaneously, their effective speed can be 10 to 20 times faster than that of a standard maxicomputer. The STAR uses a pipeline concept, which is similar to an old-fashioned bucket brigade putting out a fire. In an ordinary computer, a fireman fills his bucket from the river (reading a word from core storage), runs over and throws the water on the fire (e.g., a multiplication operation), and then runs back with his bucket to the river (result goes back to core storage). In the STAR, the fireman fills a bucket and passes it to his neighbor. While his neighbor is passing the first bucket along (beginning the six or so steps of the multplication process) the first fireman is filling another bucket (reading another word from core storage). He can dip buckets much faster than he can run back and forth to the fire. In the ILLIAC IV, everything happens at once. Sixty-four firemen fill their buckets (each using a different part of the river), run to the fire simultaneously, throw the water on their part of the building simultaneously, and then run back to the river. With apologies to the Control Data Corporation and the Burroughs Corporation, this description of these complex computers will have to suffice. One can see that programing for computer with firemen running all over the place is a substantial undertaking if you want the firemen all to be working efficiently. Reference 8 gives a structural orientation to the problems of programing the ILLIAC IV and some background on the STAR and ASC.

We will struggle for some time implementing efficient finite-element programs on the STAR or ILLIAC IV. However, if we had easy access to the supercomputers for our number crunching activities, we could use them effectively in that activity. If NASA Langley or Lawrence Livermore Labs (the owners of STAR and ILLIAC IV) provided efficient modules for finite-element generation, merging, equation solution, eigenvalue extraction, and solution of differential equations, we could pass our input files to STAR or ILLIAC IV for these specialized services. We would then pick up the solution files and process the information on our own computer. Given the ability of brand X computer to talk to brand Y, this is a possible scenario for 1985. On the other hand, the computer designers are likely to go back to the drawing boards to develop a concept that is more programmable, and by 1985 we may view the array processors as an interesting diversion in computing history.

The Minicomputer Impact

At the other end of the scale we see the minicomputers growing in power and lowering in price. Actually, many present-day minis are not so mini, with some having the power of the magnificent million-dollar IBM 704, but with far greater reliability and a price of under $100,000. Minis are proliferating at a staggering rate, and great piles of money will be wasted putting large structural programs (designed for maxis) onto minis. With a minicomputer, the greatest cost will not be the computer itself but the software development labor. This foolishness of trying to put too much on a mini should sort itself out by 1985. The maxi (here we mean either the conventional or the array

processor computers) and the mini will each have its place in structural computing, talking to one another over high-speed, high-volume telecommunication lines. A maxicomputer would rather be interrupted by a mini transmitting 10,000 words than by a human at a low-speed terminal sending 10 words. The maxi-computer probably uses 10 000 operations of overhead to process either of these interrupts. By 1985 we as users will talk to minis, which in turn will talk to the maxis. The mini will do the intermediate-size processing jobs. The number crunching will be done on large computers, and the minis will store and process our input data and the results, with appropriate prompting from us. Time sharing will be done by the mini because time sharing (the bulk of which is simple text editing) is the wrong thing to do on a maxi and the right thing on a mini.

Presently there is a wide variety of minicomputers on the market, but by 1985 there should be a shakeout similar to what we have seen with the big computers. (In the big computer market, Datamation describes the manufacturers as "IBM and the seven dwarfs." Software for the minis is primitive but will improve substantially. In 1985, the general-purpose software on the big computers will still be superior to that on the minis. We can expect good specialty software on a dedicated mini system—software designed for a special task such as servicing a group of structural engineers. For the structural analyst, the mini should provide data generation, data checking, scanning of results, presentation of results and selective output of files—all in a timesharing mode. Response will appear to be instantaneous for these operations. On some systems, small- to moderate-size analyses will be done on the mini, with the large problems going to the maxi. In other systems, the mainline analysis will be done on the maxi regardless of the size of the problem. The pros and cons of the two approaches (if they exist at all in 1985) will probably be argued heatedly.

The Work Station

We started our description of future computing hardware with the big computers. We then moved down in size and closer to the user and described the minicomputer. We now arrive at the engineer's work station, the place where all the work gets done. In the last year or two we have seen the advent of the "smart" terminal. By 1985 we will use work stations, the logical extension of the smart terminal. The work station will be controlled by a small minicomputer or a microprocessor.

The microprocessor has just arrived on the scene and is essentially a computer on a chip. These little gadgets, which now sell for less than $100, are tiny computers that process single bits rather than words. They can be designed or programed for special applications such as translating from one teleprocessing communications protocol to another. In other words, a microprocessor can be programed to allow one type of computer to talk to another even though they transmit characters in different codes. Microprocessors can take compressed bit combinations and unscramble them into plotter hardware commands. They could also be developed to perform three-dimensional rotations of graphical displays.

Each work station will have an interactive graphics display screen, which will probably combine the best aspects of our present-day storage tubes (such as the Tektronix scope) and the refresh scope. Possibly this will be done by the plasma panel, a vector graphics device, now under development, that relies on digital rather than analog circuitry. Driving this display will be the micro or mini in a box in the console. This little computer will handle many of the graphical jobs that now bother a big computer—clipping, zooming, rotation, and refreshing the screen.

The local computer will also handle some of the file editing activities, because the work station will have a modest amount of high-speed memory and the ability to read and write data from tape cassettes or floppy discs. (Floppy discs look like 7-in. diameter phonograph records, hold tens of thousands of words of information, and cost about $5.) In addition, the local computer will handle the communications protocol when the user dials up (via telephone) the central structural data-processing mini.

User input may still be via typewriter, although much work will be done with a light pen or joy stick. Displays will flash on the screen almost instantly and the central mini will constantly prompt the engineer. Input via a digitizing table (or something similar) will be available, as well as

a device to produced hard-copy plots of the images on the screen. This powerful work station will probably cost $10 000 to $20 000 in 1976 dollars. The local mini/micro computer that drives the work station will cost less than the harmoniously designed leatherette operator's chair that the buyer may choose as an option.

Operational Capabilities

The work station described above, when connected to the central mini processor, will provide powerful data handling ability so that an anlysis can be done quickly and accurately. Data generation will come in many forms. Data will be generated via a digitizing tablet, a wide variety of mesh-generating routines, "building" of the structure on the screen with a light pen, or calling up previously built structural components and "reusing" these parts. Data checking will be done graphically, with complete rotation, zooming, substructure display, and slicing. Numerical data checks will also be presented.

Once a structural data set looks good, the central mini passes it to the maxi computer for the mainline solution. When that has been completed and the files are retrieved by the mini, the engineer can examine the results. The user will call up deflected shapes, time history plots, stress contours on external surfaces and on slices through the structure, and maximum stress plots. Printouts will be done selectively, leaving the bulk of the data on storage files. Figure 6 shows one possible form that this concept of distributed computing can take.

The operational cost of one of these work stations will be about $25 to $50 an hour (1976 dollars), roughly equal to the "all-up" cost of an engineer. The engineer will do his work several times faster than he now can, making the work station very economical indeed. In addition, the reduction of errors will mean fewer bad computer runs. Results will be presented more intelligently, and dubious results will no longer remain hidden in piles of printout. The key benefits have already been stated—the reduction in flow time will help tight schedules enormously. The engineer will spend his time doing engineering design and analysis, not data generation, manipulation of control cards, and plotting data by hand.

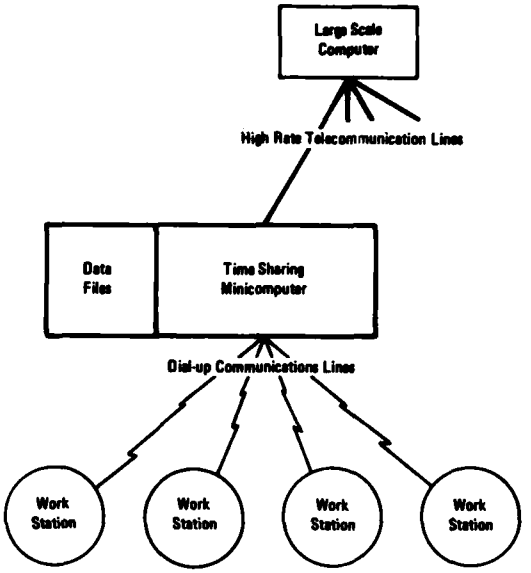


Figure 6—Schematic of a distributed computing concept

THE IMPACT ON FUTURE NAVAL SYSTEMS

Computing advances to come will significantly affect naval design and construction. Computer-aided manufacturing techniques are already affecting construction procedures. Parts are manufactured with computer-controlled machines. Preparation of drawings, parts lists, schedules, job assignments, etc., are all being done by computer. Some important design impacts of computing on naval structures are described in the following paragraphs.

Design Impact

Computerized structural analysis and design is benefiting naval structures today. The old process of design, construct, test, modify, retest, and produce (Figure 7A) is slow and expensive because of the time required for building and modifying structures. Computerized analysis has entered the picture (Figure 7B) and is reducing costs and flow time by helping eliminate costly redesign cycles in the prototype.

Computerized preliminary design will be used for navy hydrofoils. Boeing is developing a user-controlled computing system for the Naval Ship

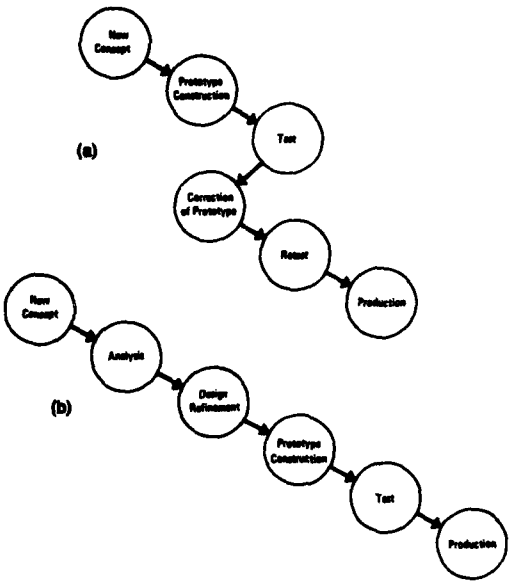


Figure 7—System development with (a) and without (b) computerized analysis

Research and Development Laboratory which will assist design in virtually all aspects of hydrofoil preliminary design. (The program is called HANDE, Hydrofoil ANalysis and DEsign.) A data base of previous design experience coupled with computing modules for propulsion, weights, control, structures, hydrodynamics, etc., will provide the designer with the capability to specify mission criteria (range, payload, speed, etc.) and develop designs that meet these criteria. One can expect that eventually different levels of analysis will be incorporated in programs like this, starting from approximate preliminary design computations and graduating to detailed analyses as the design firms up.

New materials are being considered for aircraft and helicopter components—materials that are lighter and stronger than the aluminum they replace. As new concepts are developed, a thorough structural analysis will predict the behavior before costly and time-consuming fabrication and testing is begun.

New concepts in crash protection of crew members in helicopters and airplanes are being developed. New design procedures must be implemented when the concept of crash energy management is introduced as a key design criteria. Structural elements will be chosen on the basis of their energy absorption and dynamic behavior instead of overall strength and stiffness. The computing tools for designing for crash energy management are being developed. Too often in the past we have striven to protect the structure and have thereby increased the injuries to the occupants. Accurate prediction of the behavior of new energy-absorbing devices coupled with thorough testing will produce aircraft structures that protect the occupants in a severe crash.

Futuristic Concepts

Computing technology will have impacts upon our activities that are essentially unpredictable at this time. Three somewhat speculative concepts are described below.

The microprocessor may find its way into the performance monitoring of a naval aircraft structure. Suppose a fighter plane was fitted with thousands of microprocessors, each of which had the job of detecting structural flaws and crack development at a specific point on the structure. Each microprocessor would continuously sense the local strain and compare it to the measured accelerations. After a simple filtering of the sensor data, some comparisons would be done to check that measurements were within the predicted bounds. Any deviation would produce a warning signal. We would have incipient failure detection, which would be a significant augmentation to regular inspection processes.

In the structural analysis business we have thousands of nodal points and finite elements, each interacting with its neighbors. Suppose a new form of the analog computer was devised with each nodal point and each element represented by a special microprocessor. We could have different microprocessors designed to simulate different elements: shells, solids, beams, etc. A structural analysis would consist of defining the connectivity of the microprocessors and then starting this array of specialized microprocessors iterating away to a solution.

In the engineer's office we would see on the wall a large computing panel instead of a blackboard. Input to the computer would be done

on the computing screen, and the computer in turn would produce graphical and numerical displays. The computing screen would be a sophisticated version of the work station we spoke of earlier. The screen would sense a pointerlike device, which would take the place of chalk. At the engineer's desk, a table that would function like a typewriter would provide assistance in report editing, accessing of reports and archives, and other basic word-processing functions. Calling up all of the Office of Naval Research contractors reports on crack propagation would be done at the desk tablet. So much of our time is spent with words, it seems only reasonable that the computer will be helping.

CONCLUSIONS

We have described a variety of structures, different types of structural analysis, and several types of computer hardware. We have reviewed the profession's present capability and what can be expected in the future. Throughout this paper we saw highlighted the reasons for doing computerized analysis:

- Reduction of costs by evaluating performance before prototype construction and testing
- Supplementation of safety tests by simulating a broader range of operating environments
- Reduction of flow time by minimizing retrofits to prototypes
- Evaluation of new design concepts which previously would have been discarded because the required analyses were impossible to do.

Computerized structural analysis and design has become an integral part of naval structural systems development and will be even more important in the future.

REFERENCES

1. J. H. Argyris, "Energy Theorems and Structural Analysis: Part 1, General Theory," *Aircraft Eng.* **26** (Oct., Nov. 1954\$, **27** (Feb., Mar., Apr., May, 1955). J. H. Argyris and S. Kelsey, "Part 2, Applications to Upper and Lower Limits of St. Venant Torsion Constant," *Aircraft Eng.* **26** (Dec. 1954).
2. M. J. Turner et al., "Stiffness and Deflection Analysis of Complex Structures," *J. Aeron. Sci.* (now the AIAA Journal) **23** (9) (Sept. 1956).
3. R. W. Clough, "The Finite Element Method in Plane Stress Analysis," *Proc. ASCE 2nd Conference on Electronic Computation*, Pittsburg, Pa., Sept. 1960.
4. E. L. Wilson, "Finite Element Analysis of Two Dimensional Structures." Rep. No. 63-2, University of Calif., Berkeley, Dep. of Civil Engineering, Structural Engineering Laboratory, June 1963.
5. O. C. Zienkiewicz, *The Finite Element Method in Engineering Science*, McGraw-Hill, London, 1972.
6. *Proceedings of the International Conference on Computational Methods in Nonlinear Mechanics, University of Texas at Austin*, J. T. Oden, et al., eds., Sept. 1974.
7. Werner L. Frank, "The Second Half of the Computer Age," *Datamation* May 1976, 91–100.
8. E. I. Field, S. E. Johnson, and H. Stralberg, "Software Development Utilizing Parallel Processing," in *Structural Mechanics Computer Programs* W. Pilkey, K. Saczalski, and H. Schaeffer, eds., University Press of Virginia, Charlottesville, 1974.

OCEAN SCIENCE AND TECHNOLOGY



Allan R. Robinson, Gordon McKay Professor of Geophysical Fluid Dynamics in Harvard University's Division of Engineering and Applied Physics, is Chairman of the University's Committee on Oceanography and former Director of the Harvard Center for Earth and Planetary Physics. He is Cochairman of the U.S. POLYMODE Organizing Committee and U.S. National Chairman of the U.S.-U.S.S.R. POLYMODE Organizing Committee. The latter is responsible for planning and carrying out a fully international oceanographic field experiment to determine the mechanics, geographic distribution, and physical role of eddies in the open ocean. Dr. Robinson's research is in the physics of large-scale ocean currents and the dynamics of their variabilities (eddies). His teaching includes problems of the atmosphere, and other planetary fluids. He has written numerous articles published in national and international journals. He received the A.B. (magna cum laude), M.S., and Ph.D. degrees from Harvard University. He is Cochairman of the SCOR Working Group 34 on Internal Dynamics of the Sea, is on the editorial board of the journal *Dynamics of Atmospheres and Oceans*, and is an associate in Physical Oceanography at the Woods Hole Oceanographic Institution. He was a Guggenheim Fellow at Cambridge University in 1972 and has been visiting professor at several Indian universities.

NUMERICAL MODELING AND GLOBAL OCEAN FORECASTING

Allan R. Robinson

Center for Earth and Planetary Physics
Harvard University
Cambridge, Mass.

The sea is a classical fluid system. That is, the basic physical laws governing its dynamic behavior are those of classical hydrodynamics and thermodynamics. They include the conservation of mass, momentum and energy. The equations expressing these conservations in a form appropriate to a continuum, together with the equation of state of seawater and a statement of conservation of the combined specific density of all the dissolved salts (salinity) that influence the mass density of the water constitute a system of model equations formulated so as to be complete and adequate to describe the evolution of the state of the system.

The fundamental dynamical problem is the so-called *initial-boundary value problem*. Given are the shape of the ocean basins, a description of the surface and body forces, and a knowledge of the distribution of state variables (velocity, temperature, salinity, density, and pressure) at some initial time. What is the distribution of state variables at some arbitrary future time? In principle, their evolution is governed by the basic model equations and is obtained by their forward integration in time. In practice and in principle, however, we now know that serious difficulties exist in the study of fluid behavior and in the prediction of fluid motions approached directly via the fundamental initial-boundary value problem. These difficulties can be expressed in a variety of mathematical and physical ways.

The system of basic equations is a complex one, and intrinsically nonlinear. Exact solutions are virtually nonexistent. Although approximate solutions obtained by analytical methods have provided important insights, they are difficult to obtain, and usually of restricted validity and little generality. Modern high-speed computers provide the means of obtaining solutions to numerical model equations analogous to hydrothermodynamical equations that are otherwise inaccessible. In recounting the early history of the impact of computers in meteorology, Charnev [1] recalls that in 1946 the famous mathematician John von Neumann remarked that "the success of mathematics with the linear differential equations of electrodynamics and quantum mechanics had concealed its failure with the nonlinear differential equations of hydrodynamics, elasticity, and general relativitiy. . . . To him meteorology was *par excellence* the applied branch of mathematics and physics that stood the most to gain from high-speed computation." Today numerical weather prediction is routine; numerical modeling is an invaluable research tool in contributing to research progress in atmospheric and oceanic dynamics, but not via the brute-force approach (i.e., not via the direct integration of the numerical

analog to the initial-boundary value problem of the basic model equations).

The basic model equations are general, contain a wealth of distinct phenomena, and are applicable to many special circumstances of fluid flow. For example, they contain solutions corresponding to acoustic waves, surface and internal gravity waves, and bores. They can describe the breaking of waves on beaches, the wakes of ships and fish, convective overturning in deep heated trenches and the massive coursing and transient meandering of the Gulf Stream. It is neither feasible nor desirable to obtain solutions to the basic equations for the global ocean which include a description of all the phenomena that are occurring, simultaneously, in the oceans. Phenomena that are not of interest may be removed in various ways from the basic model equations. Terms in the basic equations can be altered or removed in such a way that a less general system of equations results. For example, if compressibility effects are removed from the mass conservation equation, the simplified model has no sound waves. This is an example of a so-called filtering approximation. Small-scale and/or high-frequency phenomena can be suppressed by some process of averaging in space and/or time, applied to the basic model equations. Averaging processes again result in altered equations, in some sense simplified, but in another sense complicated by residual effects of the smaller scales, which remain in the averaged equations because of the nonlinearity of the basic equations. Even if filtering and averaging were not desirable for removing the complexity of the description, uninteresting phenomena, and fine scales from larger scale initial value problems, it would be necessary because of the impossibility of describing the initial state of the sea including the finer scale structure and phenomena. In fact, the obtainment of initial data on the larger scale remains a formidable problem for numerical ocean modeling and forecasting.

The ocean is a vast system subject to a number of variable forces and local effects that produce features such as whitecapping and wavebreaking, which require filtering or averaging out of larger scale calculations. There is, however, a very fundamental necessity for averaging, which is due to the nonlinearity of the basic equations. It is manifest in very much simpler situations such as the flow of water in a pipe or the rapid draining of a basin or tub. Unless the situation is such that the flow is very slow, the motion is turbulent. Even though on the average there is a steady rate of transport of water through the pipe or down the drain, the actual fluid motion has a rapidly varying small-scale structure which appears highly random in character. Turbulence occurs when the pure number (nondimensional Reynolds Number), formed by taking the ratio of the product of typical speed times the length scale of the flow to the molecular viscosity of the fluid, becomes moderately large. Turbulent motions are not directly forced by high-frequency small-scale external forces. Turbulence occurs under uniform steady forcing spontaneously generated by internal processes in the fluid.

The phenomenon of turbulence is related to instabilty in fluid systems. A smoothly varying large-scale flow evolving consistently with the basic model equations in any real circumstance is subject to some degree of small jiggling or perturbation. If the system were dynamically stable, such small perturbations would never alter the large-scale smooth flow by a noticeable amount. This is not, however, the case. Such perturbations will cause the flow to break down into a spectrum of intermediate and smaller scales. Although the complicated turbulent flow that results is itself consistent with the basic model equations, it is neither feasible nor interesting to describe it by a direct time integration of the basic equations. Averaging is required. At first suggested by Reynolds in 1894 [2], the total flow can be regarded as composed of two parts, the relatively slowly varying large-scale component of interest, plus a turbulent fluctuation which vanishes on averaging. But, as mentioned above, the basic model equations upon averaging do not simply become equations governing the average flow (average state variables).

Products of fluctuations occur, and if the fluctuations are correlated the averaged products do not vanish. Physically, the residual effects of the turbulent fluctuations can be of great importance and can influence the evolution of the large-scale average flow. They represent transports of momentum and heat, which contribute to the average momentum and energy conservations.

Mathematically, the averaged equations do not represent a complete set of model equations to describe the evolution of the state of the average system unless the residual effects have been expressed in terms of the averaged fields or otherwise specified. The expression or specification adopted must correctly represent the physical influence of the residual terms on the average fields. This problem is very difficult, even in the simplest circumstances of turbulent flow. It is known as the closure problem. The ocean is very large, and its physics and dynamics are influenced by special geophysical factors such as the rotation of the Earth and the vertical stratifications of the distributions of temperature, salt, and density. Turbulence occurs over many scales and in several physical forms, some of which are common to laboratory and smaller scale flows, and some of which are peculiarly geophysical or oceanic. Almost all forms are poorly understood, and the expression or specification of the effects of smaller scale high-frequency flow components in the averaged equations is speculative at best and often unsatisfactory. This represents a major problem area for ocean modeling in general and numerical modeling in particular.

An objective of turbulence theorists is to derive from the basic model equations the properties of the turbulence and thus the expression for the closure or parameterization of the turbulent effects in the mean flow equations. Traditionally, however, scientists and engineers, such as oceanographers and hydraulicists, interested in large-scale flows have by necessity taken a pragmatic approach to parameterization. They have used empirical statements or "whole-cloth" hypotheses. Indeed, in their famous monograph *The Oceans* Sverdrup et al. [3] remark that "The Navier Stokes equations [basic model equations] find, therefore, no application to oceanographic problems and have been mentioned here only as an approach to the problem of fluid resistance and for the sake of completeness." Modern oceanographic theorists and modelers share many basic problems with meteorologists and fluid dynamicists. The importance of this commonalty led to the identification in the early 1960s of the discipline of *geophysical fluid dynamics*. The rapid progress made in ocean modeling in the past two decades has advanced the subject to a position where we may anticipate an increasingly effective and sophisticated exchange of ideas and techniques with other geophysical fluid dynamicists. In particular, some real progress now appears to be occurring in some areas of turbulence theory that can be expected to affect ocean modeling [4,5]. Ocean models can also be expected to deal in an increasingly realistic fashion with the larger scale geophysically constrained range of turbulence with its special characteristics, such as the generation by turbulence of larger scales [6].

The fundamental dynamical problem and basic model for ocean dynamics is the initial boundary value problem of hydrothermodynamics. This model problem, as well as applicable and tractable simplified model problems derived from it by filtering approximations and parameterization hypotheses, are formulated in terms of partial differential equations. The solutions for the state variables representing the ocean circulation implied by these equations are continuous functions. That is to say, in principle one can solve for the velocity, temperature, etc., at every point in the ocean as a continuous function of time. In order to exploit the power of computer techniques to solve model problems, the models must be discretized, and the derivatives replaced by differences over a specified distance. In general, this discretization (or finite differencing) is done in all three spatial directions and in time. The specified distance in space is called the *grid interval* and that in time the *time step*. Thus, in the numerical model, the state variables are known only at certain points in the ocean and at certain instants, and those values are taken to represent the actual situation over the differencing distance.

Figure 1a shows how a continuous vertical profile of horizontal velocity appears as represented by six grid intervals throughout the depth of the ocean. In this case, the grid intervals are shorter in the upper ocean because the velocity profile is known to have more structure there than in the deep ocean. There are important considerations associated with the choice of differencing scheme used in the formulation of the numerical model related to a given continuous model. These include the requirement of making sure that the numerical model is the physical analog of the continuous model as well as making sure the model can be solved efficiently on the

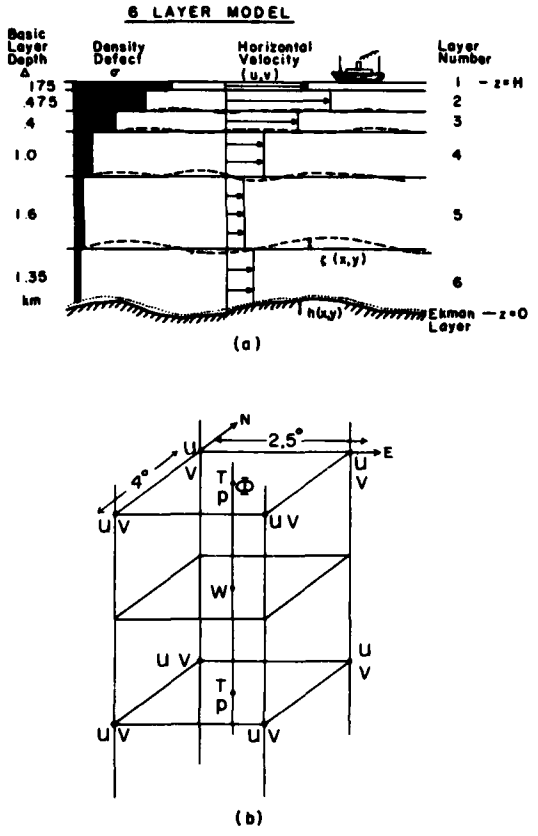


Figure 1—(a) Model vertical structure; density and mean layer depth are specified [17]; (b) Vertical and horizontal placement of grid points [8].

computer. Figure 1b shows a typical segment of a three-dimensional spatial lattice of grid points for a world ocean model. Values of horizontal velocity components u, v, of temperature and pressure T, p, and of vertical velocity w are calculated at three different types of staggered grid points. The coarseness or fineness of the grid intervals and time steps is referred to as the *resolution*. As the resolution becomes finer and finer, the numerical model solution must converge to the continuous model solution, in the sense that at the computed points the difference between the two solutions can be made arbitrarily small. The choice of grid intervals and time steps depends not only on the degree of accuracy desired, but also on internal consistency requirements of the numerical model mathematics. Bryan [9] describes in detail a numerical model that has been used extensively in numerical ocean circulation studies, and Kreiss [10] provides a contextual summary.

We have seen above that for both geophysical and physical reasons the brute-force approach via direct integration of the basic model equations is neither feasible nor desirable. Even if this were not the case, it would be impossible to proceed directly because of computing machine limitations. This is true even in the light of rapid evolution of computer technology and the introduction of fifth-generation machines [11]. Furthermore, a general-purpose direct approach for global ocean forecasting on many scales of interest in terms of the initial boundary value problems associated with a filtered and parameterized simplified model problem is also not now feasible. Nor does it appear that it will be so in the foreseeable future, in part because of machine limitations, but also because of observational data limitations and inadequacies in aspects of contemporary physical modeling.

The construction of a numerical ocean model is a scientific art. The choice of the simplified model equations, the analog numerical model, the resolution and the domain of integration (the volume of the ocean and duration of time for which the computation is carried out) are interrelated questions. Decisions depend on the special purpose for which the model is designed, as well as practical considerations of computer capabilities and the cost of long computations. The choice of resolution depends on the scales of motion required to be resolved and the physical phenomena that will then be explicitly included and described in the model results. A schematic spectral representation of some important types of oceanic processes is shown in Figure 2. Processes that cannot be resolved because of the coarseness of the resolution are called *subgridscale* processes, and their effect on resolved scales of motion must be parameterized if the physical interacton between the resolved scales and the subgridscale is not negligible. In general, because of practical computational constraints, the finer the resolution, the smaller must be the choice of domain. If the domain is the world ocean, the resolution must necessarily be very coarse. If the domain is only a piece of the ocean, then another problem arises. If the boundary, or a segment of the boundary, of

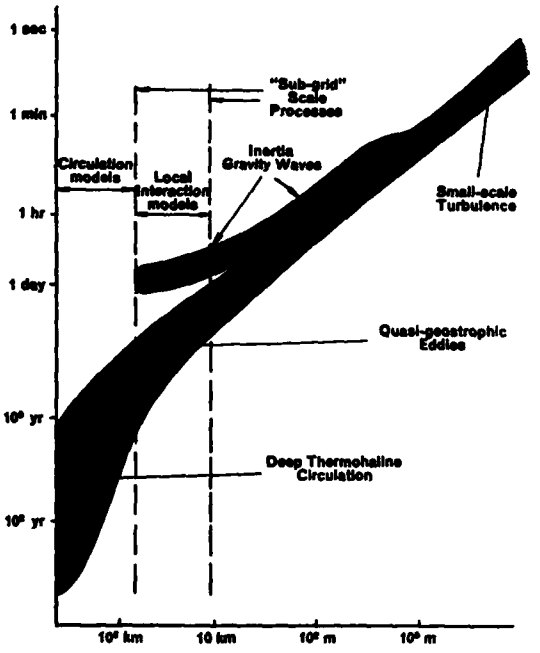


Figure 2—Principal scales of motion in the ocean. Note that in contrast to the atmosphere, quasi-geostrophic eddies must be classed as subgridscale processes for models with global resolution [12].

the domain chosen is not part of the solid ocean basin or continental margin, what is the proper boundary connection between the region of the ocean modeled and the external ocean? This is the problem of the *open boundary condition*, or the parameterization of the relationship of the resolved scales to the larger scale flow. Obviously, there are trade-offs in the construction of heuristic, special-purpose models.

Numerical modeling of the ocean has gained considerable impetus and insight from numerical modeling of the atmosphere, which is in a considerably advanced stage of development and application. The historical development has been reviewed by Phillips [13], and recent advances have been summarized by Haltiner and Williams [14]. The advanced status of atmospheric modeling over oceanic modeling is attributable to three causes: (a) the development of meteorology from a naturalistic discipline into a mathematical physical science earlier than oceanography; (b) the vastly greater data base available for the Earth's atmosphere, compared to that available for the global ocean; and (c) the practical necessity for weather prediction and the economic consequences of improving forecasting, if possible, by basing it on dynamical principles. The first unsuccessful attempt at numerical prediction was made, without the aid of a computer, in 1920 [15]. Computer modeling began shortly after the end of World War II. Numerical modeling is now vigorously pursued both for the purpose of forecasting the weather and longer term future states of the atmosphere, and for the purpose of exploring fundamental dynamical processes in the air.

The failure of Richardson's pioneering hand calculation can be interpreted in terms of the rapid amplification of internal-gravity and acoustic waves implicit in his initial state data. This difficulty was removed and the first successful forecasts carried out in 1949 [16] by using highly simplified model equations, which filtered out all atmospheric phenomena except weather (the large high- and low-pressure systems that circle the globe at midlatitudes). It took 24 hrs. of computing time to perform a 24-hr. forecast. The weather-only filtered model has since been replaced by a less severely filtered and averaged model (primitive equation model) which describes weather phenomena more accurately but also contains additional phenomena, including internal gravity waves. This approach is successful because of the so-called *initialization* of the observational data used for starting the forecast. The observed wind and pressure fields are modified somewhat so as to obey a balance-of-forces relationship that governs weather scale phenomena. This removes spurious waves, which would appear to be present due to inaccuracies in the initial observations. The U.S. National Weather Service now carries out 12-hr forecasts routinely; they require half an hour of computing time. Experimental forecasts have been carried out with some success for periods of one to two weeks. Haltiner and Williams [14] summarize the present situation as follows:

> "Numerical integration of the atmospheric equations as an initial value problem is the primary basis for the prediction of synoptic-scale disturbances for periods between 12 hours and perhaps five days, and, in addition, to some extent for smaller scales and much longer

periods. The sources of error in such prediction are a consequence of a) gaps and errors in the data which make up the initial state, b) limitations in the objective analysis-initialization schemes which are applied to the data, c) truncation errors in numerical integration schemes, d) incomplete representation of the many complicated dynamical processes at work in the atmosphere, and, finally, e) limitations imposed by the predictability of the atmosphere."

Their final point (e) is noteworthy. Meteorologists have practical and scientific reasons for attempting to extend their weather forecasts for longer and longer periods and for attempting to predict the very long term evolution of the atmosphere. The latter involves the necessity of coupling the atmospheric model to an oceanic model that is physically required for the study of climatic changes. However, there appear to be practical and theoretical limits of predictability [17] for a nonlinear fluid system such as the atmosphere or ocean. These limits of predictability are related to the instability phenomenon. In practice the limit is associated with the inevitable errors in the smaller scale descriptions of the initial state (residual observational noise). In principle, the limit is believed to be associated with inaccuracies in the physical content of the hydrothermal dynamical equation at the very smallest scales and an intrinsic lack of strict determinacy in the basic model equations. Random fluctuations on the scale of the molecular mean free path ($\sim 10^{-5}$ mm at sea level) amplify and escalate in scale, ultimately introducing indeterminacy into scales of practical interest. Charney [1] estimates that in the atmosphere errors in scale of only 1 mm progress to scales of 100 km in less than one day, and thence to scales of weather phenomena in a week or two. Uncertainty propgates from smaller to larger scales, even though at the smaller scales turbulence phenomena propagate energy in the reverse direction. The present best overall estimates for the predictability of specific weather patterns in the atmosphere is at most a few weeks. Progress in longer term prognosis is not precluded, but it must be expected to contain elements of a statistical character in its formulation, in contrast to the strictly deterministic approach.

Looking far into the future of ocean forecasting, these remarks have a twofold consequence. Similar considerations will ultimately limit the predictability of the oceanic state, although numerical estimates will differ because of the differences between the atmosphere and the sea in constitution, configuration, and specific dynamics. Moreover, in those circumstances in which atmospheric winds are directly forcing the oceanic motions, deterministic forecast of the ocean's response depends on the ability first to forecast the weather.

Numerical modeling of the ocean circulation was pioneered in the early 1960s by Sarkisyan [18] and Bryan [19]. It is now a flourishing activity which, hand in hand with analytical modeling and new observations and experiments at sea, is rapidly advancing our understanding of the physics and dynamics of the ocean. Rapid developments and the need for directional insights prompted a first major review conference, held in 1972 under the auspices of the National Academy of Sciences and with the support of the Office of Naval Research [7]. A recent summary review of numerical modeling is provided by Pond and Bryan [20] and in-depth reviews by special topics will appear in the near future in Volume VI of *The Sea* [21].

In contrast to meteorology, the introduction of numerical models in oceanography was primarily for the purpose of studying fundamental dynamical processes, not for forecasting. Because of limited oceanic data and our rudimentary knowledge of ocean dynamical processes, idealized models illustrating ocean processes in the simplest conceivable circumstances have been explored. Pond and Bryan [20] refer to these as *mechanistic* models, in contrast to *simulation* models, which model factors such as geometry of basins in greater detail and are intended to produce results that can be compared more directly with primary oceanic observations. Numerical ocean models that are formulated in terms of some version of an initial boundary value problem are termed *prognostic*. A highly heuristic, more simplified class of models, which discards considerable physics, is termed *diagnostic*. In diagnostic models momentum and mass conservation are retained, but thermodynamics is ignored, being replaced by the specification of some state variables directly from observations or assumptions. The

velocity and pressure distributions are computed from a specified field of density. The density field itself may have been derived from specification of the fields of temperature and salinity. Considerable care is required in diagnostic calculation to assure the overall physical consistency of the total set of computed and specified field variables. The development of diagnostic numerical models in oceanography is associated with the great difficulty of obtaining direct measurements of ocean currents; the classical data base consists almost entirely of hydrographic observation (temperature, salinity, and content of certain chemicals).

The initiation of numerical ocean modeling and its early application and progress were related to significant advances made in analytical modeling in the preceding decade, and simultaneously thereafter. During this period, many of the major features that appeared in the synthesis of the classical observational data were rationalized in dynamical terms. These include the existence and structure of the major circulatory gyres, the major intense currents, and the main thermocline (the strong vertical gradient of temperature that exists permanently in the upper part of much of the world ocean). Figure 3 shows the surface circulation of the global ocean. Notice that considerable symmetry exists across the equator and that the major ocean basins (North Atlantic, South Atlantic, North Pacific, South Pacific, and Indian) have similar patterns of flow. The main feature in each is a large *subtropical gyre* that includes an intense current (Gulf Stream, Kuroshio, etc.) at the western edge of the basin. That this pattern of circulation is not superficial is demonstrated in Figure 4, in which the subtropical gyre and the Gulf Stream of the North Atlantic appear in terms of the pattern of transport streamlines. The transport is defined as the vertical integral (or average) of the horizontal current. A schematic representation of the generalizable features of a typical ocean basin is shown in Figure 5.

An idealized physical model, which takes into account in very simplified forms the pattern of surface wind stress, continental boundaries, and the sphericity and spinning of the earth, reproduces the gross features of the subtropical gyre [23]. Figure 6 shows an early numerical model computation of the transport streamlines corres-

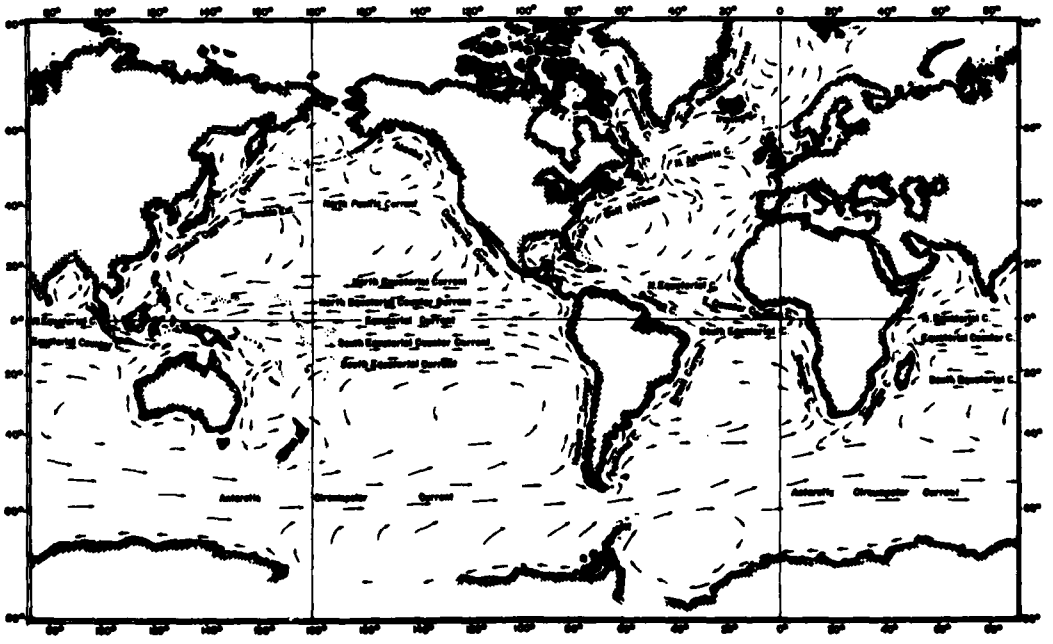


Figure 3—The main surface currents of the world ocean [22].

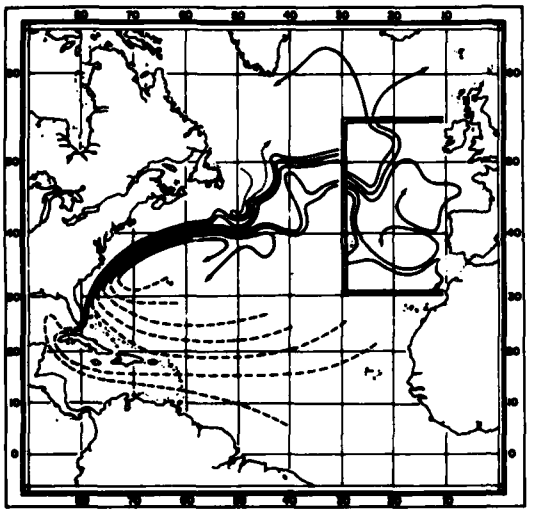


Figure 4—Mass transport streamlines in the North Atlantic [22].

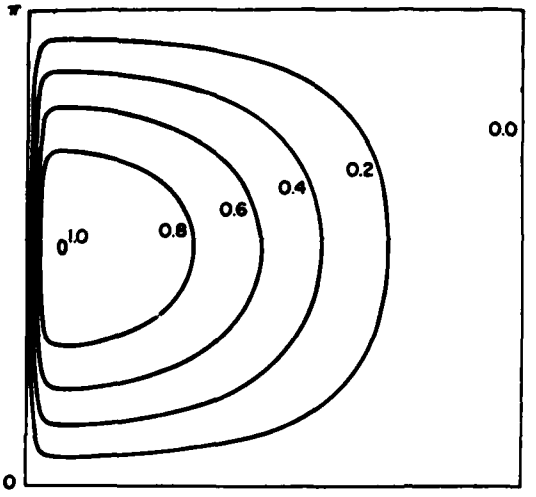


Figure 6—Nondimensional stream function for an idealized midlatitude basin calculation [24].

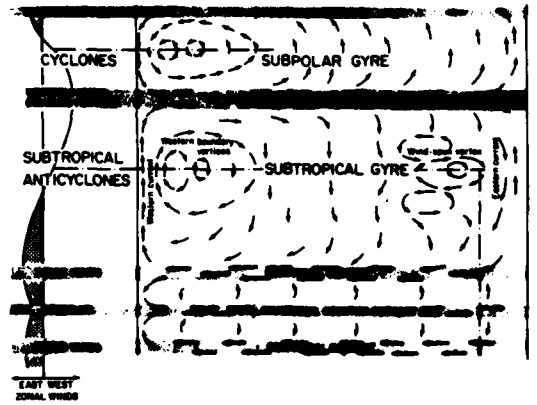


Figure 5—Schematic representation of the main current systems of a typical ocean basin [22].

ponding to such a model. The pattern is indistinguishable from that which results from the continuous function solution obtained from the analagous analytical model. The isolation of the gyre is accomplished on the eastern and western sides by solid boundaries corresponding to idealized American, European, and African contiental margins, and on the northern and southern boundaries isolation is accomplished because the assumed forcing wind pattern (Fig. 5) results in zero transport at the northern and southern bounding latitudes. The ocean is assumed to be of uniform density (barotropic), turbulence is parameterized in terms of a frictional drag proportional to the speed of the flow, and the horizontal grid pattern is 40 x 40.

The interpretation of model results, including their relationship to real ocean phenomena, requires considerable care and discussion. This is so not only because of the complexity of nature and the sparseness of field data, but also because of the number of assumptions necessary to define an interesting and tractable model problem. A model problem requires many parameters in its definition that must be assigned specific numerical values for computer computation. It is sometimes difficult to ascertain the causal factors of even gross features of the computed flow. Model verification involves both the attribution of computed effects to controlling model parameters and a study of the sensitivity of the effect to changes in the relevant parameters. Such a sensitivity study is necessary for evaluating the credibility of the physical basis for the assumptions that define the parameters of the model problem. As an example, Figure 7 shows the results of mechanistic numerical model studies of the barotropic subtropical gyre model. The primary motivation for these studies was to model nonlinear momentum transport in the boundary current more physically

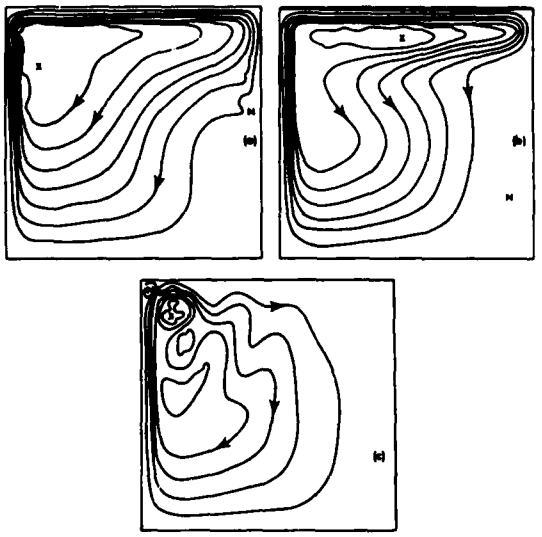


Figure 7—Transport stream function of a homogenous wind-driven ocean: (a) bottom friction, free slip on side walls; (b) lateral friction, free slip on side walls; (c) lateral friction, no slip on side walls [20].

realistically than in the simpler flow of Figure 6. In the original results, it appeared that the different flow patterns might be related to the parameterization of subgridscale turbulent momentum transport. Two assumptions were explored: (a) a drag law as in the model of Figure 6 and (b) a so-called *eddy viscosity* assumption, in which the presence of turbulence is modeled by increasing the coefficient of internal friction in the fluid to a value in the range of one-hundred million times its actual molecular value. The most important factor affecting the two different types of flow patterns in Figure 7, however, is the boundary condition along the northern edge of the gyre. In the case of drag-law turbulent paramterization, the condition of no-transport through the northern boundary is sufficient to determine the flow. In the case of eddy-viscosity parameterization, an additional condition is required. When the eastward flowing jet is allowed to slip freely along the northern boundary, the current is steady and hugs up against the northern boundary latitude. When, however, the northern bounding latitude is treated like a solid wall (a no-slip condition), the current separates and plunges southeastward in a wavelike pattern.

Many important unresolved physical questions remain even for the highly idealized class of barotropic wind-driven models. But these models exclude, even in idealized form, basic gyre and larger scale physical processes that operate oceanically. Moreover, they do not tackle the problems associated with the vertical structure of currents, nor the geographical and vertical distributions of density, temperature, and salinity. Mechanistic models in three spatial dimensions, which include effects of stratification and of the driving of currents by heating and cooling of the sea surface (in addition to the wind) have been reviewed by Bryan [25]. A schematic description of this class of model and its circulation pattern is shown in Figure 8. With thermal forcing, the subtropical gyre is no longer isolated from the rest of the ocean by the no-transport condition. Across these latitudes of no wind-forced transport there is flow that reverses with depth. Thus, a domain larger than a single gyre must be studied, or an ingenious parameterization of this effect must be introduced. A serious factor influencing effective exploitation of this class of models is represented by the very long time scales associated with the thermal circulation. The time scale for complete equilibrium to be achieved in the deep ocean in response, for example, to changes in surface forcing, is centuries. This is the so-called *spin-up time* for the initial boundary value problem for this class of models. Centuries of real time in the oceans implies days of actual computer time. Thus, only a few model solutions exist, and parameter exploration has been almost prohibited. Large-scale global modeling could draw

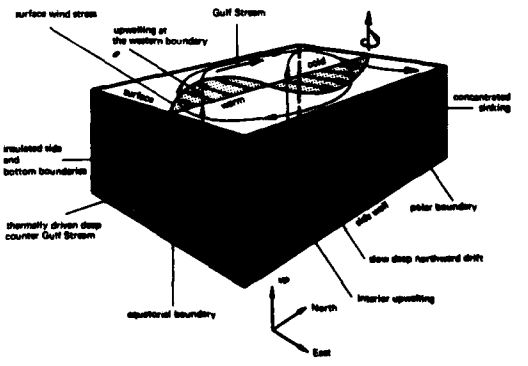


Figure 8—Schematic description of steady forcing and resulting circulation [22].

crucial insights from this class of large-scale mechanistic models. Hopefully, some combination of advances in the construction of model problems, indirect approaches to the physical spin-up problem, and the introduction of advanced and novel techniques of numerical analysis into calculations using fifth-generation computers will result in the extraction of considerable further physical results from these models. This possibility is particularly attractive because of the wealth of new global data on the distribution of the characteristics of water properties [26]. The data require such models for their rationalization and in turn provide model vertification potential.

The simulation model approach to gyre and larger scale modeling is exemplified by the calculations of Holland and Hirschmann [27]. In the attempt to model realistically the North Atlantic Ocean, certain compromises were adopted. The model is diagnostic, the resolution is rather coarse (1° of latitude by 1° of longitude), and a large value of eddy viscosity employed. The domain modeled extends from 10.5° south latitude to 50.5° north latitude. The region is connected to the northern and southern latitudes by imposing at these latitudes (a) the transport obtained from a barotropic global computation (see Figure 10a) and (b) an arbitrary assumption that maintains a simple dynamical balance between the velocity and density fields. Three results are illustrated in Figure 9. The first two cases, (a) and (b), are for a flat bottom; the third, (c), has realistic bottom topography. Case (a) is barotropic (uniform density), and cases (b) and (c) use the observed density field from classical hydrographic observations smoothed and/or interpolated to 1° resolution. Note the strong qualitative and quantitative differences in the results. The validity of the physics of the deep flow, which interacts with the bottom topography to produce the distinctive circulation pattern of the "most realistic" case, (c), depends on deep circulation processes that must be evaluated by mechanistic studies and focused observational experiments.

Some of the most advanced direct global model results published to date [28, 29] are exhibited in Figure 10. The first case, (a), is barotropic, the second, (b), is diagnostic, and the third, (c), is prognostic. Pond and Bryan [20] discuss these results. Coastline geometry and bottom topography are treated as accurately as allowed within the 2°-square resolution of the model. Again a high eddy viscosity is necessary, but no open-boundary interconnection conditions are needed. The prognostic calculation is far from equilibrium. It is the result of using for initial values in the prognostic problem the density field specified for the diagnostic case, (b), and the velocity field computed for that case. The circulation pattern, (c), results after 2.3 years of integration, which is two orders of magnitude shorter than the time anticipated to be required for final adjustment.

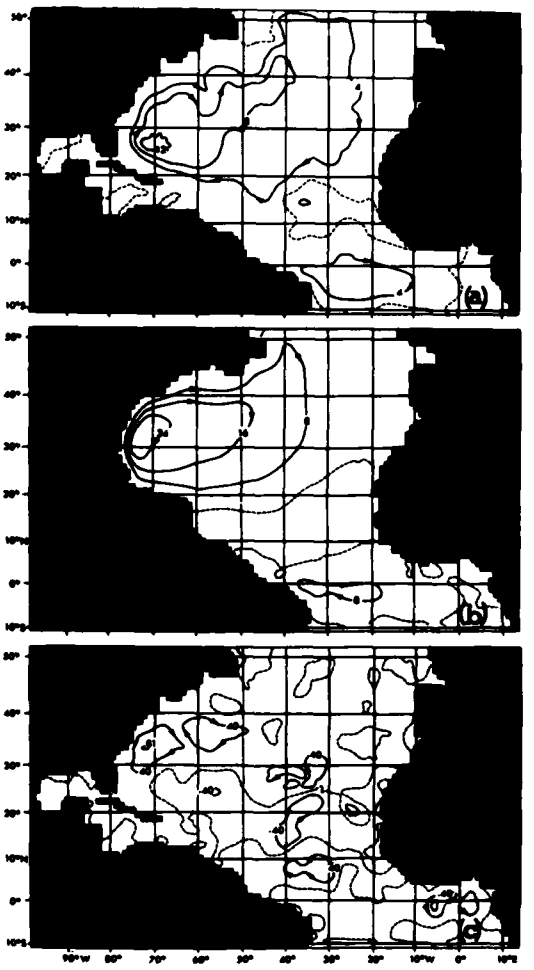


Figure 9—Transport stream function, diagnostic calculations: (a) uniform density, flat ocean bottom; (b) observed density, flat ocean bottom; (c) observed density and bottom topography [27].

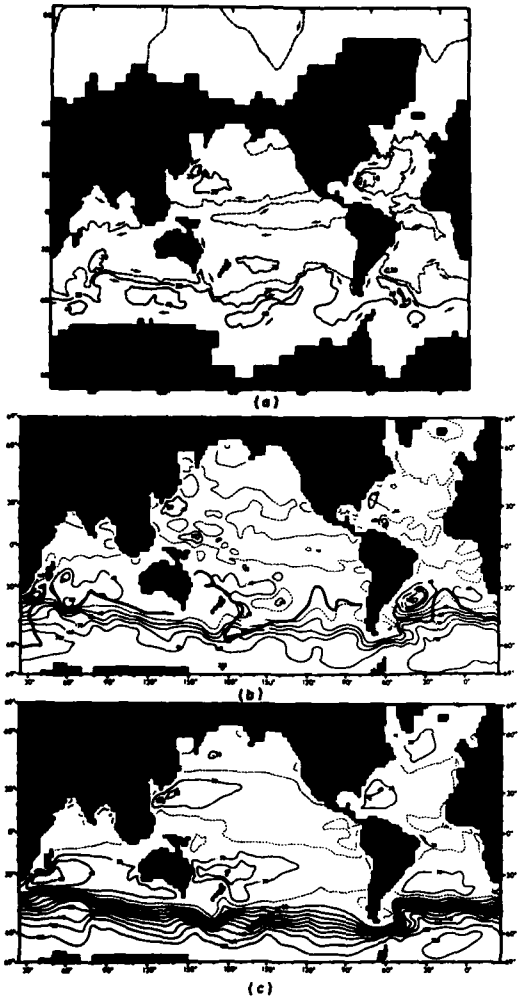


Figure 10—World ocean mass transport stream function: (a) barotropic; (b) diagnostic, baroclinic; (c) prognostic, baroclinic [20].

Each ocean-model year required 10 h of computer time. The rapid evolution of the prognostic case away from the diagnostic result indicates a degree of physical inconsistency in the diagnostic result, which may be due to the fact that the model linear viscous boundary layer width is about six times the observed boundary layer width. Patterns are smoothed, and major current transports are lowered. These results provide information both for the assessment of this approach to global modeling and for the evaluation of specific model parameters and assumptions. For simplicity and continuity of pictorial representation, maps of transport streamlines have been shown. Three-dimensional models provide, of course, distributions for all state variables. In Figure 11 we see the global distribution of temperature at a depth of 120 m obtained from the preliminary results of a world ocean model by Takano [8]. The purpose for the development of this model is to provide a simulated ocean circulation for coupled air-sea climatic studies; the horizontal resolution is 4° latitude by 2.5° longitude.

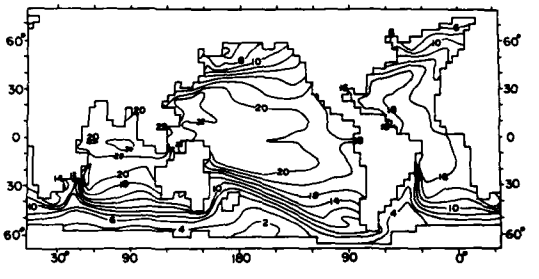


Figure 11—Temperature at 120m from a world ocean simulation [8].

We conjecture that global models of the future will be compounded from submodels of varying resolution and special purpose. The elements will be linked together by artful assumptions devised to meet the physical requirements or practical purposes of the desired computation or forecast of the composite model. The linkage parameterization and related computational schemes can be expected to be constrained by numerical analytical factors which will be discovered during the construction and development of composite models. On the global scale for certain purposes the direct approach global models, developed from the prototypes of the preceding paragraph, may be expected to provide the global skeletal framework. For other purposes, the global framework may be more schematic. For example, results or forecasts on certain scales may be required anywhere in the global ocean, rather than everywhere in the global ocean. In such a case, a mobile, limited-domain submodel could conceivably move about on a global schematic formulated with limited parametric indices of interconnection with the external ocean.

Contemporary special-purpose modeling efforts afford insight into the nature of the submodels that may be anticipated to form the elements of future composite global models. With no attempt at comprehensiveness, we cite here the following illustrative examples: (a) high-resolution gyre models (Figure 12); (b) special regional models (Figure 13); (c) local open-ocean models for the mechanistic study of dynamical processes (Figures 14 and 15) or for detailed local predictions (Figure 16); and, (d) detailed models for near-surface layer dynamics and forecasts (Figure 17).

Figure 12 shows transport streamlines for a model subtropical and subpolar gyre. Figure 12a shows the results of a classical model, demonstrated for a subtropical gyre alone in Figure 6 (here eddy viscosity replaces the low drag assumption). Figures 12b and 12c show results obtained when the resolution is significantly increased and the eddy viscosity substantially reduced. The flow is highly variable and illustrates the phenomenon of "eddying" [35]. The time-variable closed-streamline patterns are the analog in the deep ocean of atmospheric weather patterns. They are usually much more energetic than the climatic mean flow, and for many purposes they must be explicitly resolved in numerical models. Subgridscale parameterization is not yet possible, although in the future this may be done for purposes in which only the local statistical effects of this scale of motion matter.

Regional models are constructed for domains that may have special dynamics or distinctive boundary characteristics. In the model illustrated in Figure 13, these features are both provided by the equator. Moreover, the special purpose of the model is the investigation of the El Niño phenomenon. During El Niño, there is an anomalous replacement of cold upwelling water in the coastal region off Ecuador and Peru by an influx of warm water; this has had serious adverse economic effects on the fisheries and related industries. A recent hypothesis that has been

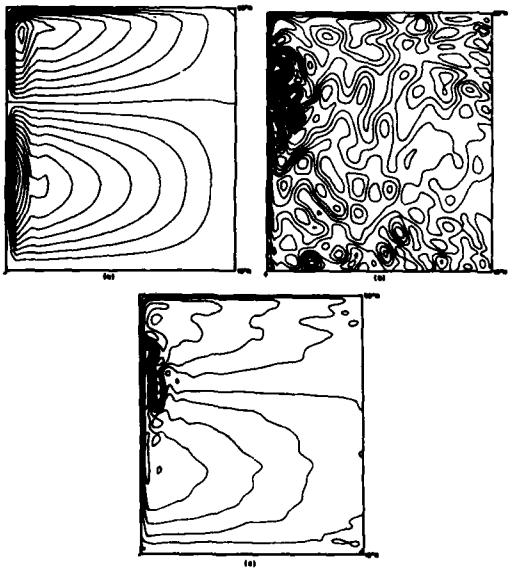


Figure 12—Transport stream function for a model subtropical and subpolar gyre simulation: (a) large lateral viscosity, low resolution; (b) smaller viscosity, higher resolution, instantaneous flow; (c) conditions similar to (b), except for time-averaged flow [30].

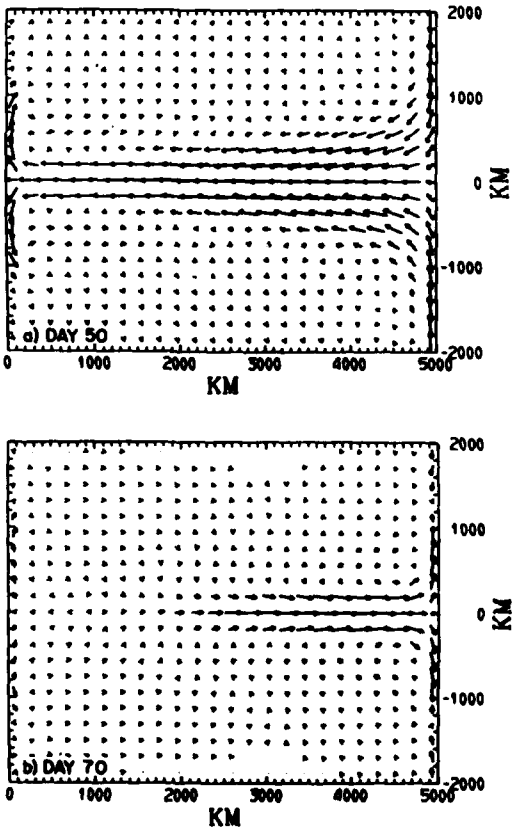


Figure 13—Velocity field in an El Niño simulation [31].

examined numerically by Hurlbut et al. [31] relates such an event to the reduction of the trade winds over the whole central Pacific Ocean. The model shown in Figure 13 extends from coast to coast, and 2000 km north and south of the equator, at which boundary latitudes the model is connected to the external ocean by a parameterization that effectively maintains the correct wind-induced transport. The figure demonstrates the reversal of the flow pattern in the upper ocean several days after the simulated trade winds have been turned off. The reversal involves the participation of a variety of large-scale subsurface waves.

Figure 14 illustrates an open-ocean model exploration of the midocean eddy flow. The internal physics is complicated, if treated in detail, and the resolution is high. The trade-off is the simplicity of the parameterization of the open boundary conditions. The flow is assumed to be periodic (i.e., the box pictured is assumed to be part of an ocean of infinite extent, which repeats the pattern shown over and over in both the north-south and east-west directions). An initial-value model problem is solved for the time evolution of an assumed initial state. The model reproduces some aspects of the observed eddy field (e.g., larger space scales are found in the upper ocean than in the deep water). Figure 15 is an example of such a model "moved to the region of an intense current" (e.g., a simulated Gulf Stream situation south of the Grand Banks). The modeling technique is similar, but the physical processes explored are greatly different. From one point of view, such models can obviously be regarded as embryonic local forecast models. For forecasting, more realistic boundary conditions are required, as well as empirical data for initialization and across boundaries. This data requirement is very stringent because the spatial gradients of the flow must be accurately specified. A direct heuristic approach to short-term, small-domain prediction of an intense current has been taken by Kollmeyer and Paskausky [33]. A 5-km grid is used over a 110-km-square domain, and the modeling forecast is based on two hydrographic surveys spaced 8 days apart (Figure 16).

The structure and the physics of the near-surface layer of the ocean are quite complicated. Features of the structure and the interconnection of this region with the deeper flow depends on the time scale of interest. Daily and seasonal variations occur with considerable regularity, but irregular events, such as the passing of severe

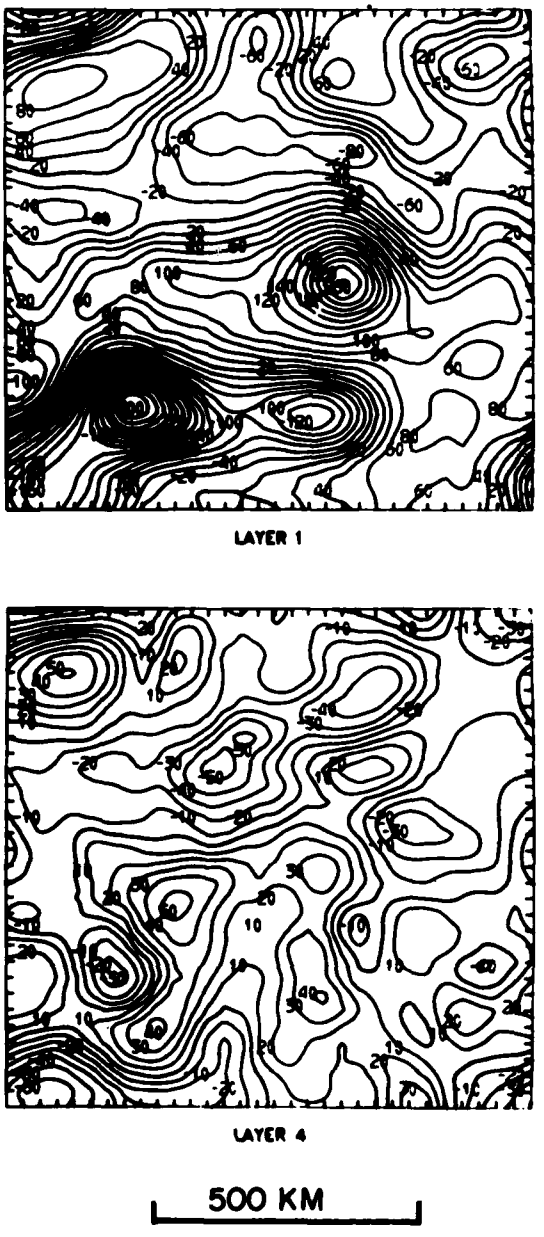


Figure 14—Layer stream function from a limited-region mesoscale simulation [32].



Figure 15—Temperature distribution with time in an intense current experiment [32].

storms, have considerable influence. Present models that incorporate sufficiently realistic physics to give results of interest are almost entirely local, i.e., they assume no horizontal variations and deal only with the vertical structure [34]. Considerable recent progress has been made in this important area of research, and an acceleration is anticipated. The first results of models incorporating effects of horizontal variations and effective coupling with the deeper flow should be obtained in the near future.

Ocean forecasting is in its infancy. Although the crystal ball is cloudy, the coming years should

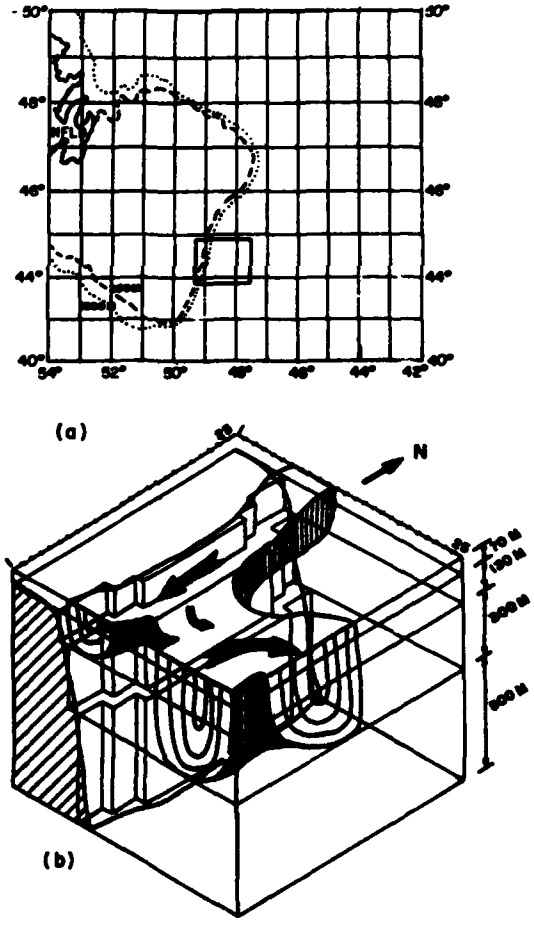


Figure 16—(a) Location of Labrador Current experiment region; (b) schematic representation of the modeled area showing the four layers, the Continental shelf and slope, and the Labrador/Gulf Stream flow pattern [33].

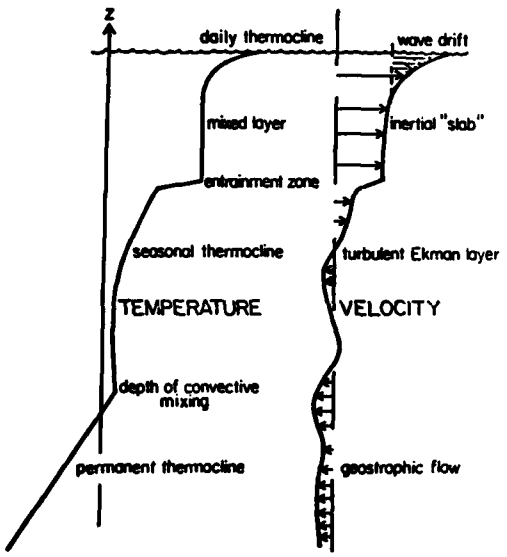


Figure 17—Schematic Representation of profiles in the upper ocean [34].

produce results that contribute significantly to the foundation of useful, albeit limited and special-purpose, forecasting. Numerical models will play a vital and essential role. The enormous difficulty of acquiring the requisite observational data plagues the potential ocean forecaster. Every effort must be made to devise and exploit novel instrumentation as well as to isolate, by research activity and trial and error, critical observational parameters for special purposes. The optimal exploitation of the potentially available data base could be of crucial importance to the success or failure of forecast schemes. The use of so-called objecitve analysis space-time interpolation techniques and updating procedures [14] should be of even greater importance in oceanography than in meteorology. Such methods attempt to combine optimally what is known about the statistics of the region, the dynamics of the flow, and the observations most recently acquired into the best possible forecast.

REFERENCES

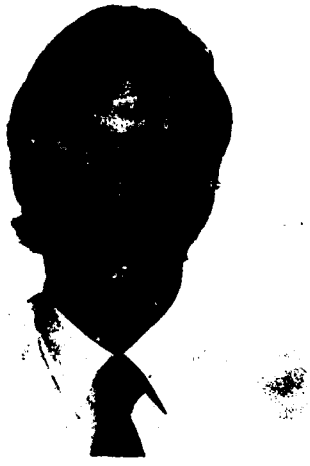
1. J. Charney, "Impact of Computers on Meteorology," *Comp. Phys. Comm.* 3 (Suppl.), 117 (1972).
2. O. Reynolds, "On the Dynamical Theory of Incompressible Viscous Fluids and the Determination of the Criterion," *Phil. Trans. Roy. Soc. Lon.* **186**, 123 (1894).
3. H. U. Sverdrup, M. W. Johnson, and R. H. Fleming, *The Oceans, Their Physics, Chemistry, and General Biology*, Prentice-Hall, New York, 1942.
4. R. H. Kraichnan, "Eddy Viscosity in Two and Three Dimensions," submitted to *J. Atmos Sci.* (1976).
5. G. L. Mellor and P. A. Durbin, "The Structure and Dynamics of the Ocean Surface Mixed Layer, *J. Phys. Oceanogr.* **5**, 718 (1975).
6. V. P. Starr, *The Physics of Negative Viscosity Phenomena*, 196 p., McGraw-Hill, New York, 1968.
7. F. P. Bretherton and M. Karweit, "Mid-Ocean Mesoscale Modelling," in *Numerical Models of Ocean Circulation*, p. 237, National Academy of Sciences, Washington, D.C., 1975.
8. K. Takano, "A Numerical Simulation of the World Ocean Circulation: Preliminary Results," in *Numerical Models of Ocean Circulation*, p. 121, National Academy of Sciences, Washington, D.C., 1975.
9. K. Bryan, "A Numerical Method for the Study of the Circulation of the World Ocean," *J. Comp. Phys.* **3**, 347 (1969).
10. H. O. Kreiss, "A Comparison of Numerical Methods Used in Atmospheric and Oceanograhpic Applications," in *Numerical Models of Ocean Circulation*, p. 255, National Academy of Sciences, Washington, D.C., 1975.
11. C. E. Leith, "Future Computing Machine Configurations and Numerical Models," p. 301, in *Numerical Models of Ocean Circulation*, National Academy of Sciences, Washington, D.C., 1975.
12. World Meteorological Organization, "The Physical Basis of Climate and Climate Modeling," GARP Publications Series, No. 16, 1975.
13. N. A. Phillips, "Models for Weather Prediction," *Annu. Rev. Fluid Mech.* **2**, 251 (1970).
14. G. J. Haltiner and R. T. Williams, "Some Recent Advances in Numerical Weather Prediction," *Mon. Weath. Rev.* **103**, 571 (1975).
15. L. F. Richardson, *Weather Prediction by Numerical Process*, Cambridge Univ. Press, London, 1922.

16. J. G. Charney, R. Fjortoft, and J. Von Neumann, "Numerical Integration of the Barotropic Vorticity Equation," *Tellus* **2**, 237 (1950).
17. E. N. Lorenz, "The Mechanics of Vacillation," *J. Atmos. Sci.* **20**, 448 (1963).
18. A. S. Sarkisyan, "On the Role of the Drift Advection of the Density in the Baroclinic Ocean," *Oceanol.* **2**, 395 (1961).
19. K. Bryan, "A Numerical Investigation of a Nonlinear Model of a Wind-Driven Ocean," *J. Atmos. Sci.* **20**, 594 (1963).
20. S. Pond and K. Bryan, "Numerical Models of the Ocean Circulation," *Rev. Geophys. Space Phys.* **14**, 243, 257 (1976).
21. E. Goldberg, ed., *The Sea*, vol. VI, John Wiley & Sons, New York (to appear).
22. A. R. Robinson, "Eddies and Ocean Circulation," *Oceanus* **19**, 2 (1976).
23. H. Stommel, *The Gulf Stream*, Univ. of Calif. Press, Berkeley, Calif., 1958.
24. G. Veronis, "Wind Driven Ocean Circulation, I and II," *Deep Sea Res.* **13**, 17 (1966).
25. K. Bryan, "Three-Dimensional Numerical Models of the Ocean Circulation," in *Numerical Models of Ocean Circulation*, p. 94, National Academy of Sciences, Washington, D.C., 1975.
26. "Circulation of the Oceans," *Mosaic* **4** (3) (1973), National Science Foundation, Washington, D.C. See also *Earth and Plan. Sci. Lett.* **22** (1974) for further GEOSECS references.
27. W. R. Holland and A. D. Hirschman, "A Numerical Calculation of the Circulation in the North Atlantic," *J. Phys.* **2**, 336 (1972).
28. K. Bryan and M. D. Cox, "The Circulation of the World Ocean: A Numerical Study," Part I. G.F.D.L./NOAA, Princeton, N.J. (unpublished manuscript), 1972.
29. M. D. Cox, "A Baroclinic Numerical Model of the World Ocean: Preliminary Results," in *Numerical Models of Ocean Circulation*, p. 107, National Academy of Sciences, Washington, D.C., 1975.
30. Y. J. Han, "Numerical Simulation of Mesoscale Ocean Eddies," U.C.L.A. Dep. of Meteorology, Ph.D. Thesis, 1975.
31. H. E. Hurlburt, J. C. Kindle, and J. J. O'Brien, "A Numerical Simulation of the Onset of El Nino," Contribution from the Geophysical Fluid Dynamics Institute, Florida State University, 1976.
32. P. B. Rhines, "Physics of Ocean Eddies," *Oceanus* **19**, 26 (1976).
33. R. C. Kollmeyer and D. F. Paskausky, "Labrador Current Predictive Model," submitted to *J. Phys. Oceanogr.* (1975).
34. P. P. Niiler, "One Dimensional Models of the Seasonal Thermocline," in *The Sea*, vol. VI, E. Goldberg, ed., John Wiley and Sons, New York, to appear.
35. "Ocean Eddies," *Oceanus* **19** (3) (1976).

Since 1954 Dr. Walter H. Munk has held the rank of Professor at the Institute of Geophysics at the Scripps Institution; since 1959 he has also been Associate Director of the Institute of Geophysics and Planetary Physics (systemwide) at the University of California, San Diego. His many honors and awards include the Arthur L. Day Medal, Geological Society of America (1965); the Sverdrup Gold Medal, American Meteorological Society (1966); the Alexander Agassiz Medal, National Academy of Sciences (1976); the Maurice Ewing Award, American Geophysical Union and the United States Navy (1976); election as Foreign Member of the Royal Society of London (1976). He is a member of the National Academy of Sciences. Dr. Munk received his B.S. and M.S. from the California Institute of Technology and a Ph.D. in oceanography in 1947 from the Scripps Institution of Oceanography. He became Assistant Professor of Geophysics at the University of California, San Diego in 1947.



Peter Worcester received his B.S. degree in Engineering Physics from the University of Illinois and his M.S. degree in Physics in 1969 from Stanford University. He served in the United States Navy from 1969 to 1972. He has received a National Science Foundation fellowship, the Churchal Fellowship (1968), and the Lisle Abbot Rose Award of the University of Illinois (1968).



MONITORING THE OCEAN ACOUSTICALLY

Walter Munk and Peter Worcester

Institute of Geophysics and Planetary Physics
Scripps Institution of Oceanography
University of California, San Diego
La Jolla, Calif.

AN APPRECIATION

In this essay on the past and future interaction of *ocean dynamics* and *ocean acoustics*, it is fitting to start with an appreciation of the Office of Naval Research. Ocean dynamics and acoustics can trace their modern development to the end of World War II, when ONR was founded. They grew up together as three siblings. The two ocean disciplines are lusty brothers under the thoughtful support of a loving sister; the brothers are rather independent and headstrong and pay scant attention to one another, though they share a deep appreciation for their sister. After 30 years, it is time the brothers showed some maturity and mutual consideration.

THE DEMISE OF ZERO-FREQUENCY OCEANOGRAPHY

The classical physical oceanographers cast their Nansen bottles and contoured dynamic heights, so that these would be available for computing geostrophic currents which are then published on permanent charts. (The Glossary to this paper contains some of the oceanographic terms used here.) The acousticians found it difficult to relate this delightfully simple view of a steady ocean interior to the complex and time-variable transmission of acoustic signals through the ocean; and so they invented their own oceans, described in terms of space and time correlations. The gap between the two ways of describing the ocean was unbridgeable.

In a sense the acoustician's ocean was ahead of the oceanographer's ocean. The acoustician had long been familiar with noisy processes and their description in terms of continuous spectra. He now applied these notions to the ocean processes themselves. The oceanographer was just beginning to do so.* He had received an early jolt when he occupied some deep-sea anchor stations for a few days and measured rather sizable variations in the temperature and salinity profiles. These were diagnosed as internal waves (the theory goes back to Stokes in 1847). The early interpretations were in terms of *discrete* tidal frequencies and the gravest one or two vertical modes. Gradually, the notion developed that internal waves occupied many modes and a *continuum* of frequencies, from the inertial frequency (2 sin latitude cycles per day) to the Brunt-Väisälä (or buoyancy) frequency (typically a few cycles per hour). This led

*Probably this delay was a matter of frequency. High-frequency acoustic spectra could be readily measured with analog devices. A similar power-spectral analyses of low-frequency ocean oscillations did not catch on until the corresponding numerical techniques became accessible.

to the picture of a steady ocean structure and associated zero-frequency circulation, upon which an internal wave noise is superimposed.

This viewpoint came into difficulties with the first direct measurements in 1962 of midwater motion, using neutrally buoyant Swallow floats that were tracked acoustically. These measurements revealed a variable structure with kinetic energy exceeding that of the mean motion by two orders of magnitude! A decade after these pioneering ARIES [1] measurements, two massive efforts were mounted to map the subinertial variable flow field: the Soviet POLYGON [2] and the U.S.-U.K. MODE expeditions [3]. We now know that a typical flow field in the upper ocean corresponds much more nearly to 1 ± 10 cm/s than to 10 ± 1 cm/s, and this has far-reaching consequences.

STATISTICAL OCEAN MODELS

Oceanographers were thus driven towards a description of ocean processes that relied heavily on the concept of continuous spectra over an enormous range of space and time scales. Above the inertial frequency, internal waves measured at different places and times were surprisingly consistent with a universal spectrum. Below the inertial frequency the so-called mesoscale eddies were found dominant (correlation scales 100 km and 60 days), and these bear some resemblance to Rossby (or planetary) waves. But the application of wave mechanics is limited here by strong nonlinear interaction between the various scales, and this had led to an alternate description in terms of a two-dimensional (or geostrophic) turbulence. Here the cascade of energy is towards large scales, thus preserving sharp boundaries of major ocean features, in contrast to the fuzzy structure of laboratory turbulence.

The reader will note that the oceanographers had begun to speak a language that was close enough to that of the acousticians that a detente was within reach. Still, there were important differences. The acousticians had become accustomed to work with homogeneous isotropic spectra of ocean variability, but ocean fluctuations (except perhaps at very small scales) are neither homogeneous nor isotropic.

MIMI

At about the time Swallow, Stommel, and their associates were acoustically tracking midwater floats to discover the complexity and variability of the ocean structure, Steinberg and Birdsall were conducting their pioneering sound-transmission experiment across the Straits of Florida [4]. They discovered tides as an oceanographic factor. (This was not surprising to oceanographers, who were quite accustomed to tidal components in all their measurements; the effect of tides on shallow-water acoustic transmissions had previously been noted by Urick.) There were some difficulties in the interpretation associated with shallow-water effects, and subsequent efforts (in which they were joined by Kronengold, Clark, and others) were shfited to a 1250-km path from Eleuthera to Bermuda [5].

An essential feature in these experiments (called MIMI for the Miami-Michigan participation) was that they gave continuous observations over many months, and this opened the way for a meaningful geophysical interpretation.* In essence the experiment consisted of transmitting a 406-Hz signal and recording the relative phase and intensity of the received signal using a perfectly synchronized 406-Hz oscillator. The resulting time series of acoustic phase and intensity are dominated by occasional fadeouts and phase jumps, which are the result of interference among the many paths (≈34) from source to receiver. This is an interesting problem in random-walk statistics, but unfortunately the ocean is involved only in a limited way (the determination of a single parameter, $\langle \dot{\phi}^2_1 \rangle$, the mean-square rate of phase along any one path).

The parameter could in principle be measured directly if such a single path could be isolated from all other paths by a suitable directional antenna, but this is not practical. What is measured is the vector sum of all paths, giving the intensity I(t) and phase $\phi(t)$ of the combined multipath signal. The multipath $\langle \dot{\phi}^2 \rangle$ is not the same thing as the

*This has not always been the case in acoustic experiments. It is amazing to us how experimenters could speak of a tidal effect in a 4-h run, when they would not think of describing an acoustic signal from observations extending over one-third the pulse length.

singlepath $<\dot{\phi}_1^2>$; it is generally larger and shows occasional near-180° "jumps" associated with intensity fadeouts (Figure 1). Over a period of a month ϕ can change by many cycles, and it is necessary to keep track of the sign of the phase jumps. The parameter $<\phi^2>$ increases linearly with record time in accordance with random-walk statistics. The spectrum of $\phi(t)$ contains high frequencies (associated with phase jumps) and low frequencies (associated with random walk) that are not contained in ϕ_1 (t).

Even though the single-path $\phi_1(t)$ is not directly measured, it can be inferred from the multipath statistics under quite reasonable assumptions [6]. The result is

$$\text{rms } \dot{\phi}_1 = 3.5 \times 10^{-3} \text{ sec}^{-1},\ 5.2 \times 10^{-3} \text{ sec}^{-1}$$

for Mid-Station and Bermuda, respectively. Now this parameter depends on the fluctuations of sound velocity along the transmission path and can be calculated if certain statistical properties of these fluctuations are known. Using an internal wave spectrum, based entirely on oceanographic measurements, and performing these calculations leads to the result [7]

$$\text{rms } \dot{\phi}_1 = 3.5 \times 10^{-3} \text{ sec}^{-1},\ 5.2 \times 10^{-3} \text{ sec}^{-1}$$

for the two stations. There are no loose parameters here, and the data sets are entirely independent, one acoustic, the other oceanographic. We would conclude that a connection has been made between the acoustician's ocean and the oceanographer's ocean. An interesting remaining question, which is being actively pursued by the MIMI group, is whether the low-frequency $\phi(5)$ variations can be entirely ascribed to random-walk statistics, or whether they are in part the result of low-frequency ocean variations. Such variations could be the result of mesoscale eddies, for example, and this brings us close to our thesis for monitoring the ocean acoustically. But first we will review briefly several other experiments that are pertinent to this topic.

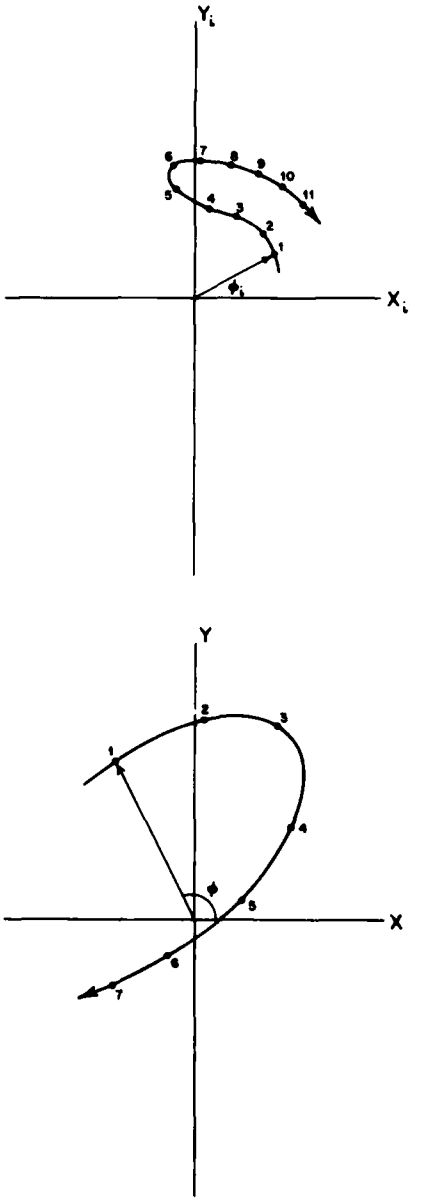


Figure 1—Phasor diagrams for single-path (left) and multipath (right) acoustic transmissions. The Cartesian coordinates designate the in-phase and quadrature components of the received pressure signal, at time 1, 2, . . . , with the vector indicating the positions at time 1. ϕ_1 and $I_1 = 10 \log (X^2_1 + Y^2_1)$ are the singlepath phase and dB intensities, and similar ϕ and I for the multipaths. Generally $\phi(t)$ varies more rapidly than $\phi_1(t)$, and I (t) is more variable than $I_1(t)$. Passage of the vector trajectory near the origin (between times 5 and 6 in the right figure) is associated with a fadeout and a rapid change of phase by almost ±180°, the sign depending on which side of the origin is passed (minus in the figure).

COBB, AFAR, AND WHOI

Ewart [8] transmitted for about a week 4- and 8-kHz pulses between source and receiver placed on Cobb Seamount at 1-km depth, separated by 17

km. Ray-tracing gives a single *downward* refracted path with a turning point at 1350 m. With this single-path geometry one can attempt to interpret the entire measured spectra of phase and intensity with those derived from an ocean model. Accordingly, much more can be learned than from the single parameter $<\phi^2_1>$ in a multipath experiment, at the expense, of course, of monitoring a much smaller ocean volume.

From the internal wave model one computes a phase spectrum bounded by the inertial and buoyancy frequencies, and proportional to ω^{-3} at intermediate frequencies. The measured phase spectrum has some of these characteristics. The measured rms (ϕ) is 1.6 cycles, and the value computed from the ocean model is 0.8 cycles. Again, there are no loose parameters. However, the measured intensities greatly exceed the computed intensities, particularly at high frequencies. R. Dashen (personal communication) has demonstrated that the intensity fluctuations are greatly influenced by interference among "sporadic multipaths" associated with the fine structure and microstructure of the sound velocity profile. This may or may not be the explanation.

Ellinthorpe's [9-11] ambitious transmission study AFAR in the Azores involves a source and receiver at 600-m depths separated by 38 km, with an *upward* refracted ray (unlike COBB) reaching an apex at a 300-m depth. The experiment was conducted over a broad range of acoustic frequencies, and involved an intensive program of ocean monitoring. The Mediterranean outflow is a prominent feature. The analysis has not been completed; Ellinthorpe's preliminary conclusion is that internal waves play a significant role in the observed acoustic fluctuations, but that in addition one must take into account the role of spatial ocean correlation structure being advected through the array by the mean currents.

Finally, we wish to refer to the ongoing Woods Hole work by Porter and Spindel [12], involving drifting and moored sensors whose motions are monitored by a bottom-based Doppler navigation system. Here again the indication is of a combined role of the time-variable internal wave effects and of the advected space-variable ocean structure (from internal waves, intrusions, or other processes). The advection is associated in one case with the movement of the water past the moored hydrophones, and in the other case with the drift of the hydrophone through the water.

MONITORING THE CALIFORNIA CURRENT

We are planning an experiment for acoustically monitoring mesoscale disturbances in the California Current. The goal is to install a moored deepsea triangle of transmitters and receivers with in-situ signal processing and data storage. The legs of the triangle would be 25 to 50 km in length, appropriate to the energetic mesoscale processes. The array would be left in place for 2 to 3 months to monitor the corresponding time scales. Each vertex of the triangle would have both a transmitter and a short vertical array of receivers, so that absolute travel times, differential travel times from reciprocal transmissions, and arrival angles can be measured. By using a broadband (1.5-3.5 kHz) transmitted signal and employing pulse-compression techniques, we can measure travel times to 10^{-4} s, or about 1 part in 10^5, while retaining the ability to resolve arrivals separated in time by about 0.6 ms. (For this precision to be meaningful we will clearly have to correct for mooring motion; Porter's bottom-based Doppler tracking system can perform this task to the required accuracy.) A travel time fluctuation of 1 part in 10^5 corresponds to a temperature fluctuation of about 0.01°C or a salinity fluctuation of about $0.01^{o}/_{oo}$ integrated along the ray path. Further, the differential travel time between reciprocal transmissions would give the mean flow velocity along the ray paths to about 1 cm/s.* The estimated precisions for temperature, salinity, and current velocity happen to be about the same as those achieved with modern instruments. The estimates may be optimistic, if for no other reason than the deterioration (spreading and wandering) of pulses by sopradic multipaths. Here it is our hope that spreading and wandering can serve to give a statistical measure of the variable fine structure and of internal wave activity in the array area.

*It has yet to be shown whether we can separate the effect of the current velocity from the nonreciprocity of paths due to current shear.

Some preliminary results have been obtained from ship-to-ship transmissions in a geometry similar to that proposed for the moored triangular array. A transmitter and a short vertical array of receivers were suspended from each of two ships at about 1-km depth and 25-km depth range (Figure 2). With this geometry a smoothed sound velocity profile constructed from data taken at the time of the experiment gives only two purely refracted ray paths: an upper path that comes within about 200 m of the surface and a lower path with a turning point at about 1500 m (Figures 3 and 4). The upper and lower paths can be separated from each other and from all the surface-reflected and bottom-reflected paths by the differences in travel time; all reflected arrivals occur much later than the purely refracted signals and were not recorded.

A phase-reversal pulse compression code (Barker code) centered at 2250 Hz was transmitted; the received signal was digitized and later demodulated and processed on a digital computer. The processing consists essentially of computing the covariance between the received signal and a replica of the transmitted signal (matched filter). The amplitude response of our processing filter was modified from that of a matched filter, however, to broaden and smooth the spectrum of the output pulse, improving the resolution in time between adjacent arrivals and reducing the sidelobes of the covariance (Figure 5).

A short sequence of the processed arrivals from one receiver at 30-s intervals is given in Figure 6. The first arrival is from the lower ray path. Its simplicity reflects the relative lack of fine structure in the deep ocean. The cluster of arrivals occurring about 40 ms later are from a number of ray paths that differ only slightly from the upper path shown in Figure 4 (micromultipaths). The micromultipaths are due to the perturbations in the sound velocity profile associated with the oceanic fine structure (e.g., internal waves and intrusions). A perspective presentation (at a different time) shows this splitting of the upper path

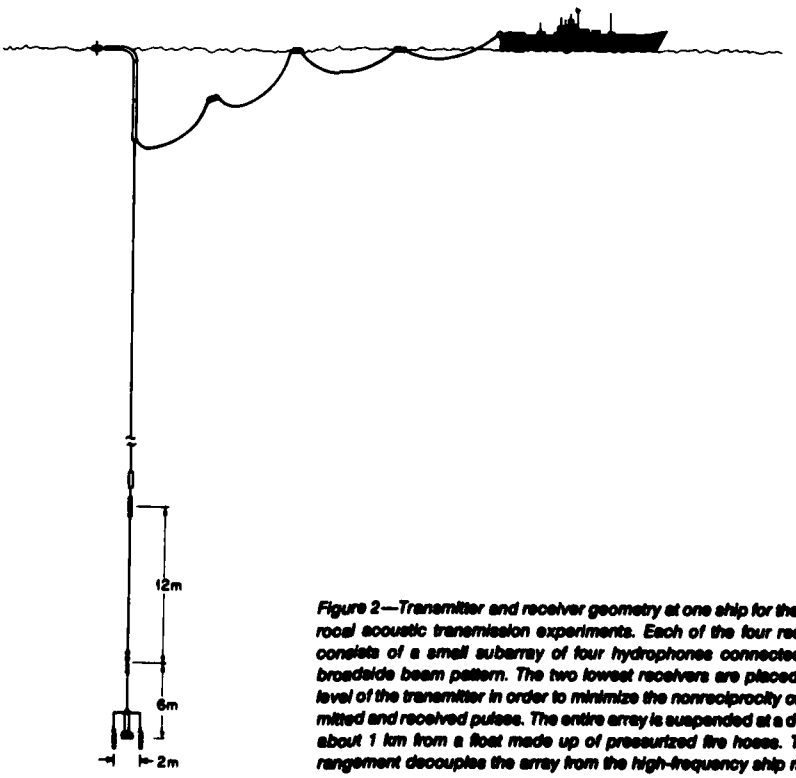


Figure 2—Transmitter and receiver geometry at one ship for the reciprocal acoustic transmission experiments. Each of the four receivers consists of a small subarray of four hydrophones connected for a broadside beam pattern. The two lowest receivers are placed at the level of the transmitter in order to minimize the nonreciprocity of transmitted and received pulses. The entire array is suspended at a depth of about 1 km from a float made up of pressurized fire hoses. This arrangement decouples the array from the high-frequency ship motion.

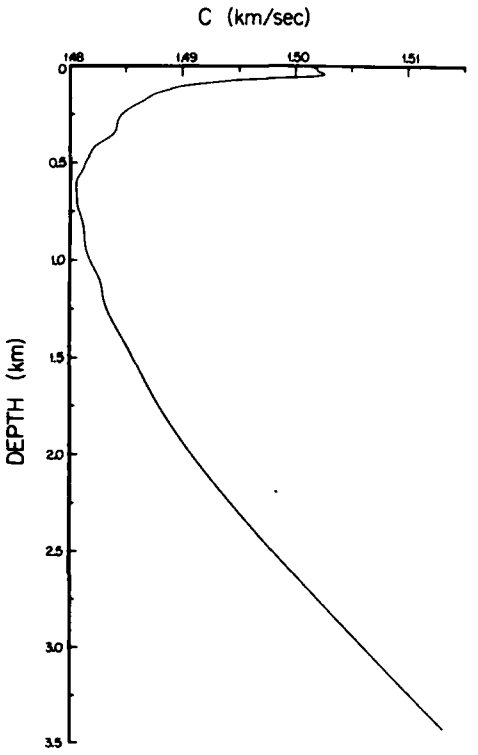


Figure 3—Sound velocity profile constructed from a cubic spline fit to data collected roughly halfway between the two acoustic stations (31°08'N, 120°09'W; April 1, 1978).

into micromultipaths even more clearly (Figure 7). A further perspective (Figure 8) shows the variations in the total travel time over a 2-h interval, due predominantly to the differential drift of the two vessels by about 300 m.

It is possible to give geophysical interpretations to the *differences* in reciprocal transmissions (Figure 9). If the sources, receivers, and medium are all fixed, then the transmission from *Agassiz* to *Scripps* and from *Scripps* to *Agassiz* should be perfectly reciprocal. We can ignore the nonreciprocity associated with a horizontal separation of 1 m between transmitter and receiver (Figure 2). In fact, there is an obvious nonreciprocity in the amplitudes of the arrivals; occasionally there are differences in the number of arrivals.

Current shear has a significant effect on the acoustic propagation. Much of the time it appears that one can identify corresponding arrivals at the two ships. For example, the lower arrival of the *Scripps* leads the lower *Agassiz* arrival by about 0.3 ms, in part as a result of the ship's drift relative to the mean water column. A simple straight-line ray calculation using this differential travel time and including the effect of differential drift between ships gives a current component from *Scripps* to *Agassiz* of 3 cm/s relative to *Agassiz*. There is some promise that we shall be able to derive currents from nonreciprocity between moored capsules.

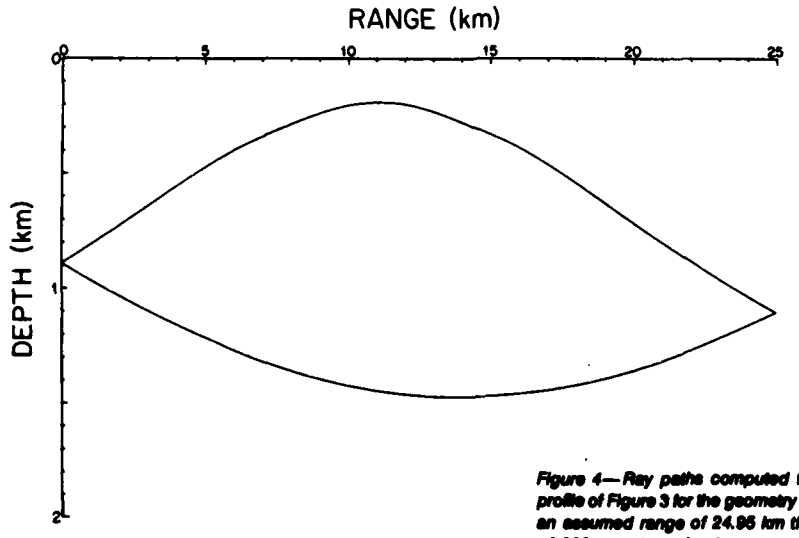


Figure 4—Ray paths computed from the smoothed sound velocity profile of Figure 3 for the geometry of the ship-to-ship experiment. With an assumed range of 24.96 km the travel time for the lower path is 16.828 s, and that for the upper path is 16.867 s.

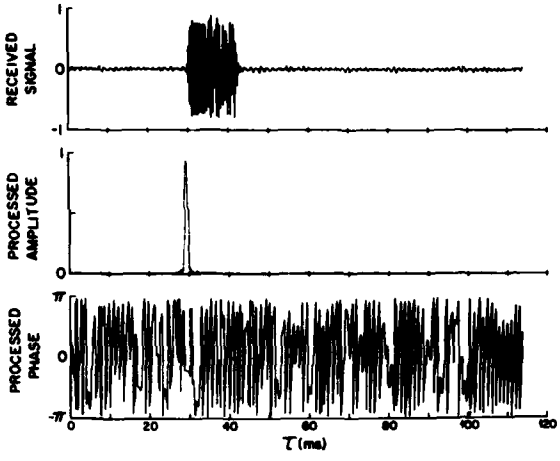


Figure 5—Thirteen-bit Barker code as received at very close range (3.5 km) before and after high-resolution processing. Note the spike in the processed amplitude and the temporary stationarity in processed phase at the first arrival time (29 ms in arbitrary units). The data are of the same duration as in the subsequent figures.

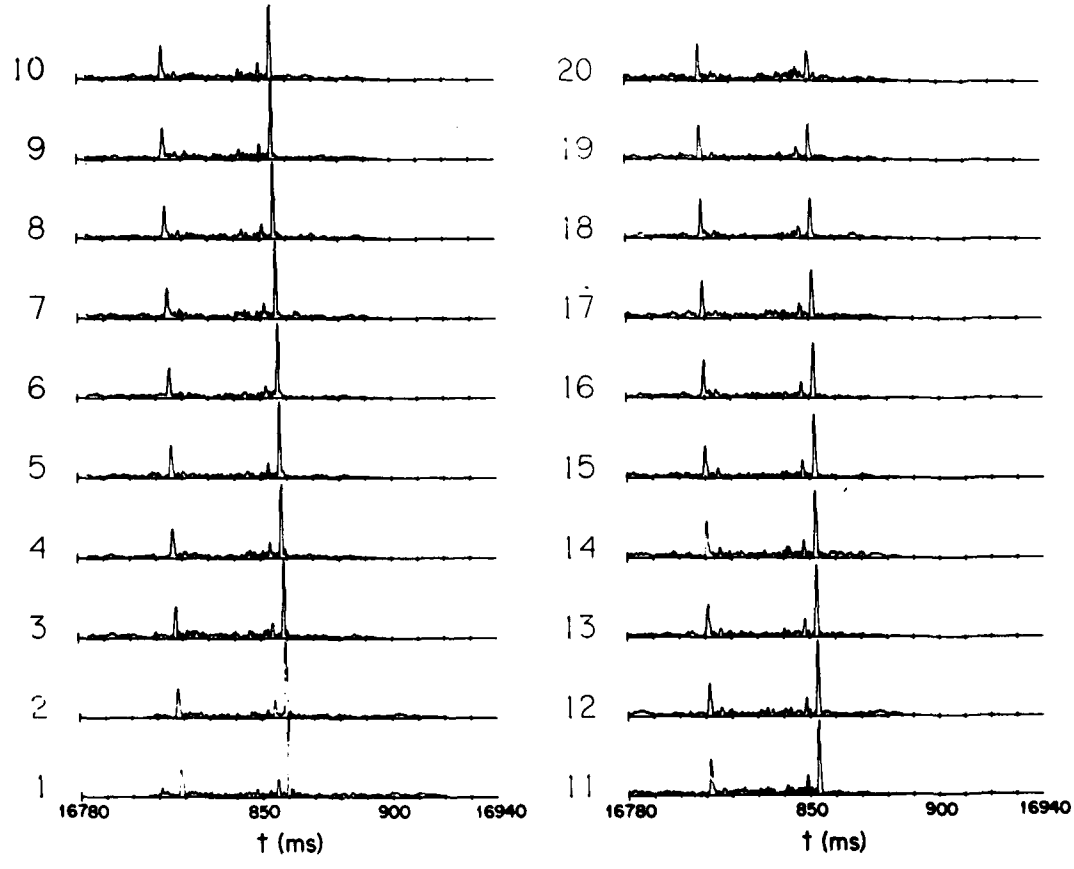


Figure 6—Arrivals at 30-s intervals after high-resolution processing (E. B. Scripps: April 3, 1978; 1444 to 1454). The time marks are in milliseconds from the start of the transmitted pulse. Note the dominant arrival along the upper path at the beginning of the series, and the dominant lower arrival 10 min later.

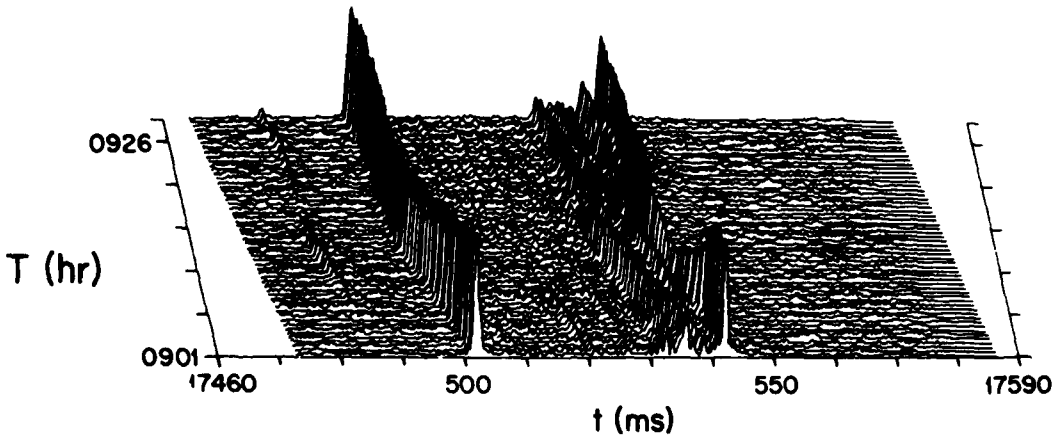


Figure 7—Arrivals at 30 s intervals after high resolution processing (Alexander Agassiz; April 3, 1976; 0901 to 0929). The time marks are in milliseconds from the start of the transmitted pulse. Note the relative complexity of the upper path. A precursor to the lower path has not yet been explained.

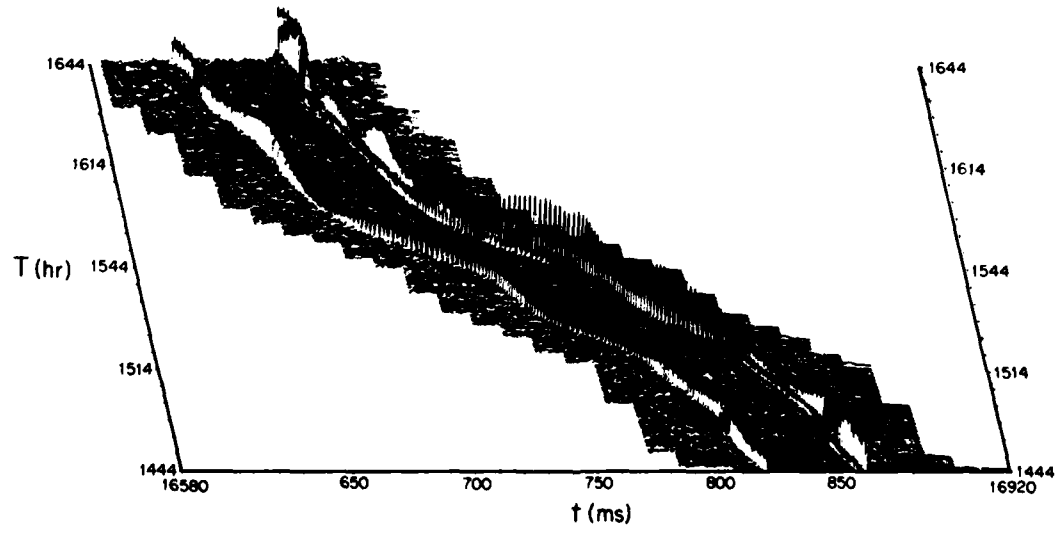


Figure 8—Arrivals at 30-s intervals after high-resolution processing (E. B. Scripps; April 3, 1976; 1444 to 1644). The time marks are in milliseconds from the start of the transmitted pulse. The gross reduction in travel times corresponds to a relative closing of the two vessels by about 5 cm/sec (0.1 Kn)

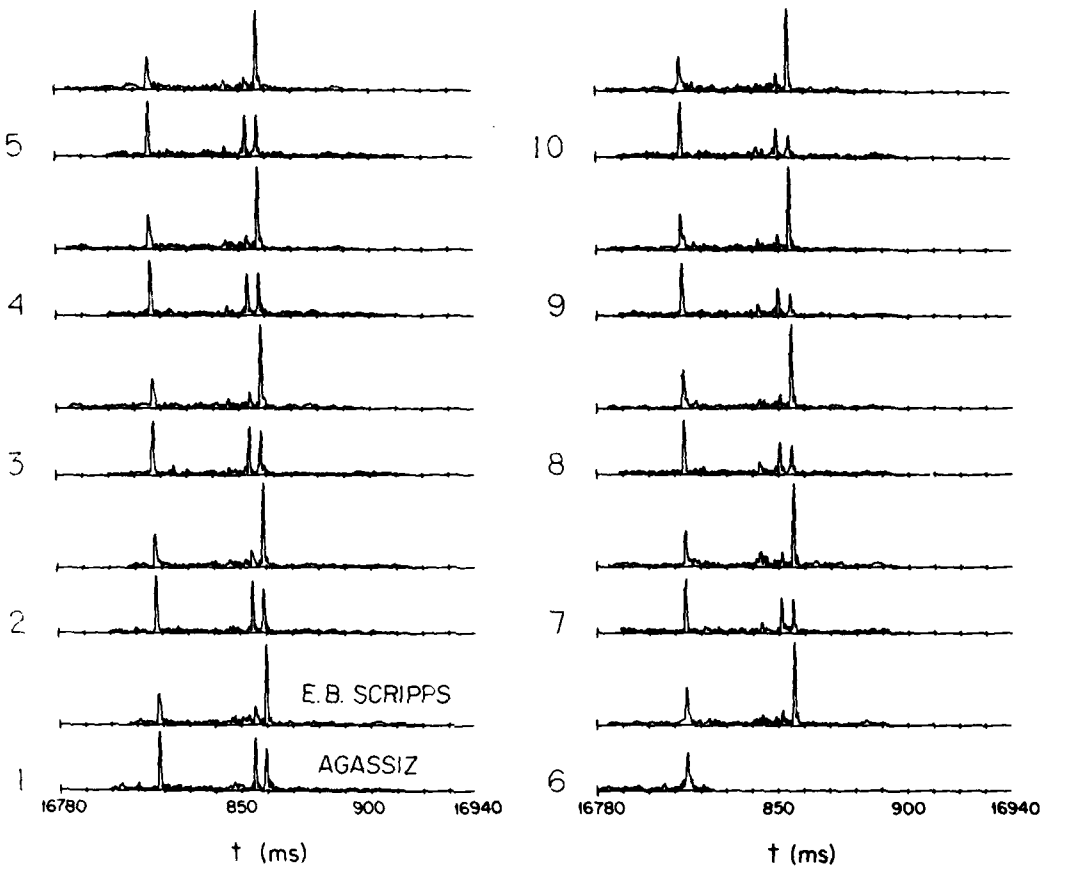


Figure 9—Scripps and Agassiz arrivals resulting from reciprocal transmissions at 30-s intervals after high-resolution processing (April 3, 1976; 1444 to 1449). Time marks are in milliseconds from the start of the transmitted pulse. Scripps arrivals slightly lead Agassiz arrivals, in part due to the ship's drift relative to the water column. Note the pronounced doublet in the upper path to the Agassiz, as compared to the upper Scripps arrivals. This may be the result of current shear.

WHY DO IT THE HARD WAY?

The question inevitably arises as to why we wish to employ such expensive and difficult methods for measuring the oceans, when all we have to do is to dunk thermometers or collect water samples. The reason is that most of the important ocean processes have large scales, 100 km or more, and these are better monitored by measuring appropriate spatial *averages* than by measuring the precise characteristics of the water that wets the thermometer. To some extent, time averages of the spot values can substitute for space averages, but this can be hazardous.

In the proposed arrangement (Figure 2) the average is taken along two ray paths, an upper and a lower. This is not an adequate vertical sampling of the oceans. Considering that the gravest two or three internal modes carry nearly all the energy, one might be able to get away with a dozen independent depth averages. There are two possible ways of doing this: (a) placing multiple transmitters and receivers at various depths along each mooring and (b) using larger horizontal separa-

tions and, accordingly, a larger number of ray paths (provided these can be resolved). The number of independent ray paths increases by about two per convergence zone (≈50 km), and the separation between arrivals is on the order of D^{-1} s, where D is the distance measured in convergence zones.

Consider relative correlation scales for ocean and atmosphere:

	Ocean	*Atmosphere*
Horizontal:	100 km	1000 km
Time:	60 days	3 days

Since it is easy to sample densely in time and difficult to sample densely in space, the oceanographer suffers an immense disadvantage relative to his meteorological colleague. Perhaps the ocean surface layers can be adequately monitored from satellites, but what about the ocean interior? It would take 10^5 stations to cover the world's oceans at 50-km spacing! Very extensive automated buoy networks have in fact been proposed. If one could measure *between* stations rather than *at* stations, then the information collected goes up with station number, like n(n-1) rather than n (allowing for reciprocals).

The discussion has dealt with features that are large compared to the buoy spacing. These are then deterministically sampled. Features small compared to the buoy spacing can be probabilistically sampled. This includes not only the internal waves, but also surface waves, using the surface scattered arrivals. With regard to surface and internal waves, such scattering experiments have the advantage of providing direct information about statistical properties, unlike the usual methods of repeated soundings (say) subsequently analyzed for statistical properties.

We have here discussed in some detail one particular experiment, selected simply because we are most familiar with it. There are, of course, many other experiments that deserve discussion. In particular, we want to mention some very recent work by R. Pinkel, who displays high-frequency, horizontally backscattered acoustic energy in range-doppler space, and from this infers the horizontal velocity field.

THE FUTURE

How do present opportunities in oceanography and acoustics compare with the opportunities when ONR was born? It would seem to us they are just what one would expect from 30-year-olds. The problems are more difficult. The approach is perhaps more responsible and better disciplined. But the opportunities for the three siblings are there, challenging as ever.

GLOSSARY

Dynamic height—The relative depths of isobars in the ocean, commonly measured in dynamic meters (gz/10). By assuming that a given reference isobar is a level surface, one can compute pressure gradients.

Geostrophic current—A current in which the pressure gradient and Coriolis forces approximately balance. The current flows along isobars.

Internal waves—Wave propagation occurring in the ocean interior, with the restoring force due to the density stratification of the ocean.

Inertial frequency—The vertical component of twice the Earth's angular velocity. A particle of water acted upon only by the Coriolis force will describe a horizontal circle in an inertial period.

Mediterranean outflow—Relatively high-temperature, high-salinity water flowing out of the Mediterranean Sea through the Straits of Gilbraltar and spreading laterally at about 1000-m depth.

Brunt-Väisälä frequency—The natural frequency of oscillation of a vertical column of fluid given a small displacement from its equilibrium position in a stably stratified medium.

Swallow float—An instrument developed by Dr. J. Swallow to provide Lagrangian velocity measurement at great depths. Since aluminum tubes are less compressible than seawater, it is possible to design a package that is heavier than water at the surface but that becomes neutrally buoyant at some predetermined depth; the instrument will then move with the water at that depth. Swallow tracked the motion from a ship by monitoring an acoustic source on the instrument.

Rossby waves—Waves below inertial frequency. These are approximately geostrophic, representing a balance between Coriolis force and horizontal pressure gradients.

REFERENCES

1. J. Crease, "Velocity Measurements in the Deep Water of the Western North Atlantic," *J. Geophys. Res.* **67**, 3173-3176 (1962).
2. L. M. Brekhovskikh, et al., "Some Results of a Hydrophysical Experiment on a Test Range Established in the Tropical Atlantic Ocean," *Izv.* **7**, 332 (1971).
3. A. R. Robinson, "The Variability of Ocean Currents," *Revs. Geophys. Space Phys.* **13**, 598-602 (1975).
4. J. C. Steinberg and T. G. Birdsall, "Underwater Sound Propagation in the Straits of Florida," *J. Acoust. Soc. Am.* **39**, 301-315 (1966).
5. J. G. Clark and M. Kronengold, "Long-Period Fluctuations of CW Signals in Deep and Shallow Water," *J. Acoust. Soc. Amer.* **56**, 1071-1083 (1974).
6. F. Dyson, W. Munk, and B. Zetler, "An Interpretation in Terms of Interal Waves and Tides of Multipath Scintillations Eleuthera to Bermuda," *J. Acoust. Soc. Amer.* **59**, 1121-1133 (1976).
7. W. H. Munk and F. Zachariasen, "Sound Propagation Through a Fluctuating Stratified Ocean: Theory and Observation," *J. Acoust. Soc. Amer.* **59**, 818-838 (1976).
8. T. E. Ewart, "Acoustic Fluctuations in the Open Ocean—A Measurement Using a Fixed Refracted Path," submitted to *J. Acoust. Soc. Amer.*
9. A. W. Ellinthorpe, "The Azores Range," NUSC, Tech. Doc. 4451.
10. A. W. Ellinthorpe and H. A. Freese, "Exploitation of the Azores Fixed Acoustic Range (AFAR) through May 1973," NUSC Conf. Tech. Rep. 4647, 1973.
11. A. W. Ellinthorpe and A. H. Krulisch, "Preliminary Account of AFAR Microstructure Measurement Operation," NUSC Tech. Memo. No. TE-105-75, 1975.
12. R. P. Porter and R. C. Spindel, "Low Frequency Acoustic Fluctuations and Internal Gravity Waves in the Ocean," in press.

Dana R. Kester is a Professor of Oceanography at the University of Rhode Island's Graduate School of Oceanography. Dr. Kester's main research interests are in the physical chemistry of metals in seawater and in chemical distributions in the oceans. He was born in Los Angeles, Calif.; received a B.S. degree in oceanography and chemistry from the University of Washington; and earned M.S. and Ph.D. degrees from Oregon State University.



IMPROVING THE CHEMICAL BEHAVIOR OF METALS IN THE OCEAN ENVIRONMENT

Dana R. Kester

Graduate School of Oceanography
University of Rhode Island, Kingston, R.I.

Deterioration of structural metals is a significant limitation to man's activities in the marine environment. Corrosion of metals presents a significant economic factor in oceanic work because it requires continual maintenance and periodic replacement of materials. In addition, there are increased costs attributable to corrosion when one considers the need for highly reliable performance of structures and devices exposed to the marine environment for moderate periods of time. As we look ahead to our future needs for structural materials in the ocean it is useful to consider the following aspects of the problem: (a) the general characteristics of the marine environment, (b) the various corrosion processes and (c) the mechanisms for preventing corrosion. This article will focus primarily on the behavior of iron in marine systems because iron is a predominant component of steel, which is widely used for marine applications.

CHEMICAL ENVIRONMENTS IN MARINE SYSTEMS

We can recognize that a large variety of environments exist in marine systems and that the behavior of materials will differ among these environments. One example is the marine atmospheric environment in which materials are exposed to sea spray, highly oxidizing conditions, periodic wetting and dehydration, and concentrated salt films. These conditions represent one extreme for material exposure. Another set of conditions is found in continuous exposure to seawater from the ocean surface to the seafloor. This is an environment in which the basic chemical constituents of seawater as well as physical and biological processes have a direct impact on the deterioration of metals. Corrosion is primarily an electrochemical phenomenon, and the electrical conductive properties of seawater are a major factor in metal deterioration. The sea floor sedimentary environment represents a third set of conditions to which materials are exposed. This region is characterized by substantial chemical gradients and relatively slow migration of chemicals by diffusion. Under these conditions it is possible for two ends of a piece of metal to be subject to different chemical environments and different corrosion results. In addition to these three general types of marine environments we must recognize the importance of microenvironments. These are localized regions near the surface of a metal which may be much different chemically from the bulk seawater in the vicinity of the metal. Microenvironments can be created beneath marine organisms that attach themselves

to metal surfaces, and they can be created in crevices, cracks, and pits on metal surfaces.

The range of chemical environments found in the ocean is important in considering corrosion processes. We normally regard seawater to be an aqueous solution with a total salt content ranging from 30-36 g of salt per kilogram of solution, with a pH of approximately 8, and normally containing dissolved oxygen. However, in considering the behavior of metals and the design of studies to evaluate corrosion and its prevention, it is important to take a broader view of environmental conditions. Systems exposed to the atmosphere may experience very concentrated sea salt solutions due to evaporation of water. In some portions of the marine environment all the oxygen is consumed from the seawater and hydrogen sulfide is produced by microbial degradation of organic material. In some microenvironments pH may range from a relatively acidic value of 2.5 to very alkaline values of 12.5. The chemical behavior of metals such as iron will vary dramatically over this extreme range of conditions.

Table 1 provides an indication of the magnitude of various parameters that are important in the corrosion process for three regions of the marine environment—the open ocean, near the seafloor, and in the interstital waters of ocean sediments. The values in this table represent typical ranges found in seawater. However, significantly more extreme conditions can be found in microenvironments and upon evaporation of water from seawater. Oxygen is a primary constituent in most corrosion reactions; it enters the ocean from the atmosphere and is consumed by biological respiration. The smallest oxygen concentrations generally occur at depths of 500-1500 m. The pH and Eh are important factors in determining the chemical reactivity of a metal. Chloride (Cl^-), sulfate (SO_4^{2-}), and bicarbonate (HCO_3^-) are some of the major components of seawater that reflect many of its chemical and physical properties, such as electrical conductivity and acid-base buffering. Phosphate (HPO_4^{2-}) and ammonia (NH_3) are biologically active chemicals in the ocean. The remaining five parameters listed in Table 1, temperature, pressure, conductivity, bacterial concentrations, and water velocity, describe some of the general environmental factors important in metal corrosion.

TYPES OF CORROSION

Corrosion is not a single process. There are a variety of different mechanisms by which a material may deteriorate. One of the most obvious types of corrosion is a general wasting away of the surface of a metal due to the chemical attack of seawater and an electrolysis of one portion of a structure relative to another. However, it is more common for corrosion to occur at specific places in a structure. This localized deterioration results from differences in the chemical environment to which the metal is exposed, such as the degree of stagnation of the water near the surface and the formation of chemical microenvironments. Alternatively the localized attack may result from differences in the quality of the metal due to inhomogeneities in its chemical properties, passive surface films, and surface protective coatings. *Crevice corrosion* represents one type of deterioration which occurs in places that, due to mechanical design, have restricted exchange of seawater with their surroundings. A second mechanism is *pitting corrosion*, in which there is a localized attack of the metal at particular locations on an otherwise flat surface. Pitting corrosion can occur in the steel plates of ships and other structures, in the structural member of devices, in piping systems that transport seawater, and in the linings of tanks that contain seawater. Pitting corrosion is particularly significant because of its highly localized nature; it is necessary for only a small fraction of the metal to deteriorate before the structure is functionally disabled.

Charles G. Munger recently reported an interesting study of pit corrosion in the tanks of oil transport vessels. He examined the corrosion pits that occur in the horizontal stiffening members of tanks that periodically contain oil, air, and seawater. The degree of pitting varied with the length of time of exposure; in the cases examined it varied from 1 to 10 pits per square foot. The size of the pits ranged from 1/8 to 8 in. in diameter. Their depth varied from 1/16 to 3/4 in. A striking observation was that the pits occurred only on the upper surfaces of horizontal structures—they were not found on the undersides of the stiffeners or on the vertical sections of the tanks. The detailed processes and the factors reponsible for this

type of corrosion are not well understood. However, it is an area of intense study at present.

A group of ocean engineers at the University of Rhode Island have developed techniques to simulate pit corrosion on a scale that is large enough and rapid enough to monitor under laboratory conditions. Some of their observations and the relationship between the chemistry of iron and pit corrosion were described in a 1974 issue of Naval Research Reviews.

In their test cells spontaneous corrosion processes occurred, which resulted in stratification of the seawater into a very acidic region and a highly alkaline layer. The corrosion reactions resulting from this system produced a bridge of iron corrosion products at the interface between the two stratified portions of seawater. The characteristics of this system closely resemble naturally produced pits.

Figure 1 is a schematic illustration of some of the individual factors in the corrosion process. Electrons in the metal are drawn away from the

Table 1

Comparison of Selected Parameters in the Ocean, near the Seafloor, and Within the Sediment

Parameter	*In the Ocean**		*Above the Seafloor** (4000 m)	*Interstitial Water*
	Maximum	*Minimum*		
O_2 (ml/l)	8.0	0.1	4.0	?
pH	8.3	7.6	7.8	7.0 to 8.4, ¶, ‡
Eh (V)†	+0.5	+0.3	0.4	−0.3 to +0.45, **, ‡
Cl^- (‰)	20.5	16.6	19.2	14 to 21¶ ††
SO_4^{2-} (M)	0.030	0.023	0.029	0.03 to 0.06‡‡
HCO_3^- (M)	0.0025	0.0020	0.0024	0.0007 to 0.007‡‡
HPO_4^{2-} (μM)	3	0	2	?
NH_3 (μM)	1.6	0	0	0 to 500‡ ‡‡
T (°C)	27	−2	1	1 to 2
P (atm)	1000	1	400	400
Conductivity (Ω cm^{-1})	0.058	0.025	0.029	?
Bacteria (cells/g)‡	10^3	10^{-1}	$1 - 5 \times 10^{-1}$	up to 1×10^8
Velocity (cm/s)	200	0.1	1	----------

*These values are taken from Riley and Skirrow (1965) unless noted otherwise.
†Garrels and Christ (1965)
‡Zobell (1946); Sieburth (1960)
‡Rittenberg *et al.* (1955)
¶Siever *et al.* (1965)
**Whitfield (1969)
††Fanning and Schink (1969)
‡‡Bruyevich (1966)

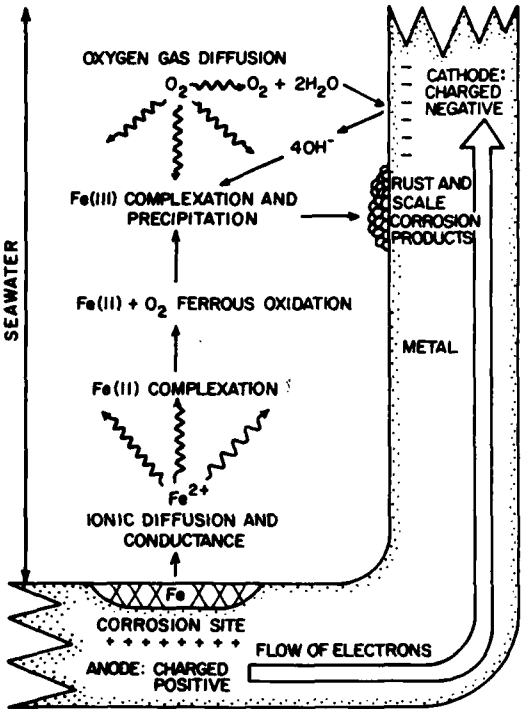


Figure 1—Schematic of iron corrosion in terms of various chemical processes.

active corrosion site (the anode), and ferrous metal ions are released to the seawater. These ions migrate away from the corrosion site by diffusion (and possibly by conductance), and then they enter into complexation reactions with seawater constituents. As the Fe(II) reaches a region where dissolved oxygen is present it oxidizes to form Fe(III), which then is subject to complexation and precipitation reactions. The result of these processes is a removal of metal from the corrosion site and deposition of rust and scale corrosion products around it. This system represents an electrical circuit in which electrons flow through the metal from anode to cathode and through the seawater by ionic transport. The process can be stopped by preventing the flow of electrons at any point in the cycle.

Improved understanding of a specific corrosion process such as pit formation can be achieved through case studies similar to that of the oil tankers, through laboratory simulations, and through applications of basic chemical knowledge. Most of the attention in preventing corrosion has been directed at the surface characteristics of the metal. Is it possible to envision a metallic material that would not permit the flow of electrons in the metal from anode to cathode? If the cathode and anode are forced to be very close together there may not be sufficient chemical energy to set up the galvanic potential required to drive the process. This could be achieved by a material in which the metal is blocked into cells separated by a nonconducting membrane, in much the same way as cellulose partitions plant material. Even though this idea presents many impracticalities such as maintaining strength while preventing conduction across the metal cell boundaries, its development might result in new insights into the corrosion process and its prevention.

Other types of corrosion include *erosion* and *cavitation corrosion* in which the force of fluid flow on the material promotes its deterioration. Another type of metal failure is *stress corrosion cracking*. Some metals are more subject to chemical attack when under mechanical stress than in an unstressed condition.

It is evident from the diverse range of corrosion mechanisms that a variety of factors must be considered in order to improve the performance of materials in the marine environment. No single approach to the problem will be sufficient, and no matter how well conceived a scheme may be for a particular type of corrosion, a lack of awareness of all the factors that may contribute to material failure can have costly or disastrous results.

PREVENTING CORROSION

We may identify four basic approaches to preventing corrosion: material science, sacrificial electrolysis, protective coatings, and mechanical design. The selection of materials for their chemical as well as structural properties can optimize their performance. The development of special alloys for applications in marine environments is a good example of this approach. From recognition of the underlying electrochemical nature of corrosion it has become common to use sacrificial electrolysis of nonstructural devices such as zinc

anodes; this technique has proved highly effective. Considerable attention has been given to developing surface coatings to minimize corrosion. These coatings may have two roles; one is to prevent fouling which leads to corrosive microenvironments, and the other is to provide a nonreactive barrier between metal and seawater, such as plastic paint. Coatings generally require careful maintenance, and it is often difficult to achieve a uniform and strongly bonded barrier. Defects in the coatings provide an opportunity for highly detrimental localized corrosion. The fourth factor in corrosion prevention is to design structures in a manner that minimizes areas of restricted water flow.

APPROACHES TO A BETTER UNDERSTANDING OF CORROSION

One approach to minimizing the consequences of corrosion is through studies of material science. It is likely that continued search for metal alloys will yield improved performance. Laminated structures in which a metal core provides the desired strength characteristics and a plastic, ceramic, or fiberglass shell provides the chemical inertness in seawater, may lead to new improved capabilities. A major difficulty to be overcome is the lack of strong chemical bonding between the metal and the nonmetallic surface layer. This approach would be of limited usefulness for applications in which mechanical wear might abrade, scratch, or chip the surface.

The most direct approach to studying the deterioration of a material in the marine environment is by exposure tests in which sample panels are immersed in the environment for a period of time and the consequences observed (this method is referred to as in-situ tests). A recent study by K. D. Efird demonstrated the relative behavior of a variety of metals after exposure to seawater. He was able to relate the results to the tendency of the metal to form passive and toxic films in seawater. While this empirical approach provides direct information, it is difficult to relate the observations to other environmental conditions and to separate the effects of various parameters on the corrosion process.

Dr. N. T. Monney has advocated the development of a complete test facility for corrosion studies so that individual variables may be controlled and altered to provide a better understanding of corrosion processes. Some of the primary environmental variables that affect metal deterioration are pH, oxygen, sulfide, chloride, temperature, pressure, and water speed. A corrosion-research test chamber could be designed to permit control of these variables. Experiments with such a system would provide a useful complement to in-situ observations, but it should not be regarded as a facility for duplicating the marine corrosion environment. In addition to the purely chemical and physical variables, bacteria and fouling organisms contribute significantly to corrosion. It is unlikely that the present technology and basic knowledge is adequate for simulating the effects of biological organisms on corrosion in a synthetic environment.

Another approach toward improving our ability to minimize corrosion is to achieve an understanding of the chemical behavior of metals in the marine environment. The chemical reactivity of a metal such as iron as reflected in its geochemical cycle and in its interactions with the various components of seawater. The complexation of iron with chloride, hydroxide, and other constituents in seawater will determine many aspects of its chemistry, such as its net electrical charge and the solubilities of its solid phases. Studies of the kinetics of oxidation-reduction reactions will provide a basis for predicting the rates of corrosion under various conditions.

Figure 2 illustrates some of the chemical characteristics of iron in marine systems. Each box represents a particular facet of iron chemistry, and the arrows between them represent chemical processes. The dissolved forms of iron are of two general types; ferrous, Fe(II) and ferric, Fe(III). Research in our laboratory has led to an evaluation of the chemical forms of these two oxidation states of iron in marine systems. The percentages in the boxes of Figure 2 reflect the distribution of ferrous and ferric iron among their principal chemical forms. Fe(II) occurs as $FeOH^+$, $FeCl^+$, and Fe^{2+}, whereas Fe(III) can be represented by $Fe(OH)_3^\circ$ and $Fe(OH)_2^+$. This illustration is for seawater at a pH = 8. It is possible to predict the changes that occur in these chemical forms over the extreme pH values encountered in marine systems. These two types of

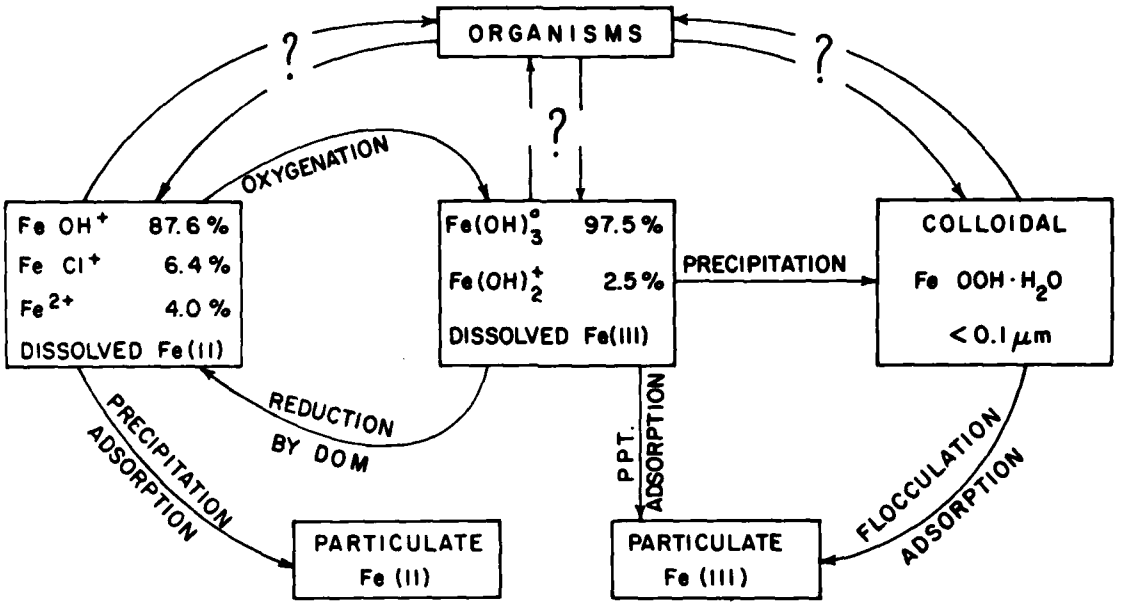


Figure 2—Chemical cycle of iron in seawater, including the dissolved forms of ferrous, Fe(II), and ferric, Fe(III), iron, the particulate and colloidal phases, and the uptake and excretion by organisms. The DOM refers to naturally occurring dissolved organic matter in seawater.

iron are linked to each other by oxidation and reduction processes. One of the current areas of research is an investigation of the rates of these reactions under marine conditions, which relates directly to some of the chemical factors involved in corrosion of metals. These dissolved forms of iron can undergo precipitation and adsorption reactions to become associated with particulate and colloidal phases. A major area requiring future work is the behavior of colloidal iron in seawater. The behavior of this material is very important in considering the dispersion of corrosion products. We can also recognize that the various chemical forms of iron can be taken up and released by marine organisms because iron is an essential element for metabolic processes. Very little is known about the relative preference of organisms for the different forms of iron, so these transfer processes have been designated by question marks in Figure 2. Studies of marine chemistry that give information on the chemical reactivity of iron in seawater provide one basis for understanding corrosion processes and their prevention.

Another area of research in the marine environment that relates to some of the characteristics of metal performance in the ocean concerns the chemical exchange betweeen sediments and seawater. This has been an area of active study in recent years. We can look forward to increased knowledge becoming available during the next several years as a result of this work. Some of the factors being considered include chemical reactions between sediments and their interstitial waters, the flux of materials from the sediments to the overlying seawater, and the role of burrowing organisms in stirring up the sedimentary environment.

FUTURE DIRECTIONS IN CORROSION RESEARCH

Advances in knowledge of corrosion processes can be best achieved through a balanced and coordinated attack on the problem. It is unlikely that any single approach will be adequate to assure improved performance of metals in the

marine environment. We can expect that studies of metal alloys, protective coatings, and nonmetallic materials will produce improved chemical and structural properties. In-situ exposure tests will continue to provide an effective means of comparing the behavior of different metals. Exposure tests are also valuable, because they reveal the net effect of the corrosion process for a particular environment. One of the limitations of in-situ studies is that it is difficult to isolate the effects of critical variables. Studies of the physical chemistry of metals in the marine environment will provide new capabilities for predicting corrosion processes. This area of work will also create a basis for corrosion prevention.

One way to accelerate the progress in a program of research and development is to form a task force of individuals who can represent the various components of the effort and have this group periodically review the status of the problem and the progress of the individual components. This group could make a particular effort to integrate the results of the various studies, observing how the pieces of the puzzle fit together, and identifying the critical gaps which require more attention.

BIBLIOGRAPHY

S. W. Bruyevich, *Chemistry of the Pacific Ocean*, vol. 3, 549 p., Academy of Sciences of the U.S.S.R., Institute of Oceanology, Moscow, 1966. (English translation by I. Evans for the U.S. Naval Oceanographic Office, Washington, D.C.).

K. D. Efird, The Inter-relation of Corrosion and Fouling for Metals in Seawater. *Mater. Protect. Performance* 15 (4); 16–25 (1976).

K. Fanning and D. R. Schink, "Interaction of Marine Sediments with Dissolved Silica," *Limnol. Oceanogr.* 14; 59–68 (1969).

F. W. Fink and W. K. Boyd, *The Corrosion of Metals in Marine Environments*, 87 p., Bayer and Company, Columbus, Ohio, 1970.

R. M. Garrels and C. L. Christ, *Solutions, Minerals, and Equilibria*, 450 p. Harper and Row, New York, 1965.

D. R. Kester, "Chemistry of Iron in Marine Systems," *Nav. Res. Rev.* 27 (9); 3–16 (1974).

T. J. Lennox, "On Marine Corrosion," *Mater. Protect. Performance* 12 (1); 6–8 (1973).

N. T. Monney, "Deep Ocean Corrosion: Simulation Facilities vs *In Situ* Research," *Mater. Protect. Performance* 12 (1); 10–13 (1973).

C. G. Munger, "Deep Pitting Corrosion in Sour Crude Oil Tanks," *Mater. Protect. Perform.* 15 (3); 17–23 (1976).

J. P. Riley and G. Skirrow, *Chemical Oceanography*, vol. 1, 712 p., Academic Press, London, 1965.

S. C. Rittenberg, K. O. Emery, and W. L. Orr, "Regeneration of Nutrients in Sediments of Marine Basins," *Deep Sea Res.* 3, 23–45 (1955).

J. McN. Sieburth, "Soviet Aquatic Bacteriology: A Review of the Past Decade," *Quart. Rev. Biol.* 35, 179–205 (1960).

R. Siever, K. C. Beck, and R. A. Berner, "Composition of Interstitial Waters of Modern Sediments," *J. Geol.* 73, 39–73 (1965).

M. Whitfield, "Eh as an Operational Parameter in Estuarine Studies," *Limnol. Oceanogr.* 14 (4), 547–558 (1969).

C. E. Zobell, *Marine Microbiology*, 240 p., Chronica Botanica Co., Waltham, Mass., 1946.

John D. Costlow, Jr., is Director of the Duke University Marine Laboratory at Beaufort, N.C. Dr. Costlow was Resident Liaison Scientist in Marine Biology and Oceanography (1965-1966) and Visiting Liaison Scientist (1968-1974) at the U.S. Office of Naval Research in London. He has served as a member of the planning committees for the First Estuarine Research Conference (Jekyll Island, 1964) and the Second Estuarine Research Conference (Myrtle Beach, 1973); of U.S. Scientific Committee on Ocean Research; of the Subcommittee on Biological Oceanography, National Academy of Sciences; of the U.S. Delegation, International Association of Biological Oceanographers; of the U.S. Working Group, U.S.-U.S.S.R. Cooperative Program in Ocean Sciences; of the Panel on Underseas Facilities, National Academy of Engineering; and of the North Carolina Marine Fisheries Commission. He is the author of more than 100 scientific publications on development and growth in barnacles, larval development of marine invertebrates, larval physiology, and endocrinology. At ONR London, he contributed more than *100 articles to European Scientific Notes.* He also edited several volumes of proceedings of international symposia. Dr. Costlow was born in Brookville, Pa. He earned a B.S. at Western Maryland College and a Ph.D. at Duke University.



MARINE BIODETERIORATION

John D. Costlow, Jr.

Duke University Marine Laboratory
Beaufort, N.C.

MARINE BIODETERIORATION

As man continues to expand his abilities to utilize the marine environment, as well as actually working and living in the oceans, he becomes more aware of the deficiencies in his understanding of marine organisms, the environment in which they live and breed, and the way in which some of them deleteriously affect the structure he places in and under the oceans. Over the past 30 years considerable progress has been made in identification of marine organisms in the estuarine and coastal waters of the continents of the world. Basic information is available on the species found in specific geographical areas and the way in which seasonal variations in the marine environment may contribute to their spawning, development, and survival. Within those species which comprise the "fouling community," additional information is becoming available on the tolerance of adult animals to a variety of physical and chemical parameters of the marine environment, the way in which these factors contribute to the establishment and maintenance of the fouling community, and an increasing awareness of the fact that complete control of the adults in the fouling community, by biological or chemical means, is virtually impossible at present. On the basis of our present knowledge eradication of fouling communities from most large, manmade structures can be accomplished only by periodically removing the organisms from the underwater surfaces.

Information derived from research over the past two decades indicates that there are, at specific points in the life histories of most fouling and boring organisms, a number of physiological processes that are beginning to be better understood. Complete understanding of these processes could permit interference that would lead to control of the major species responsible for the destruction of manmade structures and of the tremendous reduction in efficiency of surface and underwater vessels. In the life history of virtually all marine animals it is important to consider not only the general relationship between the animal and the environment but also to consider the biological mechanisms, at all levels, that regulate the capacity of the animal to adapt to various physical, chemical, and biological factors in the environment.

For adult animals, a number of general aspects of the reproductive processes have been described, and information is available on the seasonal variations in spawning, the morphological adaptations that enhance the process of fertilization, improve the capacity for retention of eggs and embryos, and permit the release of larvae into the marine environment at a time that is most

favorable for their survival and dispersal and subsequent settling, attachment, and metamorphosis. It is only through such successful adaptation in each phase that the organisms can survive as adults and produce successive generations. Apart from general descriptions, however, there is a paucity of information on those mechanisms known to regulate the various phases of the reproductive process. A number of workers have described the sites of endocrine activity in higher Crustacea, which determine sex in the adult animals. Virtually nothing is known, however, of how the determination of sex is regulated in the lower crustaceans such as the barnacles or in those bryozoans, tunicates, and molluscs that are frequently found as members of the fouling community on submerged and intertidal surfaces. We are beginning to understand how hormones control the proliferation of gonadal tissue and the subsequent development of eggs and sperm. These processes in many species of the fouling community, however, are still undescribed. The process of fertilization and the release of eggs and sperm into the water is known to be controlled in some marine species by hormones and pheromones, but little information exists on the sites of synthesis of these compounds, their chemical nature, or the location and function of the chemoreceptors necessary to detect the pheromones in the water column and effect the release of gametes.

For insects, which represent the arthropods of the terrestrial environment in the same way that crustaceans represent arthropods in the marine environment, the presence of a particular chemical, the juvenile hormone, has been clearly demonstrated. The way in which this hormone regulates the transition from larval stages to the juvenile and adult stages of insects is well documented. Although substances possessing an activity similar to the juvenile hormone of insects have been described for a few marine crustaceans, virtually nothing is known about how it may regulate development of those larval stages that abound in plankton and are responsible for the distribution of many of the fouling organisms. The recent development of juvenile hormone mimics, compounds that simulate the activity of the natural juvenile hormone, has led to a number of studies on the way minute amounts of these mimics may inhibit the normal sequence of development of insects. This inhibition culminates in the prevention of metamorphosis to the adult stage and is looked upon as having potential for control of such undesirable insects as mosquitoes and flies. Virtually nothing is known, however, about how these mimics might be used to inhibit the successful completion of all larval stages of marine crustaceans or if developmental stages of other groups in the fouling community may be regulated by similar chemicals. These same mimics have been shown to interfere with the orderly progression of events in the development of gonads, eggs, and sperm in a number of the insects as well as in a few of the marine crustaceans, but their effects on the reproduction of barnacles and other members of the fouling and boring communities have not been considered.

The success of the major species of marine invertebrates associated with the fouling communities as adults depends on the success of the planktonic larvae released by the millions at periodic intervals each year. Not only must the larvae survive and complete a number of stages leading to settling, metamorphosis, and growth as sessile juvenile animals, but they must also be successfully distributed in areas in the natural environment that are conducive to normal growth and maturation of successive generations. It is in the larval stages that we have the least understanding of the many processes that are normally integrated and regulated so as to ensure continuation of the species, insuring its distribution into new areas where possible extension of its geographical range may occur.

Although we are beginning to have some appreciation of the morphological characteristics of larvae in the major groups of invertebrates, especially those normally found as representatives of the fouling and boring communities, our understanding of their microstructure, physiology, behavior, endocrinology, and distribution is still quite limited. Gross morphological features of the nauplii and cyprid stages of the acorn and stalked barnacles have been known for more than 100 years. It has only been since the development of the transmission electron microscope and the scanning electron microscope that we have begun to appreciate the complexities of the microstructures and their possible role in the adaptive pro-

cess of the larval stages (Figures 1 and 2). Recent studies have elucidated the details of morphology of many of the appendages of the cyprid stage, the final stage of the barnacle prior to settling and metamorphosis. The function of most of the detailed anatomical structures that have been described, however, is still unknown. The behavioral response of planktonic organisms to temperature, light, and pressure, including those of larval stages of species in the fouling community that spend only a part of their life in the plankton, has been the subject of a number of investigations but we still lack a clear-cut picture of how these stimuli and the responses of the planktonic organisms relate to their survival and distribution. Although it is generally recognized that planktonic forms exhibit extreme sensitivity to light, little is understood of the photoreceptors which are present in virtually all larval stages, the way in which the nervous system and endocrine systems integrate the stimuli and coordinate the response, or why extreme differences may exist in the responses of different species of planktonic organisms normally found in the same portion of the water column. The basic phenomenon of diurnal vertical migration, displayed by both zooplankton and phytoplankton, has been known for many years, but additional information is still needed to differentiate the relative contribution of rhythms and activity as contrasted to responsiveness to light, gravity, currents, and pressure. Laboratory investigations thus far have concentrated largely on the mechanisms of orientation to light as well as on the physiology of photoresponses, but they have demonstrated little interest in contributing to the overall picture and describing the real ecological implications. It has been difficult, therefore, to determine the factors that initiate, control, and orient the zooplankton associated with diurnal vertical migration. Recent studies on the response of crustacean larvae to light have clearly demonstrated that the normal response, depending on the stage of development, can be completely reversed by the presence of certain chemical compounds in the water column. The extent to which this reversal could be used for antifouling purposes, however, remains to be determined.

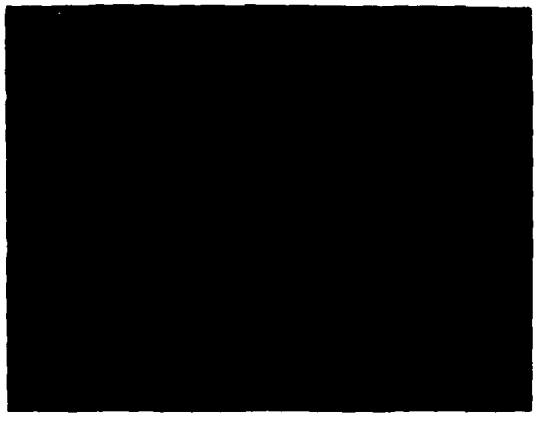


Figure 1—Scanning electron micrograph of cyprid of the barnacle, a common fouling organism. Times 150. Courtesy of T. West, Duke University Marine Laboratory, Beaufort, N.C.



Figure 2—Scanning electron micrograph of cuticle and specialized seta on surface of cypnd of barnacle. Times 7200. Courtesy of T. West, Duke University Marine Laboratory, Beaufort, N.C.

Although considerable effort has been expended to delineate and understand the hormonal or endocrine mechanisms responsible for regulating development in insects, virtually nothing is known of similar mechanisms in the development of marine invertebrates. The regular sequence of ecdyses, or molting, the casting off of the old exoskeleton prior to increase in size, is a well documented occurrence for many crustaceans both during larval development and as adults. Molting in the larval stages is known to be accompanied by a regular and sequential increase in morphological complexity leading to the final larval stage, which then metamorphoses to the juvenile. In the case of most crustaceans that are

fouling organisms, a sessile adult animal results. It has been demonstrated that the removal of certain portions of the nervous system in the developing larvae of higher crustaceans results in the disruption of the regular and sequential occurrence of morphological features. The absence of portions of the central nervous system, perhaps through removal of a regulatory hormone, can also disrupt a number of physiological processes, including the capacity of the larva to regulate the chemistry of the blood relative to the chemistry of the external seawater environment. Virtually nothing is known, however, of the chemical nature of this compound or where it is synthesized in the larval stages during development.

We do not know how and when the synthesis of these compounds is activated in the larval stages, the way the sites of synthesis may be modified at the time of metamorphosis to the juvenile stages, or the physiological and biochemical pathways that are followed and that might be used to interfere with or inhibit these natural stages of development. It has also been shown that removal of specific portions of the nervous system during the larval stages of higher crustaceans can result in the production of extra larval stages and in the precocious development of reproductive tissues during the early juvenile stages. Studies over the past decade have made it apparent that in the relatively small and insignificant larvae a variety of processes are controlled and regulated in such a way as to achieve the ultimate goal of the mature animal. Virtually none of these processes, however, are described or understood, and until we have a complete picture any effort to interfere with the sequence or inhibit some specific link in the process will be impossible.

On the successful completion of the larval stages in the plankton, two remaining processes are crucial to continuation of the species: (a) settling in an environment favorable to the adult and (b) successful metamorphosis from the final planktonic stage to the sessile juvenile. Workers for many years have tried to determine and describe those stimuli, chemical or tactile, that induce successful settling and cause the initiation of metamorphosis in a number of the sessile forms within the fouling and boring community. The detection of stimuli that lead to settlement is generally assumed to be a function of sense organs found in the appendages of the last larval stage, although the evidence thus far is largely circumstantial rather than experimental. Virtually nothing is known of the location and integration of chemoreceptors of the barnacle larva, although a number of studies have demonstrated that these microscopic animals do respond to environmental stimuli, including chemicals, pressure, and specific wavelengths of light. Their ability to detect and settle in an area previously populated by adults of the same species has been documented but the chemical and biological interactions necessary for such a behavioral response are virtually unknown.

In the process of settling, the successful activation of adhesive glands for attaching the organism to the substratum has been extensively studied. The investigations to date have concentrated largely on the source of the cementing substance and the way it is applied to the surface. Only recently has progress been made toward understanding the chemical composition of the cementing substance and the way this substance is synthesized by the adult animal as it increases the size of its attached surface over the substratum. The mechanisms that activate periodic secretion of the cementing substance, thought to be hormonal, are not known, although there are suggestions that they may be associated with the regular sequence of molts, which continues throughout virtually the entire adult life of the barnacle. The adhesives that permit the noncrustacean members of the fouling community to attach to the substratum are also completely unknown, as are their means of synthesis. The process of molting, essential to the growth and maturation of most crustaceans, is known to be controlled in the higher forms by hormones, synthesized and stored in specific portions of the central nervous system. There is evidence that at least one of the same hormones is also responsible for regulation of molting in juvenile and adult barnacles. Thus far, however, sites of synthesis and mode of action of the molt-accelerating hormone are not known for any of the lower crustaceans.

The complete reorganization of all internal body systems and the elaboration of a new integument or shell is another phase in the growth of many marine organisms that is only partly understood. A complete understanding of growth of

many fouling organisms is, to a considerable extent, an understanding of the mechanisms of calcium carbonate deposition and factors that may affect the rate at which it occurs. Although a considerable body of information is available on both the ultrastructure and mechanisms of calcification, many of the basic features are still not well understood. Several areas of investigation appear to be promising approaches to a level of understanding of calcification and growth that would allow the development of effective measures of control. The formation of calcified exoskeletons of crustaceans and molluscs depends on the transport of calcium across epithelial layers to the actual site of calcification. A number of studies have demonstrated that the enzyme carbonic anhydrase plays an important role in ion movement across the shell-forming tissues and that the hormone ecdysterone increases the rate of calcium transport in crustaceans. Studies of the mechanisms of calcium transport should be combined with investigations of inhibiting agents to determine just how they may affect the basic mechanism of transport. The continuous growth of calcified skeletons in barnacles and molluscs proceeds in small increments or growth cycles rather than as a continuous process. The time required for the formation of a growth increment may be of the order of 1 hour or more than 1 day, depending largely on the organism itself and the environmental conditions. Little is known about the biochemical aspects of incremental growth. It appears, however, that the ratio of crystal formation to organic matrix secretion may change during a single growth increment, resulting in a discrete calcified increment that is evident under the light microscope or the electron microscope. The day-night photoperiod, diurnal changes in the frequency of stimuli to those tissues responsible for calcification, and possibly hormonal mechanisms may influence the formation of these growth increments. A more complete understanding of the physiological and biochemical processes, including the enzymatic and hormonal mechanisms involved, could lead to techniques that would permit partial or complete inhibition of shell development and growth.

Within those animals known as "borers", including several species of shipworms found in temperate waters, many of the same areas in the life history that have been touched upon relative to animals in the fouling community apply equally well and demand a more complete understanding. Most of the boring animals are unique in that rather than being attached to the surface of a substratum or structure, they have successfully evolved mechanisms for drilling into all but the hardest manmade structures and for throughout their adult lives continuing to grow and expand their protected interior habitat with deleterious effects on the structures themselves. As with the fouling organisms, boring organisms are dependent on planktonic larvae for their continued existence (Figures 3 and 4). It has been demonstrated that wood-boring molluscan larvae (i.e., the shipworms) do not survive on wood impregnated with rosewood extractives, obtusaquinone, or obtusastyrene. This appears to be related to an inability to form the calcified structures used in the boring process itself. Obtusaquinone and obtusastyrene are known to inhibit the enzyme phenoloxidase, which is important in the formation of the outer sclerotized, cross-linked, organic layer of molluscan shells, on which calcium carbonate crystals are first deposited. In the absence

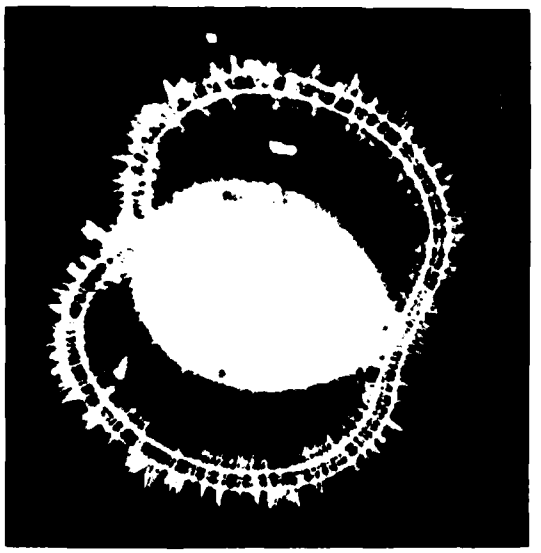


Figure 3—Ventral view of a swimming larva of the wood-boring bivalve Xylophaga atlantica, *showing the wide expanse of the velar lobes, the characteristic notch in the anterior velar margin, and radiating fibrils in the velum tissue (scale bar = 50 um). Courtesy of Dr. Ruth Turner, MCZ, Harvard University.*

of this sclerotized layer, a calcareous shell will not be formed. Additional information is needed on cross-linking of the proteins of the sclerotized outer shell layer and the effects of obtusaquinone, obtusastyrene, and related compounds in interfering with this process.

Although the importance of currents to the dispersal of these larval stages has been recognized for almost a century, remarkably few investigations have been carried out to show specifically how coastal and oceanic currents may affect the dispersal of larvae over broad areas of the world's oceans. In part this is because the circulation in the coastal regions and estuaries, where man has concentrated his most intensive development, is very complex. Investigations have demonstrated that tropical larvae can be successfully transported over long distances and, further, that the completion of morphological development, from hatching to metamorphosis and settlement, may require periods of time as long as 6 months. Even in temperate and cold water forms, the period of larval development may extend for two to six weeks, a sufficient period of time for the larvae to be transported over considerable distances and introduced into new areas. An extremely important question, as yet unanswered, concerns the probability that larvae will remain within the influence of coastal currents rather than being swept offshore where there is little likelihood of survival or of a surface suitable for settling. In recent years we have become more conscious of the complexities of offshore currents, especially as they apply to such areas as the southeastern coast of the United States. The recent discovery of eddies, variable currents in midoceans, counter-currents along the coasts, and the particular type of eddy described as a "ring" emphasizes an extremely important aspect of physical oceanography, which relates directly to our complete understanding of distribution of planktonic forms, including those larval stages associated with both fouling communities and boring animals.



Figure 4—Side view of a crawling pediveliger larva of the wood-boring bivalve Xylophaga atlantica, *showing the ciliation on the foot (scale bar = 50 um). Courtesy of Dr. Ruth Turner, MCZ, Harvard University.*

Although it is recognized that larval dispersal through ocean and coastal currents introduces those larvae that survive into new geographical regions, a number of fundamental questions remain concerning the taxonomic differences between geographically separated populations of adults, not only in terms of morphology but also in terms of physiological, biochemical, and immunological characteristics. Differences between spatially separated populations of marine organisms, described thus far, have frequently been so slight that it has been *im*possible to detect significant differences by the conventional and traditional taxonomic methods. Electrophoretic procedures are now routinely used to differentiate species and as tools in studying genetic complexity at the intraspecific level. In the marine habitat, conditions for genetic differentiation of populations are most favorable in intertidal regions, similar to those frequently occupied by representatives of the fouling and boring communities, as well as in other marginal areas such as brackish waters and salt lagoons. It is within these marginal areas that the relationship between populations and the environment can be best expressed at a microgeographical level. Here one may begin to consider the feasibility of attempting partial control of physical and chemical conditions of the animal populations themselves through a better understanding of the genetic mechanisms involved. The questions of speciation in marine animals, the extent of genetic variation and polymorphism within morphological species, and the possibility of interfering with genetic lines of species as they apply to a number of marine populations and species need to be further studied. An

understanding of the evolutionary processes that control speciation, not only in the coastal areas, where geographically isolated habitats exist, but also in the deep ocean itself, where there are suggestions that the vertical structure of the water mass may serve as a potential isolating mechanism, will require a thorough, detailed, long-term basic study involving collaboration between scientists representing a number of traditional disciplines.

Within the last three decades we have extended our understanding of coastal and estuarine environments, but we have only begun to extend investigations into the fauna and environment of the deeper portions of the oceans. Virtually all of the unanswered questions relative to fouling communities in shallow waters remain unexplored in the deeper waters. The study of the performance of materials in the deep sea is relatively new, but as with research in the shallow waters, it has all too frequently been largely empirical. It has been shown that microbial activity within the deep sea is virtually nonexistent and that many infaunal molluscs in the deeper portions of the oceans have a very slow growth and reproduction rate. Exceptions, however, exist, and some wood-boring bivalves (Figure 5) have been found to grow and reproduce at such a rate that wood 1 in. thick may be completely destroyed after a submergence of only 3 months. Studies conducted at 3600 m have demonstrated that the rate of attack by wood-boring organisms increases with the continued presence of the wood and that the settled borers demonstrate a much more rapid increase in size than had been expected for "normal" deep-sea species. From a biological point of view, it would appear that wood plays a far more prominent role in nutrition in the dep sea than had been expected and that the development of food chains based on wood could be of considerable interest and importance. Continuous studies, involving positive identification of exact locations of experimental sites and provisions for long-term studies at these sites, will be of extreme importance to a further understanding of a number of biological and biochemical problems in the deep oceans. It will be only through concerted efforts and a considerable expansion of the technology and instrumentation necessary for work in the deeper oceans that we can come to an even partial understanding of the reproduction, larval development, physiology, and growth of those animals that are found at great depth.



Figure 5—Metamorphosed shell of the wood-boring bivalve Lyrodus pedicellatus *(family Teredinidae) showing the first three rows of imbrications, or boring teeth. Note the clear, concentric growth rings on the larval shell growth (dissoconch shell,... ch lacks clear growth lines (scale bar = 50 um) Courtesy of D. Ruth Turner, MCZ, Harvard University.*

An ultimate description and understanding of the exact mechanisms involved in biodeterioration will depend on a more complete and interdisciplinary understanding of the biochemical or molecular level of the numerous processes in question. The capacity of marine organisms to grow, reproduce, develop, and survive will ultimately require an understanding at the molecular level. Even the ability to successfully settle on a variety of submerged surfaces, the utilization of a variety of nutrients available within the various depths of the water column, the capacity to adapt to the tremendous variety of fluctuating factors found in the coastal and oceanic depths (changes in temperature, pressure, salinity, available oxygen, CO_2, light, dissolved minerals, etc.) depend further on the organization of the organism at the biochemical level. In the realm of genetics, the capacity of many marine fouling and boring organisms to adapt to "novel substances" placed in the ocean (i.e., plastics, certain metals, petroleum products, and synthetics) will require an understanding of physiological and biochemical properties at the basic level of organization. In many

cases an understanding of the basic biochemical process could provide a "common denominator," which could be applied to a broader understanding and resolution of problems involving a number of unrelated organisms. For example, if it can be shown that there is a common biochemical basis for the thread that connects various fouling organisms (plants as well as animals) to the substrate it would be possible to develop a broad and logically sound approach to all antifouling organisms. A common biochemical pathway leading to the production of the adhesives that weld the organisms to the new substrate might then effectively be blocked by a simple compound that either inhibits the synthesis of the adhesive at a particular point in the biochemical pathway or, through the use of a different chemical that mimics the hormone occurring naturally in the organisms, prevents the activation of the process necessary for release of the adhesive.

It is apparent, from the brief discussion of the many areas of biodeterioration in which our understanding is only partially complete, that a concerted, multidisciplinary, long-term program could contribute greatly to an understanding of the various complicated mechanisms and their regulation, with the final goal of reducing or eliminating those animals which are responsible for the deterioration of manmade structures in the marine environment.

TECHNOLOGY



Fred N. Spiess is a Professor of Oceanography and Associate Director of the Scripps Institution of Oceanography, and Director of the Marine Physical Laboratory of the University of California, San Diego. Dr. Spiess served in 1974–1975 at the Office of Naval Research, London, as Scientific Liaison Officer in oceanography and acoustics for Western Europe. His present research interests include underwater sound, ocean technology, and marine geophysics. He leads about two seagoing expeditions each year. Dr. Spiess received an A.B. from the University of California, Berkeley, an M.S. from Harvard, and a Ph.D. in Physics from the University of California, Berkeley. He is a member of the American Physical Society, the American Geophysical Union, the American Association for the Advancement of Science, the American Association of University Professors, and the Marine Technology Society. He is a Fellow of the Acoustical Society of America. He has received the Wetherill Medal of the Franklin Institute, the Marine Technology Society Distinguished Achievement Award, and the U.S. Navy Conrad Medal for administration of research.

UNCONVENTIONAL VEHICLES FOR OCEAN RESEARCH

F. N. Spiess

University of California, San Diego
Marine Physical Laboratory
Scripps Institution of Oceanography
San Diego, Calif.

Ocean science and technology advance in part through innovations that open up new ways of working or making observations at sea. In many instances this has meant the development of new vehicles to support people and instruments in or on the ocean. While there have been many supporters of such developments, the Office of Naval Research has played a major role throughout its existence. In particular in the early 1960s there was a very fruitful rush forward. The story of half a dozen concepts, first brought to reality in that period, is the subject of this paper. Its concern is not only with origins, however, but with the impact these craft are having today, as well as the forms and uses their descendants may have tomorrow.

A look at this hardware-oriented side of an organization perhaps most widely known for its contributions to basic knowledge emphasizes an essential role of ONR, that of contributing to improved operational capabilities for the Navy. Such a contribution of course arises from the assimilation of basic understanding by those who devise new operational systems and those who use them at sea. Such diffusion of knowledge and concepts must often be encouraged explicitly, however, and one among many ways of doing this is to give the research sponsor the administrative responsibility for some developmental programs. The converse approach, in which primarily development-oriented organizations carry out some research, is also helpful. Looking back, it appears very often that the major innovations really occur at this diffuse interface between science and engineering.

ONR has for many years had responsibility for a modest amount of work funded in the Navy's exploratory and advanced development programs. This activity has interacted strongly with development projects administered by the System Commands (and before them the Materiel Bureaus) to the benefit of both the operating Navy and U.S. ocean science. Most of the vehicle innovations discussed below have supported research programs that have contributed on both sides of the ledger. Overall, in fact, the ONR undersea exploratory development programs have made continual operationally significant contributions, from the explosive echo ranging work of Vine and Hersey at Woods Hole Oceanographic Institution (WHOI) in the early 1950s and the ambient noise and propagation coherence studies of Frosch and Berman at Hudson Lab, through the digital multibeam signal processing innovations of Anderson and the sonar bearing-fluctuation studies of Fisher at Marine Physical Laboratory (MPL), to the most recent contribution to undersea surveillance system design from the Long Range Acoustic Propagation Project.

One hopes that this borderline activity between research and development and between environment and system viewpoints can somehow survive in an era of increasing compartmentalization of functions and areas of interest.

The vehicle innovations to be discussed below all arose from various visualizations of research needs. In some there was a very specific initial requirement, but in all there was an awareness of a multiplicity of possible uses. None originated from programs to design vehicles per se. In every case the designers knew that future experiments are best left to future decisionmaking. They proceeded to bring to reality the simplest systems to do the jobs that were initially important and left substantial growth potential to accommodate needs not then clearly perceived.

Three of the six examples are surface craft (FLIP, ORB, Monster Buoy)—none self-propelled, one unmanned. The other three operate submerged, one (ALVIN) being free and manned, the other two (RUM and Deep Tow) being cable-connected to the tending ship and operating unmanned on the sea floor and in the water column respectively.

All of the vehicles included seem to have a future, since most have been continuing to develop over more than 10 years and there are still new jobs appearing. Some, particularly FLIP, suggest other forms that would be applicable to yet unsolved problems.

Succeeding sections will review the origins, general concepts and some of the scientific or engineering applications of each vehicle, and the final sections will discuss future possibilities. Since some of these craft have generated substantial numbers of publications, the insertion of references in the text would be a complicated matter. Instead there is a minimum bibliography in the final section, separated by topic. In making this compilation emphasis has been given to recent or review papers that will lead the reader back into earlier literature.

RUM and ORB

The first of these six vehicles to materialize was also the one with the least precursor activity or precedent. Historically it has also, because of constraints which should be overcome in the future, been the least productive of scientific output. This is probably in part because its initial conception was tied to a particular piece of sea floor work rather than an observational program.

RUM (Remote Underwater Manipulator) originated in 1958 in Project Artemis. The Project was developed to investigate the possibilities of very long range active acoustic detection of submerged submarines; essentially, this meant the establishment of fundamental information on which to build an undersea radar that could cover a million square miles of ocean with a single station. The program was administered by ONR, directed by R. A. Frosch at Columbia University's Hudson Lab, and involved a large fraction of the underwater acoustics community.

A major problem was the installation of a large number of hydrophones on the sea floor at intermediate depth. A number of approaches were suggested. Among them was a proposal by V. C. Anderson (who was primarily involved in directing the signal processing portion of the program) that one should put the elements down in approximately the proper spots and then drive out from the beach an electrically powered, remote controlled sea floor tractor that would put the units, one by one, into their correct locations.

He was supported to build a trial unit. A surplus Marine Corps tracked vehicle (Ontos) was obtained and gutted to leave only the body and the treads. Two d.c. motors were installed, one to power each of the two tracks. The hull of the vehicle was made watertight and filled with oil to isolate electrical components from seawater. The primary electrical power supplied was about 30 kW at 4800 V, with appropriate transformers and rectifiers to turn it into 0 to 24 V d.c.

Since it would be about 5 mi (8 km) from the point of entry into the water to the work site, it was decided to build a wire storage reel which would ride on the vehicle to carry the necessary 9 km of connecting cable. This led to a substantial total load—60,000 N in-water weight—for the modest-sized vehicle (total track area 2.6 m^2). To move the hydrophones, a manipulator arm and hand were purchased and substantially modified to operate in the saltwater environment.

The length of the cable—used simultaneously for primary power, vehicle control, and data

telemetry—also presented some substantial challenges. For example, it was essential that some sort of viewing capability be provided, in spite of the narrow frequency band the cable could pass. The result was a slow-scan TV system, utilizing a bandwidth of 1 MHz. Two cameras were provided, with illumination from a pair of mercury vapor lamps. A short-range (100 m) high-frequency scanning sonar was built and installed to provide an image of the scene at ranges beyond TV.

As this development was proceeding, other hydrophone placement techniques were also investigated, and eventually an approach was devised by which the installing ship could do the complete job unassisted. RUM was thus brought to an initial checkout point, at which time it appeared as in Figure 1. Anderson returned to concern himself solely with Artemis signal processing, and the vehicle reposed quietly in the Marine Physical Laboratory storage area.

About 1965, with Project Artemis a thing of the past, Anderson began to turn his attention again to remote-controlled vehicles, and ONR assisted in RUM's rebirth. This time the goal was to use the tractor for sea floor work in a more general sense. It was clear in this context that the concept of crawling all the way from the ocean's edge to the work site was too restrictive. What was needed was a tending craft that could carry the wire and lower the vehicle to the sea floor. This would also provide means for righting the tractor in the event it might roll over; the tender could simply reel in the cable to do the job. In fact, it was felt that with a properly designed constant-tension winch, it would be feasible to carry some of the vehicle's weight on the suspension line and thus allow it to operate on much softer sea floor than in its previous incarnation.

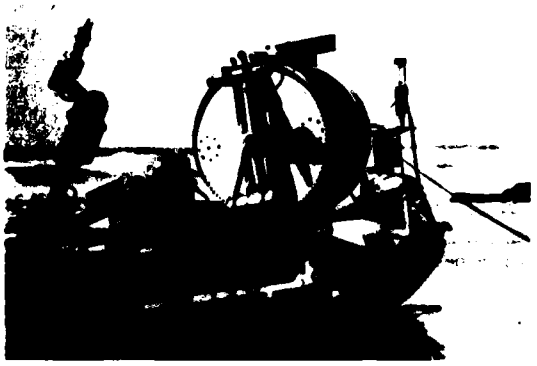


Figure 1—RUM (Remote Underwater Manipulator) as configured in its form for use in Project Artemis

It was thus that the second of the six special craft—the closest to conventional technology and experience—came into being as tender for the most innovative vehicle. ORB (for Ocean Research Buoy) was originally built as a 45-ft-square (14 m) barge, almost completely taken up by about a 20-ft-square (6 m) central well and a large winch with a pneumatic accumulator system, to maintain constant wire tension (± 10% with the vehicle on the sea floor). Doors are provided in the well; RUM rests on these while in transit to the work area. It is then lifted up by its support cable, being held laterally by hydraulic snubbers. The doors are opened, the vehicle is lowered through the well, the snubbers are removed, and RUM is ready for its trip to the sea floor.

Although RUM itself is primarily an oil-filled vehicle and thus has no real depth limitation, its television cameras are in pressure cases limited to 2500 m, and only 10000 ft (3000 m) of support wire can be handled on the winch drum. Operations have thus been limited to the continental borderland off Southern California, where there are many useful work sites at depths of 2000 m and less.

In addition to developmental and training operations, including multipoint mooring of ORB in these depths, RUM has supported two major research programs. The first was concerned with direct in-situ measurement of the mechanical properties of sea floor sediments, and the second, in collaboration with R. Hessler of Scripps Institution of Oceanography, was a study of biological phenomena at the water/sediment interface and in the mud.

RUM has also been useful in its primary role as a sea floor work vehicle, installing sea floor hydrophones, recovering cables, and the like. In this she acts rather like a small manned submersible, but with the ability of exerting a greater pull on items in the mud, since it can take advantage of its weight and the reaction of the sea floor. Because it is unmanned, the safety restrictions are minimal, and this aspect sometimes makes RUM preferable to a manned craft. For example, the U.S.

Navy Civil Engineering Laboratory placed some hollow concrete spheres on the sea bed for test. Small manned submersibles were willing (and able) to attach the retrieving lines to those that were well above the concrete sphere collapse depth. They were not willing however, to operate close to those near their failure limit, since an implosion might endanger the submersibles themselves. RUM was thus successful for that phase of the recovery operations.

At this point RUM is in standby status without a current program. It may not go to sea again, but it has produced both useful research output and a background of engineering and operational experience on which we can build more versatile sea floor vehicles. Options for the future will be discussed below in the section on future sea floor vehicles.

ORB, on the other hand, continues to be used fruitfully. Its well, its winch, and its low on-station cost (relative to conventional ships) make it a useful support craft for other operations. With a 12-ft (3.6 m) section added to each end to make a water plane 45 × 69 ft (14 × 21 m), she is now as shown in Figure 2. With its shallow draft and large metacentric height it has an interesting type of stability, essentially, following the slope of the waves without its own natural roll period being excited. Where a conventional ship of much larger dimensions would roll or pitch heavily in intermediate seas, ORB rarely rolls more than the actual face angles of the waves. As a result one has a suspension point for submerged loads which moves primarily with the heaving motion of the sea surface, plus a capability of maintaining good compensation with the constant-tension winch for heavy loads. Its chief disadvantage is its low towing speed, 3-4 knots (1.5-2 m/s), forced by its small dimensions and boxlike form.



Figure 2—ORB (Ocean Research Barge) under tow for operations in 1972

The most exciting program for ORB is yet to come, with the advent of the ADA (Advanced Detection Array) vehicle, which ORB will tend. The vehicle is a submersible structure having a deck about 9 × 21 m and a thickness of 3 m. On the deck, covered by a water-filled rubber dome, will be an array of 720 hydrophones for use in studies of sound propagation and ambient noise. The plan is that it will be towed, in tandem with ORB, to the work area. There it will be flooded, retaining slight positive buoyancy and flipping onto its side, so that the plane of the hydrophone array will be vertical. It will then be attached at a predetermined point to one of the lines of ORB's deep-water three-point mooring and pulled under to the desired operating depth by adjusting the lengths and tensions of the mooring lines.

At this time it appears that ORB has a number of years of work ahead of her in a variety of underwater acoustic research programs.

ALVIN

The research submersible ALVIN is the first of the modern, deep-operating American submersibles to have real research impact. ONR's initial venture into research submersible activities came with its charter of Piccard's bathyscaphe *Trieste* for operations in the Mediterranean Sea in 1957. In the same period, Edward Wenk, who had led submarine pressure hull development work in the Navy, moled to Southwest Research Institute and began to work, with backing from Reynolds Aluminum Co., on the design of a 4000-m-operating-depth submarine, *Aluminaut*. The Navy was quite interested in using this ship for research, although its configuration limited its near-sea-floor observational capabilities. ONR thus budleted exploratory development funds in hope of arranging to charter this new craft through Woods Hole Oceanographic Institution. This did not materialize, and the funds were shifted to allow Woods Hole to oversee the design and construction of a new submarine from the keel up. ALVIN (in honor of Allyn Vine, one of the major proponents of new craft of many types) was the result. As with others in this list, this craft has grown in capabilities throughout the years since

its initial oprations in 1964, although its outward appearance is still substantially as in Figure 3.

As originally built it could operate to nearly 2000 m. Since that time a new titanium pressure hull has been substituted for the original steel one, and with appropriate alterations of essential fittings its working limit has been increased to a very useful 4000 m. It carries two men and has, over the years, built up an associated acoustic transponder navigation capability as well as equipment for photography and bottom sampling. Under development is an auxiliary rock drill, which will improve the sampling situation, particularly for obtaining fresh rock for magnetic and chemical analysis and for emplacement of measuring instruments on the rocky sea floor at rise or ridge crests.

The principal limitations on the use of manned submersibles arise from the difficulties of handling them at the sea surface near the work site. The launch and retrieval operations are quite weather sensitive, and since vehicle endurance is normally limited to less than one day one must be very conservative about predicting the weather at retrieval time, before launching. In fact, the only severe accident in the many years of ALVIN operations occurred during the launch process.

Woods Hole has built a special-purpose craft, LULU, as the support ship for ALVIN operations. While this has been satisfactory in the past, LULU is slow (6 knots, 3 m/s maximum) and is herself somewhat weather dependent. As recognition of problems ALVIN could help solve grows, it becomes increasingly clear that it must be operable from large, but more or less conventional research ships such as *Melville* and *Knorr* (AGOR 14 and 15). This would make it much more feasible to think of the submersible as part of the equipment for learning about the sea and to use it in close conjunction with other tools, with consequent reduction of overhead costs.

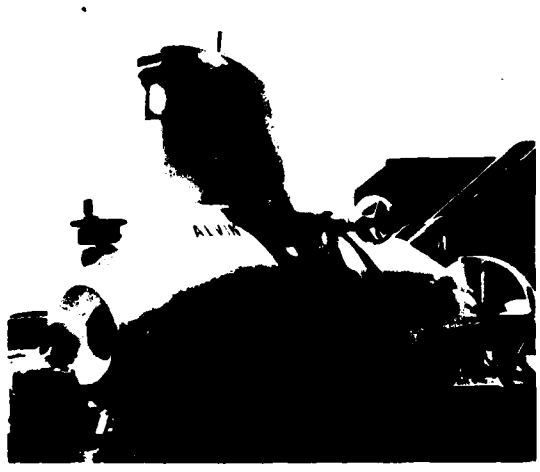


Figure 3—Research submersible ALVIN, operated by Woods Hole Oceanographic Institution

Initial design work is under way at WHOI on adapting handling equipment developed for commercial submersible work in the North Sea to provide this capability. Actual construction of the necessary equipment will help bringing about more effective submersible utilization.

ALVIN has played a major role in a number of sea floor study programs. In these her major contribution is in being able to go to a place that previous, less direct observations have shown to be particularly complicated to understand, and by persistent, direct observation to pick up clues that would be difficult to find or interpret with less flexible unmanned survey equipment. The best publicized of these operations was the dive sequence at the crestal valley of the Mid-Atlantic Ridge as part of the 1973–1974 French-American cooperative program (FAMOUS). Similar operations have been carried out in the Cayman Trough, and a further, geochemically oriented one is planned, with more sophisticated water and sea floor sampling systems, for the crest of the Galapagos Spreading Center (near latitude 0° 1°N and longitude 85°W) in 1977.

Looking to the future, the major new potential that may come into play is the ability to place and monitor complex local measuring systems on the sea floor. These could shed considerable light on very-near-sea-floor water circulation and on the motions of the sea floor itself.

While very little of ALVIN's research activity has yet had any direct impact on the Navy's operational capabilities, ALVIN has, as a hardware development, opened up new pathways. Following its initial successful operations, the Navy (which in fact owns ALVIN, as it does all the other vehicles discussed in this paper) built two more units, *Sea Cliff* and *Turtle*, very similar to the original prototype. These have been operating

for several years as units of Submarine Development Group I, based in San Diego. They have supported some scientific work but have also contributed in a more applied manner by retrieving key parts of downed aircraft and weapon components, inspecting and repairing or replacing elements in the Navy's underwater tracking ranges, and assisting in other ocean engineering activities. Their success and acceptance as useful elements of the naval force is indicated by the fact that both are being modified to provide greater operating depth. *Turtle* is planned to have a 3000-m capability by 1978, and *Sea Cliff* should be able to go to 6000 m by 1980.

FLIP

The most obviously strange craft of the group, to the layman's eye, is probably the tiltable manned spar buoy, FLIP. From an oceanographer's viewpoint, however, the concept of using a vertically floating spar as a stable platform is very old. Such devices were used as wave gages many years before FLIP was designed, and manned spar buoy laboratories were recommended as part of the marine science fleet in the National Academy of Sciences Committee on Oceanography reports of 1959. In this context FLIP represented three major innovations. First and most important, she was built, while all other manned spar buoys before her existed only on paper or in model form. Second, she introduced the capability of carrying out a 90° change of attitude as a routine maneuver, going from horizontal to vertical and back again in a matter of minutes. Third, she introduced the concept of shaping the underwater profile of the spar in such a manner as to give not only a lower frequency (longer period) resonance than a simple spar, but in fact to reduce the driving force of the waves, by a cancellation process.

FLIP's origin, was in the Navy's SUBROC program, as was that of the Deep Tow (discussed below) and several other interesting major equipment developments (e.g., the University of Washington Applied Physics Lab's remote-controlled deep ocean probe) carried forward outside of ONR. This proposed a submarine-launched antisubmarine missile, to be fired at submerged targets at ranges substantially greater than those for conventional torpedo attack. The firing was to be controlled solely from underwater acoustic information, and thus it became necessary for the first time to be able to put the closest possible limits on the azimuth errors that might be introduced into the sound path by lateral refraction.

A research program was thus launched to determine what the magnitude of such propagation direction fluctuations might be. The author and F. H. Fisher of the Marine Physical Laboratory of Scripps Institution of Oceanography, after considerable investigation of the problem, proposed the construction of a manned spar buoy as an ideal platform for such measurements. Hydrophones could be mounted at reasonable depths on the rigid structure, from the upper part of which optical or microwave direction measurements to the source could be made for comparison.

The result was the opportunity to design and build FLIP. This long (100 m) slim (6 m) craft is towed, at speeds of 7 to 8 knots (3.5 to 4 m/sec) out to its work area. Onsite the ballast tanks are flooded, and it swings to the vertical attitude, with 90 m draft. In the vertical position its heave resonance period is about 27 s, an it is shaped to have a null response at about 18-s wave period. The result is that over most of the normal ocean wave spectrum, the ship's heave response is about 5% of the wave amplitude.

The craft was kept as simple as possible. It was clear that she could carry out a variety of other tasks, but we felt that an evolutionary approach to them would be better than trying to anticipate other requirements in a detailed way. A comparison of the configuration of FLIP's upper portion as she flipped during trials in 1962 (Fig. 4a) with her appearance during the OWEX operations 10 years later (Fig. 4b) gives some feeling as to her growth in capability.

She can carry 16 people, including 4 to 6 crew. Although she has operated for 30 days at a time with full 16 on board, a more comfortable loading is 10 or 11. For operations in excess of 3 weeks she is usually resupplied at sea. It would be quite feasible for her to stay at sea for 1 or 2 years at a time. The principal problem for long operations is personnel transfer. At present this is done by small boat and requires some degree of agility and

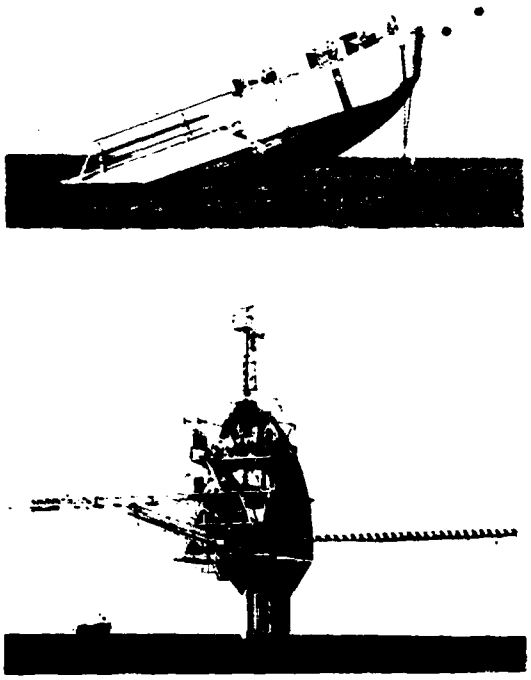


Figure 4—FLIP (a) in the process of making the transition from horizontal to vertical, 1962; (b) in the vertical, as configured for Ocean Wave Experiment (OWEX) of 1972

awareness of the sea on the part of the person being transfered, since FLIP scarcely responds to the waves while the small boat follows them vigorously. The result is substantial relative motion, occasionally prohibiting transfer altogether. Studies of the possibility of using helicopters or other means are being initiated.

The two principal fields in which FLIP has been used are underwater acoustics and physical oceanography/meteorology. Initially, most of the acoustic work was in the shallow part of the water column, using receivers fixed to the hull. As the ship's postulated stability was verified it became clear that it would be fruitful to suspend the listening elements far below the ship, since they would not be subject to appreciable motion of the suspension point with resulting generation of interfering noise. Until that time, all operations had been carried out in a freely drifting mode. This was no longer satisfactory, since as FLIP drifted with the wind and surface currents, the deep hydrophones were towed (with noise production) through the much more nearly stationary deep water.

At this stage (1969), we mastered the art of making three-point moors in 5000-m depths, and since then we have carried out this operation over 30 times. With this capability, it became quite useful to deploy multielement vertical arrays of receivers that would allow us to gather sound-propagation and ambient-noise data over the entire water column.

The physical oceanographic and meteorological work began early (1963), when FLIP was used by W. H. Munk as the mid-North Pacific observation point in his study of trans-Pacific propagation of surface waves generated by southern ocean storms. Attention then shifted to the internal motions of the water, particularly internal waves. At the same period, scientists interested in working on the problems of wave generation and transfer of momentum, heat, etc., between air and sea began to use FLIP as a convenient, nearly stable platform for making microscale observations in the open ocean close to the air-sea boundary. In this context, FLIP supported investigations during BOMEX (Barbadoes Oceanographic Meteorological Experiment, 1969) in the Atlantic and OWEX (Ocean Wave Experiment, 1972) in the Pacific. The former emphasized near-surface micrometeorology, and the latter, radar backscatter effects.

At present the principal programs are in sound propagation and internal motions of the water, particularly as viewed acoustically. One particularly exciting new facet is the possibility of observing the horizontal motions of individual elements of the water out to as much as 1 km away from FLIP with spatial resolution cells of 5 to 10 m and velocity precision of in the millimeter-per-second range. This is in major part feasible because of the ability to hold a Doppler-measuring, horizontally viewing sonar on a stable platform. It will make it possible for the first time to observe internal waves in the mixed layer and to have much better resolution for directional studies than in the past.

In the first few years after FLIP's completion several other manned spar buoys were built, including SPAR of the U.S. Naval Ordnance Lab

(also for SUBROC-related work), Cousteau's Isle Mysterieuse and its CNEXO successor Bouée Laboratoire, and POP of General Motors. Of these, only the Bouée Laboratoire, moored in the western Mediterranean, is currently in use.

Looking farther into the future one can visualize other craft utilizing the flipping and spar buoy concepts, as well as interactions between FLIP-based observations and satellite programs. These will all be addressed in sections below.

MONSTER BUOY

This is the only vehicle development discussed here which originated ONR's research (as opposed to exploratory development) program. There had been a long history of development of small to intermediate-size, deep-ocean moored buoys, used by the oceanographic community as position reference points and to record near-surface phenomena (wind, temperature, nuclear bomb test radioactive fallout, etc). Major innovators in this field were Richardson (then at Woods Hole) and Isaacs at Scripps. While these had, in some configurations (e.g., Isaacs' Bumblebee, Figure 5), remarkable survivability, they were all internally recording. ONR staff personnel in 1959 felt a need for units that could not only survive but could mount a radio transmission system to transmit data in nearly real time to shore from midocean.

A program of engineering development was started at General Dynamics/Convair in San Diego; it resulted in a design for a large discus-shaped buoy (Figure 6) containing a diesel engine for primary power and providing a mast for a radio antenna, a high air intake, and instrument mountings. Principal dimensions of the units were 12 m in diameter, 2.5 m thick, with a mast height of 10 m. This size dictated that the units be towed to the work site, where they would be moored.

Initially only a very simple instrument suite was mounted—air and sea temperature, atmospheric pressure, wind, humidity, rainfall, and solar radiation. The telemetry system (operating in the HF band), however, had considerably greater capacity. Data from the sensors were stored in the buoy until an interrogation was received from the control station or a preset scheduled transmission time was reached, and then the information was transmitted to shore.



Figure 5—Bumblebee ocean measurement and data recording buoy

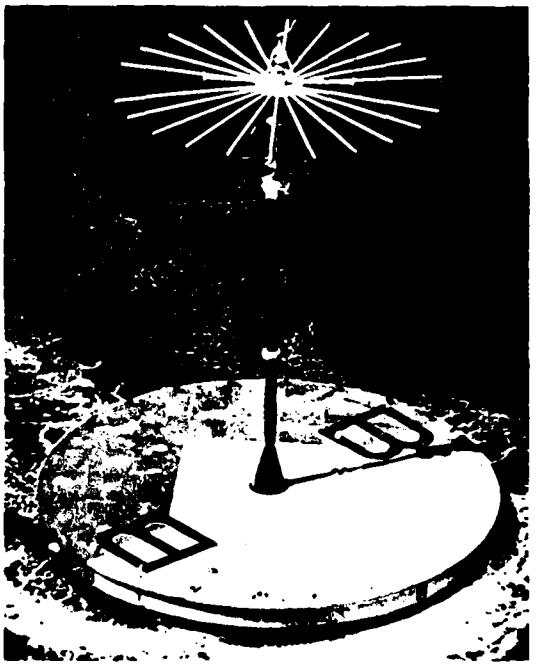


Figure 6—Monster buoy in its earliest form.

The program basically achieved its goals in the hardware sense by 1968, and two of the buoys became the major observational platforms (augmented by a number of Isaacs' Bumblebees) of the North Pacific Buoy Project, a short-lived precursor of the present, broader NORPAX air-sea interaction program.

At about this time the National Oceanic and Atmospheric Administration (NOAA) began to be interested in the air-sea interaction problem and established a data buoy section (now at the NASA test facility in Bay St. Louis, Mississippi), which immediately began buoy development and took over the Monster Buoy design as one of its major vehicle types. This organization now has several such buoys, operating essentially as weather ships, in the Pacific.

The design has also shown itself to be useful in more applied contexts. For example, the British have replaced many of their lightships with buoys of essentially the monster-buoy configuration, although moored in shallow water and not requiring the radio telemetry link.

It thus appears in the long run that this development has had its most direct impact outside the Navy sphere, although the data the buoys currently provide gives additional backing to a variety of Navy, as well as civilian, meteorological and oceanographic prediction activities. The possibility of direct interaction between such vehicles and the NASA Sea Sat observational program will be discussed further below.

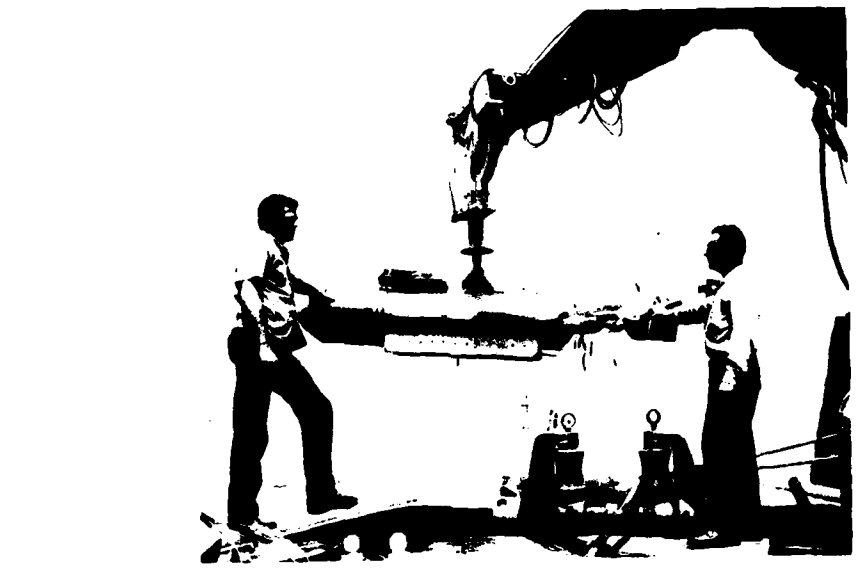
DEEP TOW

Probably the most successful vehicle in this group in terms of scientific output, the Deep Tow system is perhaps the least clearly a vehicle, as differentiated from a research instrumentation system. Its initial funding came, as in the case of FLIP, from the SUBROC program. Again, the direction of propagation of sound was the problem, but in this case it was the interaction with the sea floor that introduced the error. If sound bounces on a sloping surface the resulting reflected ray will lie in a differential vertical plane from the incident ray, with the result that a bearing error is introduced, depending primarily on the slope of the sea floor across the line of sight.

As this problem was examined it became clear that the scale on which the slope should be measured involved averaging over lateral distances of less than 100 m. Unfortunately, the usual survey echo sounder of those days insonified a patch about half as wide as the water was deep and recorded only the times of the major arrivals (not necessarily from directly below the ship); thus they were capable of making slope measurements only on a lateral scale of kilometers. Two approaches were proposed for gathering the necessary data. Our group at MPL suggested simply towing a precision sounding system close to the sea floor, while Vine of WHOI suggested developing a multiple, vary narrow beam, surface-ship-mounted sounder. Both approaches were followed. The MPL one, being simpler, produced data on this problem in a number of areas before the multibeam sounder (developed eventually under Naval Oceanographic Office cognizance) was in being. The latter, of course, has a much higher rate of coverage. In recent direct comparisons, it does not appear that the ship-mounted system provides good data for bottom slopes up to about 50°, although it occasionally misses small but significant topographic features (e.g., cones 50-100 m diameter, 20-30 m high).

The original deep tow system was quite simple. A pressure-proof case for the electronics, an up- and a down-looking sounder, and a transponder call-and-receive transducer made up the entire unit (Figure 7a). It was towed close to the sea floor with an electromechanical cable (about 9 km long) having a coaxial core over which power and both outgoing and returning signals were transmitted. All data storage and display was done on board the ship. The transponder navigation system had to be designed and built in house since in the early 1960s there were no adequate commercial systems available.

Considerable growth potential was provided, and the first (predictably) added capability was a proton-precession magnetometer to allow near-bottom investigation of the curiously lineated magnetic anomalies, which had been mapped by Mason and Raff and eventually became a cornerstone of the evidence for sea floor spreading and plate tectonics. Following loss of the submarine *Thresher* in 1963, there was a buildup of interest in sea floor search techniques and en-



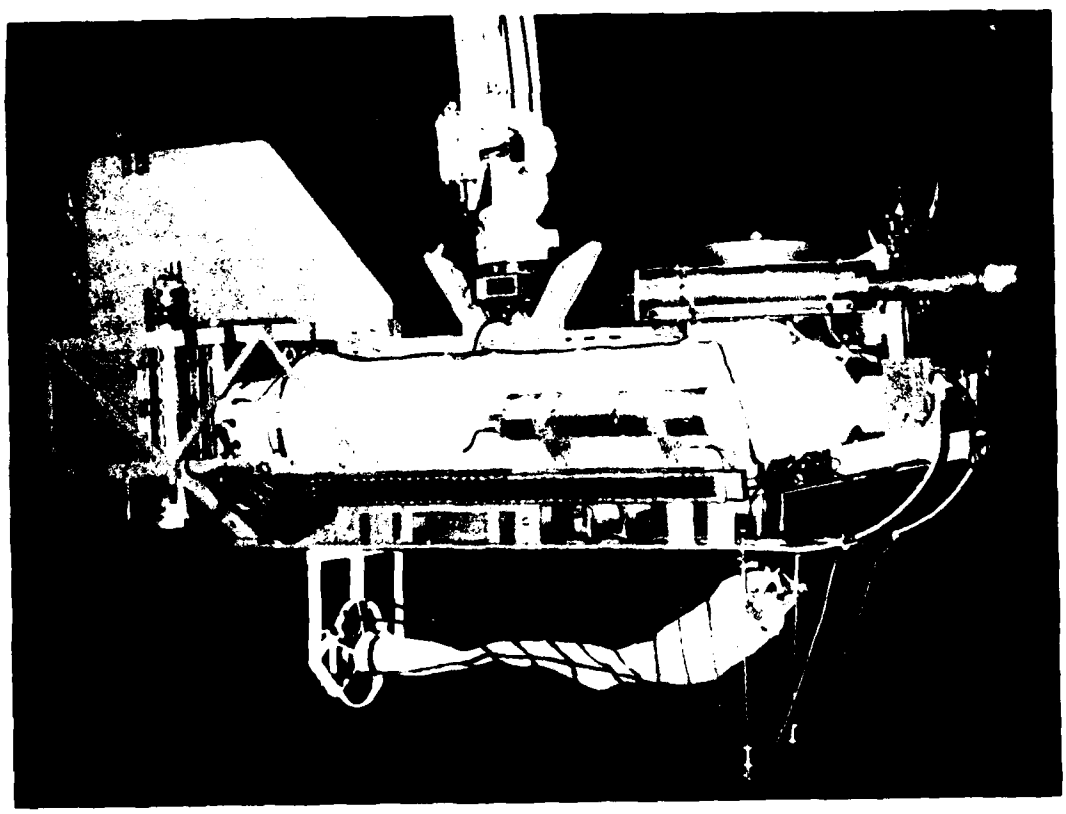


Figure 7—Deep tow sea floor survey unit (a) as operated in 1964 and (b) as currently configured

vironmental limitations. During the late 1960s the system grew substantially and rapidly in its capabilities, adding cameras, side-looking sonar, and a bottom-penetrating 4-kHz echo sounder. While earlier we had operated 100 m off the bottom, with these systems we moved closer; the side-looking sonar and the 4-kHz system operate best about 30-40m up, and the stereo wide-angle photo system requires about 10 m altitude.

Since the period of the most rapid system expansion there has been a continuing, but slower, development of further capabilities (Figure 7b), many of which are not used on every lowering. A number of these represent the inclusion of interest in the water column as well as the sea floor. First to be added was a precision (0.001°C) temperature measuring capability. This was followed by a device, still in the developmental stage, to measure the optical properties of seawater in situ. In 1973-1974 two systems were added to sample suspended or living materials in the water. One of these is a modification of a standard upper-ocean plankton net. This one, however, has three sections and can be controlled from the ship to provide consecutive net tows close to the deep sea floor with continuous monitoring of vehicle locations in both horizontal and vertical coordinates. The second sampling system is of the pumping type, with the water being drawn through a millipore filter. The system is completely external to the fish, being powered by a propeller driven by the vehicle's motion through the water. Starting and stopping the pump is controlled from the ship.

The most recent addition was assembled early this year jointly by one of the SIO geochemical groups (Craig et al.) and ourselves. It provided a salinity/temperature/depth measuring capability and a set of of remotely triggered 10-l water-sampling bottles. The investigation that promoted this was a search for plumes of warm water we suspected were emitted from the broken rocks at some sea floor spreading centers. An expedition, just completed (June, 1976, Lonsdale and Weiss chief scientists), did in fact sample such water at the Galapagos spreading center, and chemical analyses are in progress. It has already been determined, for example, that the sample is very highly enriched in helium three.

A major step forward, still in progress, is a move to record and utilize the returning acoustic amplitude information quantitatively. This started a few years ago with the 4-kHz system and is providing substantial information on sound absorption and the fine-scale variability of surface reflectivity. A similar capability will be in action during the latter part of this year to make quantitative measurements of acoustic backscatter at our side-looking sonar frequency (110 kHz). This will not only provide design information for proper construction of such systems, but it will also allow interpretation of the resulting data in terms of bottom roughness parameters.

The scientific programs in which the system has been involved are a mixture of Navy Deep Submergence Program interest in sea floor search technology, ONR interest in sea floor acoustic properties, and NSF support for geological and geophysical work. Both Navy-sponsored activities result in development of useful equipment and production of operationally relevant information (e.g. sea floor slope statistics as related to sonar performance, variability of bottom reflectivity, etc). A dozen Ph.D. theses (nine at Scripps and three at other institutions) have been based substantially on data from this system.

These programs has been carried out in a variety of sites (Figure 8), chosen to cover the various topics of interest. In the early stages the emphasis was on typical deep sea areas, particularly abyssal hills. Subsequently this shifted to sites primarily

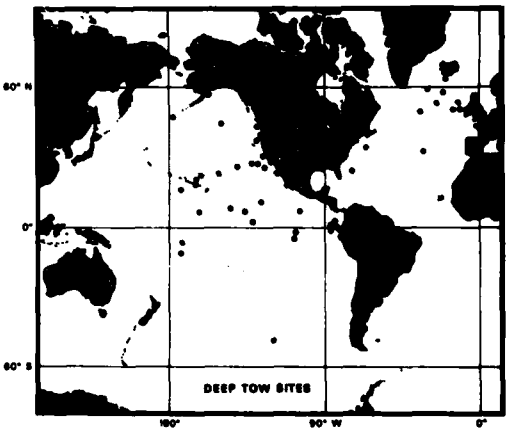


Figure 8—Deep tow survey and research operating sites, from 1964 through 1975

important in relation to plate tectonics (particularly spreading centers) and those of interest relative to deep sea sedimentation and erosion. At present there is a return to less dramatic sites, as we concentrate on sea floor acoustics and manganese nodule programs.

Operationally the most challenging was towing in the Aleutian Trench at a depth of 7 km. This required the use of two winches and a total of 11 km of wire in two sections, coupled together as part of the launching operation. On the other hand, the most directly rewarding was our recovery of one of our vehicles 6 months after it had been dropped due to a broken wire in 3 km of water.

The system has been amazingly useful. Some new facet of sea floor information usually appears on every operation, often on one of the subsystems that one felt would not perhaps be particularly important in that operation. This makes it difficult to simplify the growing complexity of the system; once one aspect has proved its worth, we are unwilling to go into further expeditions without it. As least several more years of fruitful operation seem likely, however, before complete redesign is required. In the meantime other groups (Naval Oceanographic Office, with Teleprobe and WHOI with Angus) now have their own versions, while commercial organizations, the U.S. Geological Survey, and CNEXO are making plans for building theirs. Our eyes are on a new configuration, to be discussed below.

FUTURE CONCEPTS

Given the inherent, unpredictability of research activity, at least on a long-term basis, it seems unlikely that one can predict what new vehicles will emerge as important in the next decade. One can, however, see ways in which some present unconventional craft may contribute, and even visualize forms that do not now exist but follow in some logical progression from the craft discussed above.

One interacting complex of existing and apparently unrelated craft could well form a very powerful whole, particularly in the study of air-sea interaction and the related field of computer modeling of internal ocean dynamics. The driving force is the growing use of satellites for observation of the sea surface. The NASA Sea Sat program will provide the most complex instrument suite for this purpose, providing systems that respond to sea surface conditions—temperature, roughness, foam, etc.—in a variety of ways. With the capability of observing large ocean areas, this opens up the possibility of developing much more comprehensive pictures of physical oceanographic phenomena than ever before. Unfortunately, however, it is not clear that the sensor outputs will yield unambiguous descriptions of conditions at or near the surface. It is even less certain whether the observations can, by themselves, be interpreted in ways that will help us understand what is happening throughout the volume of the ocean.

One can, however, take a pragmatic view of the situation and establish continuously operating stations of varying degrees of sophistication, to measure waves, slicks, air and water temperatures, humidity, air turbulence, and the like, and then consider the satellite data as a means of interpolating among these observation points. While some of the surface vehicles may be in the nature of well-equipped monster buoys, some should surely be manned stations on which more complicated instrument suites can be mounted. FLIP-type platforms would be particularly useful in this context, since they would open the next layer of the problem, providing information as to the manner in which the surface phenomena interact with those in the volume, producing internal waves, mixing, microstructure, wind-driven currents, and the like.

The emphasis here would also be on long observational sequences to match those of the satellites. This would be a new aspect of oceanography, since for most ship-dominated work an observation period of a few weeks at a single site is considered fairly long. Only a very limited number of sequences of observations at sea have been made for longer periods. In the present context, the emphasis is not only on the need for long enough time series to allow valid statistical studies (spectra, etc.), but on the nonstationary phenomena of the sea—the manner in which change takes place. The vehicles involved must be durable and reasonably inexpensive to maintain on station—attributes which the manned spar buoy possesses.

As complementary vehicles to the on-station observation platforms, there must be some means for carrying out resupply and personal rotation. Conventional oceanographic craft, with their 10 to 15-knot (5-7.5 m/s) maximum speeds, would hardly be appropriate. Instead it would be desirable to bring much faster craft into the picture. Surface-effect ships would provide one option; another, slower but better matched for personnel transfer, would be the semisubmersible ship.

Finally, one further class of vehicles, well used in the military context, would be essential. A reasonably fast, properly equipped nuclear submarine would provide a capability of intercepting major storms and making observations from the relative calm that always exists well below the interface. Acoustic equivalents of the satellites' electromagnetic remote-sensing suite could provide most of the information, with limited use of direct measuring units floated up from the submarine. Multichannel Doppler systems (of the type described in the FLIP section above) with up-looking echo sounders of various degrees of resolution should provide a major part of the information.

This interplay among a variety of unconventional craft could lead to a far more comprehensive picture of the constantly changing condition of the ocean than we could ever hope to accumulate by reliance on any one vehicle alone. Beyond this, there will clearly be new vehicles produced. Some possibilities are covered in the next three sections.

Sea Floor Work Vehicle

A number of research and engineering problems seem to dictate that a deep sea version of RUM, including attributes of some of the other cable-connected vehicles such as the Deep Tow, would be of considerable value. One can visualize a number of research problems in which it would be most desirable to be able to emplace instruments on the deep sea floor in well-chosen location or in complex local configurations relative to one another. One example would be installation of ocean bottom instruments for direct measurement of sea floor spreading. Functions involved in this problem would be the selection of solid rock sites, drilling of holes for securing instruments, equipment installation, replacement of power packs, and the like. A second class of experiments are those concerning hydrodynamic effects close to the deep sea floor. Numerous erosional and depositional features whose origins have not been explained have been observed on the sea floor. For example the hundreds of furrows seen near the foot of the continental slope and in other similar locations apparently involve roll vortexes along the sea floor. With an appropriate vehicle, families of instruments could be carefully emplaced to make the long-term observations necessary for describing the interactions of sediment and water that create such bedforms.

On the engineering side there are problems (not unrelated to those above) associated with evaluation of the feasibility of disposing of radioactive waste in the sea bottom. Salvage operations or detailed examination of wrecks to learn their causes and to retrieve essential elements also indicate a need for a vehicle that can make good observations in the water column and then land to move about firmly on the sea floor carrying out the necessary operations. If such operations were well planned it would not be necessary to use a manned vehicle, with its attendant limitations arising from launching problems and short on-bottom working times in deep water.

It appears that it would be a reasonable engineering feat to create a smaller yet more capable version of RUM that could be operated from some of our existing larger research ships, to carry out most of the tasks indicated above.

Large-Area Platform

A number of research problems suggest the need for a sea-surface structure that need not have much bulk but which would have considerable lateral extent. Such a craft would be useful for suspending midwater hydrophone arrays for studies of sound propagation and noise. They could also provide the mounting structure for radio antenna arrays for over-the-horizon radar or astronomical research. If the structure were open enough it could also carry instrumentation for the study of internal waves and upper ocean mixing processes.

These options suggest an array of spar buoys, interconnected to produce a large, open framework. Such assemblages are clearly in some sense feasible and have been studied at modest scale in tanks of various dimensions, usually in the context of supporting midocean aircraft landing strips. An important aspect, relating to the practicability of using craft of this kind, is the question of how one might go about assembling such a thing at sea.

This has been addressed in a program sponsored by ARPA and administered through ONR. Paper studies and small models can show the way in matters of this kind, but true practicality is not really demonstrated until one actually copes realistically with the details of carrying out the operation.

Thus, the culmination of the investigation was the assembly of a three-element, open frame, big enough to provide an approximation to reality. In this case the three vertical elements (a bargelike structure to be discussed below and two spar buoys), each about 13 m long, were assembled at sea into a rigid triangular configuration. The units rode about with about 10 m draft and were connected top and bottom by rigid horizontal members, with diagonal bracing provided by chains. Figure 9 shows the structure in its completed form.

Since our concern was with still larger elements, capable of supporting an airfield, the assembly operation was carried out as a one-eighth-scale model test. The full-scale spars would be as big as FLIP. In this context,then, we used outboard-motor-powered skiffs as tugboats and prerigged lightweight lines, devising a variety of connecting couplings. No divers were used except as observers. All elements involved were brought to the area as deck load of the flippable barge and swung into the vertical as a single unit.

Assembly was accomplished in a matter of hours, in a seaway which scaled to a mean wave height of about 5 m. The spectrum of the sea in this instance, considering the scaling factor, was much more severe than that which one would encounter in the real ocean for the same mean wave height. In this case there was significant swell energy in the same frequency regime as that of the spar buoy heave resonances, which would be at periods longer than 20s at full scale.

The approach used was such that it could be generalized to encompass a much greater number of spars and thus an overall structure of much greater lateral extent. For example, the horizontal members were sized to cope with the much greater bending moments that would be encountered if a larger structure were assembled.

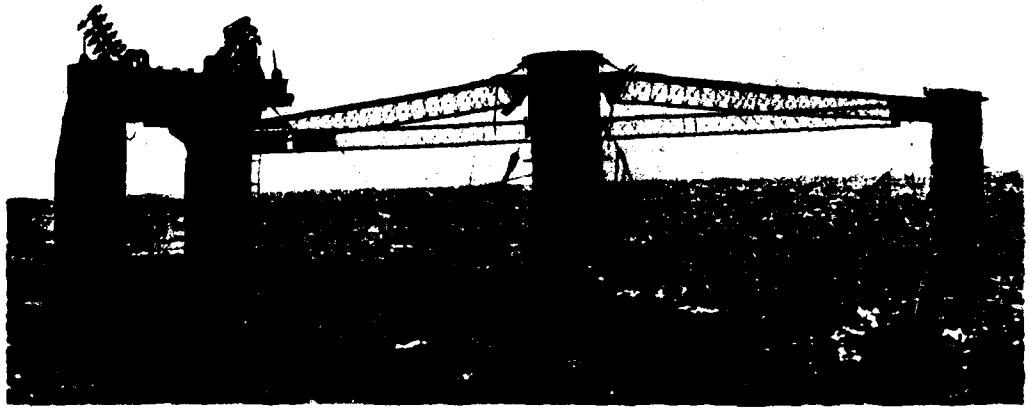


Figure 9—One-eighth scale model of three-element stable large-area platform, as assembled at sea off San Diego

With this operation successfully completed, one could propose to build and assemble a full-scale structure with confidence that it could support whatever program might need it.

Flippable Barge

Many research possibilities require that one move large objects through the air-sea interface and suspend or tend them in the water during the course of an experiment. In the course of the ARPA floating platform investigation mentioned in the previous section we developed a concept that seems to provide a good capabilty for carrying out such operations at modest cost.

A drawing of the flippable barge, in its vertical attitude, is shown in Figure 10. It is an amalgam of the FLIP concept with modern large barge construction concepts. In its gross form it resembles the 100-m-long craft built to carry pipe manufactured in Japan to the North Slope for construction of the Trans-Alaska Pipeline. The FLIP concept enters in two ways. First is the operational aspect of being able to flood or blow ballast to go from the horizontal, towed attitude to the vertical, tending position. Second, the cutout portion gives the proper underwater shape to reduce the waterplane area (with resulting long natural period for heaving motion) and to minimize the driving force of the waves.

A discussion of two potential applications, one in underwater acoustics and the other in support of submersible operations, should provide some insight to the usefulness of such a craft.

Only one major program has been mounted to study the problems and potential of long-range active sonar. Project Artemis, discussed above, gathered during its lifetime, significant but limited information on sound propagation and reverberation. At some time it will be essential to initiate a follow-on program, and in it there will be the problem of handling a very large, heavy acoustic transmitting transducer system. Such a device could most easily be operated from this type of barge.

In this case the transducer, with enough attached buoyant material to leave it with only slight negative buoyancy, would be mounted as deck load in such a way that personnel could work on it with the barge horizontal. It would be in quiet

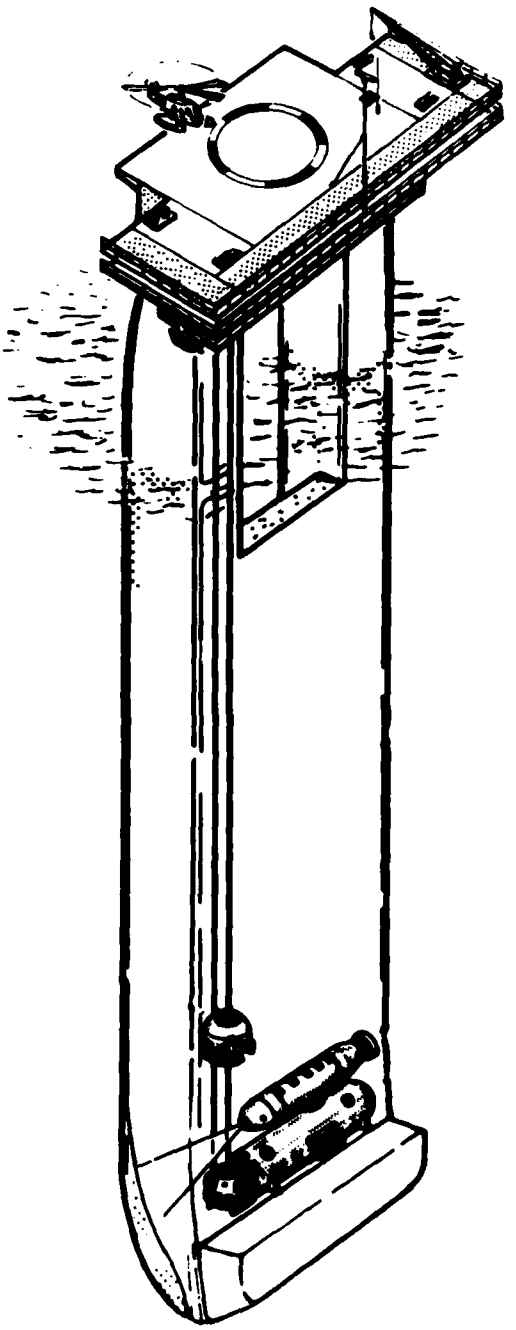


Figure 10—Artist's drawing of flippable barge in vertical. Configuration is arranged for support of small submarines.

water 50 to 100 m below the surface once the craft had flipped to the vertical. In this position the transducer could be lowered to any operating depth from this reasonably stable suspension point. Since the winch and wire would not have to cope with the dynamic loads associated with controlling the massive unit in the surface or near-surface environment, they could be optimized for power transfer (a major engineering constraint in itself). An accumulator adequate to compensate for the limited heaving motion the barge would experience (a few meters in the most extreme sea) would eliminate the need for the suspension system to handle any load other than the small net negative buoyancy.

As a large platform extending into the air and with more than adequate fuel tankage, the barge would be able to handle the prime power requirements and house the personnel and topside equipment needed to carry out the program at sea.

The same craft would provide an ideal tending capability for small or intermediate-sized submersibles. Providing adequate cradles to hold the submarines during the flipping operation would allow launching in considerable higher sea states than is now possible. Once the barge was vertical and the submersible well below the surface it would be straightforward to arrange for dry underwater replenishment, battery charging, and personnel transfer using a variety of configurations. It would thus not be necessary for the submersible itself to return to the surface until completion of the entire sequence of dives.

A wide variety of other tasks that can be visualized would use the full capabilities of a craft of this style. In addition, it could support tasks that might be carried out from a more conventional barge.

CONCLUSION

This account has emphasized unconventional vehicles developed in the framework of ONR's exploratory development program and possible follow-on craft that have their roots in these concepts. Several other classes of vehicles have emerged from other Navy activity. Most notable of the unmanned undersea craft are the tethered, hovering types, such as CURV and RUWS (developed by the Naval Undersea Center and its precursor laboratories), the free vehicles for use at the sea floor (e.g., Isaacs' monster camera and a variety of current meters, seismographs, and the like), and others which hover in midwater (Munk's oscillating temperature measuring system, for example). Major surface craft types such as the hydrofoil, surface-effects craft, and semisubmersibles have been developed in the more ship- and hydrodynamics-oriented community.

The most impressive point is that all of these are basically Navy programs. Essentially no innovative instrument deployment or vehicle concepts (except for deep sea drilling) have originated and been brought to fruitful research use in the ocean science or technology programs of any other agency (NSF, NOAA, etc.). These have in many instances funded programs to use craft developed under Navy sponsorship, but even in those instances have contributed little to major improvements in their capabilities.

It is to be hoped that these other groups will stimulate use of new techniques in the future, but it also seems logical that the Navy, as the principal U.S. user of the sea, should continue to take the lead in learning to work more effectively in its native environment.

BIBLIOGRAPHY

RUM/ORB

Alexander, C. M., "Sea Floor Technology report No. 5, Sea Floor Effectiveness of RUM II," *Mar. Technol. J.* pp. 9-15 (Aug. 1975).

Anderson, V. C. "Vehicles and Stations for Installation and Maintenance of Sea Floor Equipment," *IEEE Spectrum* 1 (11), 104-108 (Nov. 1964).

Anderson, V. C. "Maintenance of Sea floor Electronics," *IEEE Trans. Aerospace Electron. Syst.* AES-4 (5), 650-658 (Sept. 1968).

Anderson, V. C., "Spatial and Spectral Dependence of Acoustic Reverberation," *J. Acoust. Soc. Amer.* 42 (5), 1080-1088 (Nov. 1967).

Anderson, V. C. and D. K. Gibson, "An Experience with the ORB-RUM Sea Floor Work System,"

Handbook of Ocean Engineering (Japan) (in press), 1972.

Anderson, V. C., D. K. Gibson, and O. H. Kirsten, "RUM II—Remote Underwater Manipulator (A Progress Report)," Marine Technology Society, June 29-July 1, 1970, Washington, D.C., reprinted from Vol. 1, 6th Annual Preprints, 15 p., 1970.

Anderson, V. C., D. K. Gibson, and R. E. Ramey, "Electronic Components at 10,000 psi," SIO Ref. 65-6, May 20, 1965.

Ayala, F. J., J. W. Valentine, D. Hedgcock, and L. G. Barr, "Deep-sea Asteroids: High Genetic Variability in a Stable Environment," *Evolution* **29**, 203-212 (1975).

Gibson, D. K. and V. C. Anderson, "Sea-Floor Soil Mechanics and Trafficability Measurements with the Tracked Vehicle "RUM," in *Deep-Sea Sediments, Physical and Mechanical Properties, A. L. Inderbitzen, ed.*, Plenum Press, New York, 1974; *Mar. Sci.* **2** 347-366 (1974).

Noorany, I., O. H. Kirsten, and G. L. Luke, "Geotechnical Properties of Sea Floor Sediment off Coast of Southern California," Paper OTC 2187, *Proc. Offshore Technol. Conf.* 1975, vol. 1, pp 389-398, 1975.

Smith, K. L., and R. R. Hessler, "Respiration of Benthopelagic Fishes: *In Situ* Measurements at 1230 m," Science **184**, 72-73 (1974).

Thiel, H., and R. R. Hessler, "Ferngesteuertes Unterwasserfahrzeug erforscht Tiefseeboden" ("Remote Underwater Craft Explores Deep-Sea Bottom"). *Umsch.* **74** (14), 451-453 (1974).

ALVIN

Backus, R. H., et al., "Ceratoscopelus Maderensis: Peculiar Sound-Scattering Layer Identified with this Myctophid Fish," *Science* **160** (3831), 991-993 (1968).

Ballard, R. D., and K. O. Emery, "Research Submersibles in Oceanography," 70 p., Marine Technology Society, Spec. Publ., 1970.

Ballard, R. D., "Summary of the Geologic Dives Conducted in the Gulf of Maine during 1971 and 1972 by the Research Submersible ALVIN," 73 p., WHOI Ref. No. 74-29, 1974.

Breaker, L. C., and R. S. Winokur, "The Variability of Bottom Reflected Signals Using the Deep Research Vehicle ALVIN," 22 p., Naval Oceanographic Office, IR No. 67-92, Dec. 1967.

Donnelly, J. D., "1967—ALVIN's Year of Science," *Nav. Res. Rev.* 21, 18-26 (Jan. 1968).

Ellinthorpe, A. W., and R. G. Malone, "A Visual Ocean Bottom Survey off the Island of Santa Maria, Azores," 11 p., Navy Underwater Sound Laboratory, USL Report No. 1017, Apr. 1969.

Emery, K. O., "Positions of Empty Pelecypod Valves on the Continental Shelf," *J. Sed. Petrol.* **38**, 1264-1267 (1968).

Heirtzler, J. R., and X.Le Pichon, "FAMOUS: A Plate Tectonics Study of the Genesis of the Lithosphere," *Geol.* **2** (6), 273-378 (1974).

Jannasch, H. W., and K. Eimhjellen, "Studies of the Bio-degradation of Organic Materials in the Deep-Sea," in *Marine Pollution and Sea Life, FAO Conference on Marine Pollution*, M. Ruivo, ed., London, 1972.

Sanders, J. E., and C. S. Clay, "Investigation of the Ocean Bottom with Side Scanning Sonar," Proc. of Symposium on Remote Sensing of Environment, Institute of Science and Technology, University of Michigan, Willow Run Laboratories, 1968.

Schlee, J., "Geology from a Deep-Diving Submersible," *Geotimes* **12** (4), 10-13 (1967).

FLIP

Bronson, E. D., and L. R. Glosten, "FLIP Floating Instrument Platform," SIO Ref. 73-30, Nov. 15, 1973.

Bronson, E. D., Three-Point Anchoring in the Deep Ocean," *Proc. U.S. Nav. Inst.* **101** (2), 101-103 (Feb. 1975).

Fisher, F. H., and C. B. Bishop, Letter to the Editor: "FLIP as a Fleet Training Platform," *USN J. Underw. Acoust.* **25**(2), 525-530 (Apr. 1975.)

Fisher, F. H., and F. N. Spiess, "FLIP—Floating Instrument Platform," *J. Acoust. Soc. Amer.* **35**, 1633-1644 (Oct. 1963).

Fisher, F. H., and R. B. Williams, "Acoustic Bearing and Amplitude Measurements in the Thermocline of the Open Ocean," *USN J. Underw. Acoust.* **19** (3), 295-304 (July 1969).

Fisher, F. H., R. B. Williams, and P. Cushing, "Puerto Rican Experiments, Part I, short Range Acoustic Amplitude and Bearing Fluctuations of the Open Ocean in the Thermocline," 28th USN Symposium on Underwater Acoustics, Naval Research Laboratory, Washington, D.C., Nov. 17-19, ONR Report ACR-170, Vol. II, pp. 323-334, 1970.

Fisher, F. H., R. B. Williams, and F. M. Phelan, "Fluctuations in Surface Duct Propagation, *USN J. Underw. Acoust.* **25**(2), 373-383 (Apr. 1975).

Pinkel, R., "Upper Ocean Internal Wave Observations from FLIP," *J. Geophys. Res.* **80**(27), 3892-3910 (Sept. 20, 1975).

Pinkel, Robert, "Space-Time Measurement of Oceanic Motions from a Range-Gated Doppler Sonar," *J. Acoust. Soc. Amer.* **59**(1), S58 (Spring 1976).

Williams, R. B., F. H. Fisher, and P. Cushing, "Puerto Rican Experiments, Part II, Bearing Fluctuations in Short Range Bottom Bounce Propagation," 28th USN Symposium on Underwater Acoustics, Naval Research Laboratory, Washington, D.C., Nov. 17-19, 1970, ONR Report ACR-170, Vol. II, pp. 335-343, 1970.

Monster Buoy

Ender, A., "Environmental Data Buoys," *MIT Technol. Rev.* **76**(4) (Feb. 1974).

Gaul, R. D., and N. L. Brown, "A Comparison of Wave Measurements from a Free Floating Wave Meter and the Monster Buoy," Marine Technology Society Transactions, 2nd International Buoy Technology Symposium, Washington, D.C., Sept. 18-20, 1967.

Kosic, R. F., K. A. Morgan, and L. A. Scott, "Long Range Telemetering from the Monster Buoy," Marine Technology Society Transactions, 2nd International Buoy Technology Symposium, Washington, D.C., Sept. 18-20, 1967.

Morgan, K. A., L. A. Scott, and R. F. Devereux, "The Monster Buoy, its Data Acquisition and Telemetry/Command Systems," Marine Technology Society Transactions, 2nd International Buoy Technology Symposium, Washington, D.C., Sept. 18-20, 1967..

Deep Tow

Atwater, T., and J. D. Mudie, "Detailed Near-Bottom Geophysical Study of the Gorda Rise," *J. Geophys. Res.* **78**(35), 8665-8686 (Dec. 10, 1973).

Boegeman, D. E., G. J. Miller, and W. R. Normark, "Precise Positioning for Near-Bottom Equipment Using a Relay Transponder," *Mar. Geophys. Res.* **1**, 381-396 (1972).

Ivers, W. D., and J. D. Mudie, "Towing a Long Cable at Slow Speeds: A Three-Dimensional Dynamic Model," *Mar. Technol. Soc. J.* **7**(3), 23-31 (May-June 1973).

Johnson, D. A., "Ocean-Floor Erosion in the Equatorial Pacific," *Bull. Geol. Soc. Amer.* **83**, 3121-3144 (Oct. 1972).

Larson, R. L., "Near-Bottom Geologic Studies of the East Pacific Rise Crest," *Bull. Geol. Soc. Amer.* **82**, 823-841 (Apr. 1971).

Lonsdale, P. F., and B. Malfait, "Abyssal Dunes of Foraminiferal Sand on the Carnegie Ridge," *Bull. Geol. Soc. Amer.* **85**, 1697-1712 (Nov. 1974).

Lonsdale, P. F., and F. N. Spiess, "Abyssal Bedforms Explored with a Deeply Towed Instrument Package," submitted to Elsevier Scientific Publishing Co., Geology Science Section, (in press).

Luyendyk, B. P., and K. C. Macdonald, "Physiography and Structure of the FAMOUS Rift Valley Inner Floor Observed with a Deeply Towed Instrument Package," submitted to *Bull. Geol. Soc. Amer.* dedicated issue on FAMOUS (1976).

Normark, W. R., "Growth Patterns of Deep-Sea Fans," *Amer. Ass. Petrol. Geol. Bull.* **54**(11), 2170-2195 (Nov. 1970).

Spiess, F. N., "Recovery of Equipment from the Ocean Floor," *Ocean Eng.* **2**, 243-249, (1974).

Spiess, F. N., B. Luyendyk, and M. S. Loughridge, "Bottom Slope Distributions and Implied Acoustic Bearing Errors in Abyssal Hill Regions of the North Pacific," *USN J. Underw. Acoust.* **19**(2), 183-196 (Apr. 1969).

Spiess, F. N., J. D. Mudie, and C. D. Lowenstein, "Environmental Limitations to Deep Sea Search," *Proceedings of the 4th U.S. Navy Symposium on Military Oceanography, May 10-12, 1967, Washington, D.C.*, Vol. 1, pp. 69-80, Naval Research Laboratory (1967).

Spiess, F. N., and R. C. Tyce, "Marine Physical Laboratory Deep Tow Instrumentation System," SIO Ref. 73-4, Mar. 1, 1973.

Future Concepts

Apel, J. R. "SeaSat: A Spacecraft Views the Marine Environment with Microwave Sensors," AOML/NOAA, OCEAN 75 - MTS/IEEE Conf. Rep., Sept. 1975.

Lang, T. G., W. J. Sturgeon, and J. D. Hightower, "The Use of Semisubmerged Ships for Oceanic Research," Naval Undersea Center, OCEAN 75 - MTS/IEEE Conf. Rep., Sept. 1975.

May, A. E., and L. S. Tomooka, "Flippable Barge for Ocean Engineering Support," Scripps Institution of Oceanography Rep., SIO Ref. No. 74-30, Oct. 1974.

Spiess, F. N., "Stable Floating Platform Project," AEOL Report 60—Final Rep., SIO Ref. 74-17, May 1974.

Spiess, F. N., A. E. May, L. S. Tomooka, and D. R. Bellows, "A Flippable Barge for Ocean Engineering Support," MTS Conf. Rep., Sept. 1974.



T. G. Muir has been employed since 1961 at the Applied Research Laboratories of the University of Texas at Austin. During this time he has specialized in nonlinear acoustics, emphasizing practical problems of naval sonar applications. Dr. Muir has conducted a variety of measurements at sea and in the laboratory. He received a B.S. in Physics, an M.A., and a Ph.D. in Mechanical Engineering from the University of Texas. Dr. Muir is a member of the Acoustical Society of America, the British Institute of Acoustics, the U.S. Naval Institute, and the Society for Historical Archeology.

NONLINEAR ACOUSTICS: A NEW DIMENSION IN UNDERWATER SOUND

T. G. Muir

Applied Research Laboratories
University of Texas at Austin
Austin, Tex.

When the sound intensity in some regime of an acoustic process becomes so high that the principle of superposition no longer applies, we enter the domain of nonlinear acoustics. The limit of proportionality in the stress-strain or pressure-density relationship will have been exceeded, and the resultant vibrational disturbance at any point in the medium will no longer be equal to the sum of its individual components. When this happens, one must consider the nonlinear interaction of waves with themselves, with other waves, and ultimately with the medium.

The essence of the nonlinear acoustic mechanism is that an intense wave perturbs the medium in which it exists and this perturbation alters the natural order governing the wave's behavior. The wave then changes to accommodate new rules, further altering the perturbation and consequently the prevailing rules.

As one might expect, such a hectic state of affairs can exist only as long as the wave continues to significantly alter the medium. Beyond that point, the process returns to the domain of ordinary linear acoustics familiar to conventional expectations. Only a short passage through a nonlinear regime, however, is often sufficient to totally revise the original acoustical problem, yielding a new one that is as interesting as it is complex.

The fundamental mathematical foundation of nonlinear acoustics dates back at least to Euler [1]. Stokes' [2] ingenious realization of how an acoustic wave distorts nonlinearly, forming a shock wave, was a genuine highlight of the early period.

The subject began to develop sporadically in Europe and America during the 1930s and 1940s with further examination of this problem by Fay [3], Fubini [4], and Thuras, Jenkins, and O'Neil [5]. A theoretical paper by Eckart [6] dealing with nonlinear field theory opened up a new era bolstered by similar work by Lighthill [7].

According to Beyer [8], ". . . the use of perturbation theory in Eckart's paper was picked up and exploited by the Russian school, led by Andreev at the Acoustics Institute in Moscow. Although some American experimental work appeared in the middle 1950s, beginning with Fox and Wallace [9], the experimental work of Krasil'nikov and coworkers at Moscow Unviersity and of Mikhailov and his group at Leningrad University have led the field."

The Office of Naval Research played the major role in developing nonlinear acoustics in the United States during the 1950s, through the establishment of extremely productive research programs at several American universities. The investigators at Brown University, the Massachusetts Institute of Technology, the University of California at Los Angeles, Harvard University, Michigan State University, and Catholic University should be given special recognition for their singular contributions in this era.

The recent history of nonlinear acoustics clearly indicates a scientific renaissance, in direct relation to Westervelt's formulation of the parametric array [10]. His first discovery of this effect was made during an ONR tour of duty in London in 1952. Berktay's early papers on potential applications [11-15] were instrumental in calling parametric arrays to the attention of underwater acousticians and engineers.

Today nonlinear acoustics continues its expansion. Beside the Soviet and American efforts, there exists what may be called a European school of nonlinear acoustics with centers in England, Norway, Denmark, France, and Germany. Seven international symposia have been held on the subject, leaving it with a clearly multinational image. It is becoming more than just a science or a technology as nonlinear effects and techniques are recognized and integrated into the world's most powerful navies. Nonlinear acoustics is therefore a serious subject whose impact in these forces is of more academic interest.

For this article, I have been asked to introduce the subject of nonlinear acoustics for the purpose of discussing possible future directions of research. This task is not easy. One can only guess about the future, knowing full well that it most certainly won't work out as any one individual has planned.

I begin by discussing some of the most famous problems in nonlinear acoustics. A brief review of the history of each problem will serve as a reminder of the trials and tragedies experienced by those engaged in research. The approaches taken to problems and solutions of the past are indispensable to the appraisal and efficient execution of future investigations.

Archival papers are referenced where appropriate to provide useful leads for those interested in further study. All figures are from the author's files, except where indicated otherwise.

FINITE AMPLITUDE DISTORTION

Since sound propagation is characterized by the elastic transmission of disturbances among the fundamental particles of a supporting medium, it can be reasoned that increasing the particle density increases both the efficiency and speed of sound transmission. The disturbance itself c[illegible]ies pressure and therefore offers a self sustaining means of altering the density. The result is that segments of the disturbance in compression travel faster than those in rarefaction. When this happens, the acoustical wave alters its shape, steepening as it travels along. The alteration is infinitesimal for weak disturbances and gradually increases in importance with increases in both amplitude and frequency of sound.

This expectation seems reasonable and even elementary, but it apparently agonized some of the greatest minds in classical physics. According to some recent treatises on the history written by Blackstock [16] and reviewed later by Bjørnø [7], both the great French mathematicians Lagrange [18] and Poisson [19] obtained mathematical proof of this phenomenon but just couldn't believe their own results. Lagrange discarded his solution because "the new formula would destroy the uniformity of the speed of sound and would make it depend in some way on the nature of the original disturbances; that which is contrary to all experiments." Poisson similarly failed to fathom the consequences of his own findings, suppressing their real meaning with the conclusion that "all sound, loud or faint, is transmitted with the same speed."

Almost half a century passed until the great British physicist Stokes [2] came up with the correct interpretation of what can now be called the first problem of nonlinear acoustics. Involved in a debate with his colleagues over whether or not a plane wave of sound could even exist, he gave the first clear description of the progressive waveform distortion implied by Poisson's solution and even produced a sketch of the process. Blackstock observes that "Stokes' paper touched off a torrent of controversy; a total of twelve (argumentative) papers by Challis, Stokes, and Airy followed during the next twelve months." The correct interpretation was eventually accepted and further delineated in the work of Earnshaw [20], Riemann [21], and Rayleigh [22]. (Figure 1).

ACOUSTIC NONLINEARITY

The first problem of nonlinear acoustics opened up a wide variety of new questions that got only

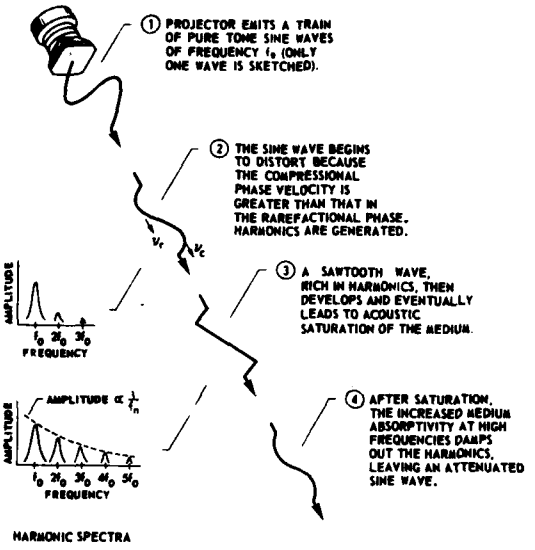


Figure 1—Case history of a high-intensity sound transmission, illustrating finite-amplitude distortion and harmonic generation.

sporadic attention until the 1930s. The early investigators worked primarily with gases, where the density is low and the particle velocities are so high as to require consideration of the sonic "wind" or "convection" in addition to the elastic nonlinearity wrought by self-imposed alterations in density. Although due to a strictly linear phenomenon, the behavior of gas particles carried along by the wave's own velocity manifests itself in an identical nonlinear fashion.

In simple algebraic terms, the effect of convection on the velocity of a particular point on a waveform is given by

$$v_{(x)} = c_0 + u_{(x)},$$

where c_0 is the well-known sound speed constant (1100 ft/s, or 335.28 m/s, in air) and $u_{(x)}$ is the oscillating particle velocity at the point in question. The particle velocity and sound speed are therefore linear components of the wavelet velocity. It can thus be seen that steepening of the waveform can result from the accentuating effect of the particle velocity component $u_{(x)}$, whose periodic changes in direction and amplitude in the waveform cause the compressional phase to hurry up and the expansionsal phase to slow down.

In most media, however, the elastic nonlinearity must also be considered. When this is done, Eq. (1) becomes

$$v_{(x)} = c_0 + (1 + \tfrac{1}{2}\, B/A)\, u_{(x)}.$$

Here, B/A is the constant specifying the elastic nonlinearity. It arises from a mathematical definition of convenience in which a series expansion is made of the pressure-density relationship for an acoustic wave.

The French physicist Biquard [23] was the first to quantify the nonlinearity constants for liquids. His work, which many subsequent investigators have apparently missed, contains several other first achievements of note. It turns out that water, including seawater, has a value for B/A of 5.2, while air has an equivalent B/A of 0.4. Thus, as Eq. (2) shows, the convection effect dominates in air (83%), while the elastic nonlinearity dominates in water (72%). Some media are more nonlinear than water. Mercury has a B/A of 7.8; ethyl alcohol has a B/A of 10.4.

Tabulations of fluid nonlinearities have been made by Zarembo and Krasil'nikov [24], Beyer [25], and Mikhailov and Shutilov [26]. These tables have been quite useful in nonlinear acoustics research. However, the nonlinearity of only 40 or so fluids has been determined to date. The B/A parameter can be calculated quite accurately from knowledge of some of the thermodynamic constants of a medium, with the aid of measurements on how the mean sound speed varies with temperature and static pressure [27]. It can also be measured directly by both acoustic [25] and optical techniques [28].

Considering the appropriateness of such an investigation for physics and standards laboratories, it is difficult to understand why so few media have actually been examined. The nonlinearity of the media under extreme environmental conditions (involving state and phase changes, ionization, etc.) would also appear to offer justifiable grounds for future research.

Basic information of this type is needed for the development of nonlinear science and technology in such fields as fluid and solid-state electronics. The nonlinearity parameter also appears to offer a means of characterizing a wide category of materials and substances. For liquids, the sensitivity of

B/A to the amount of entrapped gases suggests techniques for making remote, nondestructive measurements of gas content, a topic of great interest in diver medicine.

Thus, the door remains open for physicists and engineers to develop new avenues of research and development around the nonlinearity problem. The first step for this particular example and for many others is to broaden our understanding of the basic nonlinearity for various media and for new configurations.

SHOCK FORMATION

The culmination of nonlinearly induced distortion is the formation of discontinuities in the waveform pressure profile. The waveform train then resembles the teeth of a saw blade; hence the name "sawtooth" waves.

A similar wave results from the flight of supersonic aircraft. In this case, the over-and-under pressures caused by the nose and tail of the airplane cause a head and tail shock resembling one cycle of a sawtooth wave.

Actually, perfect discontinuities do not form in any real shock wave. Historically, this phenomenon has amounted to quite a bit more than a superficial qualification because a respectable amount of physics transpires at a shock front. For example, given that a wave distorts and steepens, what keeps it from overshooting the zero crossing and becoming multivalued? This question is not insignificant, as Stokes recognized, because a multivalued waveform implies the seemingly impossible situation of having more than one amplitude at a time.

FINITE AMPLITUDE ATTENUATION AND SATURATION

This difficulty stymied the classical investigations throughout the 100 years or so following Earnshaw [20]. Blackstock writes, "At the bottom, the trouble lay in the neglect of dissipation (the conversion of sound to heat). Dissipation prevents the formation of discontinuities. Put another way, shock propagation is always accompanied by energy loss." (Ref. 16, p. 15.) Following up the work of Rankine [29], Rayleigh [22], and Taylor [30], Fay [3] showed that the tendency of a wave to steepen indefinitely is balanced by dissipation at the shock front, and this phenomenon prevents the development of multivalued waveforms (Figure 2).

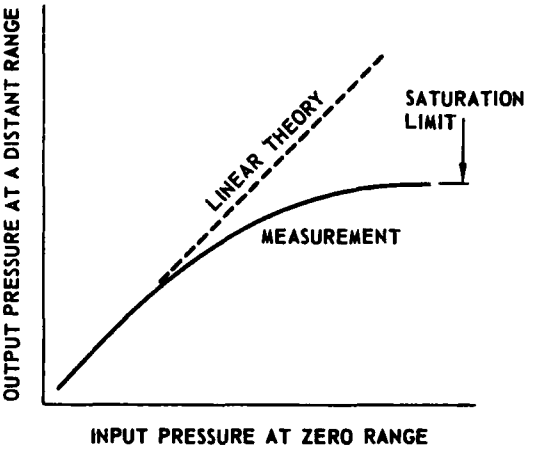


Figure 2—An example of saturation. As one increases the source level, more and more energy is wasted in nonlinearly induced absorption. There is a maximum attainable pressure level permitted by saturation for each range and frequency.

The Office of Naval Research played a major role during the 1950s and early 1960s in focusing attention on the extra absorption induced by dissipation in nonlinear waveforms. Their impressive list of sponsored work includes papers by Mendousse [31], Fox and Wallace [9], Towle and Lindsay [32], Narasimhan and Beyer [33], Rudnick [27], Cook [34], Blackstock [35], and Barnes and Beyer [36]. A noteworthy development in the ONR program came toward the end of this era when Lester [37] reported the first definitive measurements on acoustic saturation in water.

Acoustic saturation occurs when the dissipation at the shock fronts, in its role of countering further distortion, limits further increases in sound pressure amplitude. In other words, the dissipation simply sidetracks any and all brute-force efforts to "break" the sawtooth waveform by channeling any additional input source levels into heat. Thus the amplitude of a wave cannot increase indefinitely, giving some justification to

the term "finite amplitude." In one of the latest examinations of this phenomenon for ONR, Shooter et al. [38] extended theory and experiment to the spherical waves used in naval sonar systems.

This problem has only recently been appreciated by systems engineers involved in high-frequency sonar design. Many sets, for example, are extremely overpowered, carrying huge transmitters in the tens of kilowatts range when only a few kilowatts would have produced the same sound pressure level at the target.

HARMONIC RADIATIONS

Although acoustic saturation is a deleterious effect in sonar and in most other practical applications of high-intensity sound, there is one potentially useful aspect of this entire process that has not yet been fully exploited.

If one considers the frequency domain explanation of the distortion or waveform steepening effect, it is easy to show that progressive distortion is accompanied by the progressive generation of harmonic components in the distorted waveform [4]. Each component is an integral overtone of the fundamental or original frequency, and measurements have shown that each grows with increase in propagation distance [39]. The growth of harmonic components is abated when shocked waveforms are formed and the wave passes into a regime characterized by increased dissipation [40]. Before the harmonics are eventually dissipated at long range, their amplitudes go as 1/n times that of the fundamental, where n is the harmonic number. Thus, the second harmonic is only 6 dB less than the fundamental, the third 10 dB less, and so on (Figure 3).

The distorted finite-amplitude waveform is therefore quite rich in harmonics and possesses the ability to irradiate samples and targets over an extremely wide range of frequencies. This ability appears to be very useful in a wide variety of acoustical measurements because it is always difficult to make sound radiators operate over large frequency ranges. Sound receivers, on the other hand, can have wideband capabilities because they do not have to be tuned to their electrical terminations for maximum power transfer. One develops the wideband transmitter system by simply driving a tuned projector at some fundamental frequency, allowing the harmonics to be generated in the medium [41]. The harmonics can then be passed through a sample or can be reflected from a target and then received with a wideband hydrophone [42].

This technique has obvious future potential as a measurement tool for a wide variety of problems in applied acoustics. One such application arises in biomedical acoustics, where it is often desired to examine the absorptivity and sound velocity of tissue over a wide frequency range. Similar measurements in marine geophysics appear to provide

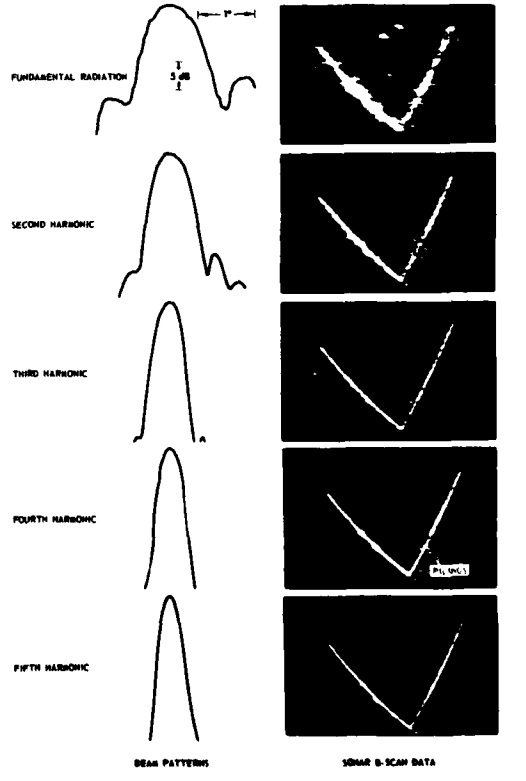


Figure 3—Harmonic radiations and parasitic sonar. The nonlinear harmonics generated in a high-intensity sonar beam "go along for the ride," propagating with the fundamental and reflecting off the target, in this case the corner of a barge. Although each successive harmonic has less source level, their directivities increase progressively giving better delineation of the field of view. The harmonic radiations have not been used in scanning sonar, because of unresolved signal processing problems.

useful information concerning the structure and compositon of sedimentary strata [43].

Naval sonar applications have similar potential for further study, especially when one considers the characteristics of the harmonic beams when the fundamental beam is radiated from a directional source [44]. A recent study by Lockwood et al. [45] shows that each successive harmonic beam pattern is related to that of the fundamental raised to the nth power, where n is again the harmonic number. This relationship causes the beamwidth and minor lobe level to be reduced as one goes from second to third to fourth harmonic, etc. These factors (beamwidth and minor lobe suppression) are all important in sonar because they ultimately limit the system's angular resolution and suppression of false targets. Thus, by using the fundamental frequency for course detection, one can go higher in the harmonic sequence as the range is closed to realize better angular discrimination. At present, the high-speed techniques now used to scan sonar beams across the target field are not compatible with the single-beam operation to which the presently configured harmonic radiations appear limited. It will be interesting to see what solutions to this practical problem the signal processing community may offer in future research.

It should be mentioned that the harmonic radiations are amenable to study with some interesting optical techniques [46, 47]. Further, the reflection of intense harmonic radiations from various surfaces produces unique phase and propagation properties characteristic of the boundary condition [48, 49]. These effects clearly offer future researchers some unusual tools capable of synthesizing and isolating special problems in nonlinear acoustics.

PARAMETRIC ARRAYS—A CHRONOLOGY

The recent history of nonlinear acoustics shows that a veritable renaissance in research, development, and application has occurred in direct relation to Westervelt's formulation of the parametric array [10]. Although the subject of finite-amplitude propagation was already a respectable topic, as the rich history mentioned on the previous pages indicates, the potential advantages of parametric array applications in ocean acoustics soon captured the imagination of a wider group of scientists and engineers (Figure 4).

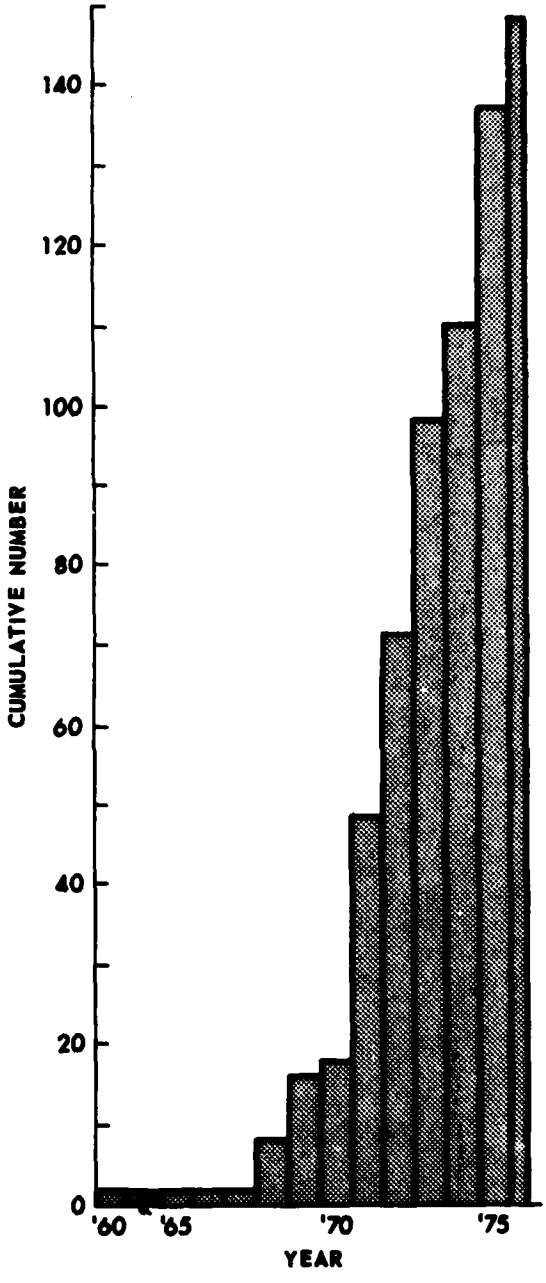


Figure 4—Papers on parametric arrays presented at acoustical society meetings and special symposia demonstrate the popularity of this subject.

The parametric array uses the nonlinear properties of the medium to generate superdirective sound beams at low operating frequencies. Albers' [50] brief account of the first observation of a parametric array bears repeating here:

> While Dr. Westervelt was stationed at the London, England branch office of the U.S. Office of Naval Research in 1951, he met the late Captain H.J. Round, English pioneer in the development of the superheterodyne receiver. Captain Round was carrying out experiments with an underwater magnetostriction projector in his private laboratory. The work was being done for Dr. Paul Vigoureux who was then at the Admiralty Research Laboratory. Captain Round happened to have an 18-kHz transducer operating in air and when Dr. Westervelt walked in front of the beam, he was startled to hear a loud low-frequency hum, rich in harmonics, but highly directive, coming from such a tiny projector. The fundamental he heard seemed about 100 Hz, while the emitter was not more than about six inches on a side. He immediately concluded that Round was supplying his RF driver either with an unfiltered power supply or at worst raw a.c., and that the demodulation was occurring either in the air or in his own ears. It was at this moment that the concept of an end-fire array first occurred to him.

Once again, the Office of Naval Research played a key role in a discovery having a momentus impact on science and technology. Here was also a clear example of a scientist coupling his keen observation with a highly developed sense of physical intuition. A scientific purpose was served first as the phenomenon was carefully put into the most elegant theoretical formulation, so elegant, in fact, that it took less than two and a half pages when finally published in 1963.

This paper showed how two high-frequency radiations, each confined to a narrow beam, interact with each other to produce sound at the sum and difference of the two original frequencies. The interaction takes place over a relatively long pathlength in the medium, lending an array like or antennalike aspect to the entire process. The longer the interaction distance, the longer the array and the more directivity the beam has. Later, it was shown that the parametric array was perfectly shaded by the exponential absorption of the original radiations, enabling the difference frequency beam to be completely free of undesirable diffraction lobes. Still later, the wide-frequency band capability of the parametric array was discovered. At first, however, the unique feature of the parametric array was its ability to develop a very narrow difference frequency radiation from a small ultrasonic sound source, driven hard at two frequencies (Figure 5).

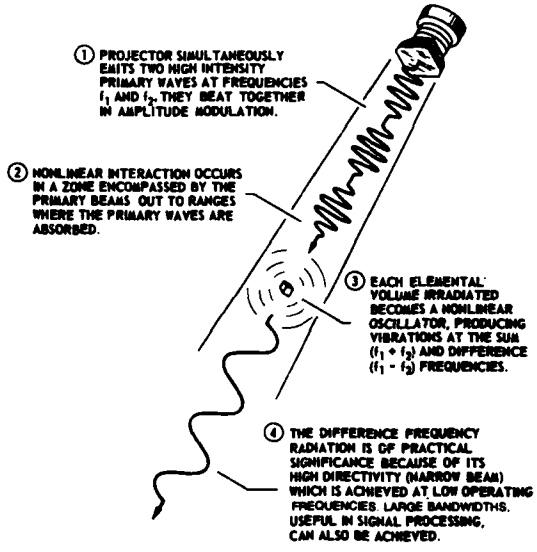


Figure 5—Processes in a parametric transmitting array.

Although others had predicted and measured the fundamental interaction phenomenon [51, 5], it was Westervelt who saw the array aspect of the problem (Figure 6).

When Westervelt first read his paper on the parametric array at the Providence meeting of the Acoustical Society of America in 1960, he received a mixed reception. One of my former acoustics professors was impressed—". . .that's clever, wish I'd thought of that." Others were more satirical—". . .it must be nice to have the time to play around with second-order effects." The vast majority, however, shared some skepticism about its practicality, especially since the

parametric process is so inefficient. The conversion of energy from primary to secondary sound depends in a rather complicated way on many parameters, the ratio of frequencies, the input power, etc., and the efficiency has never been much more than 1%, although this may change with further research.

Despite the fact that Bellin and Beyer [52] produced experimental evidence of the existence of the parametric array at the same meeting, the efficiency issue proved to be too big a stumbling block for immediate consideration of parametric arrays, at least in the United States.

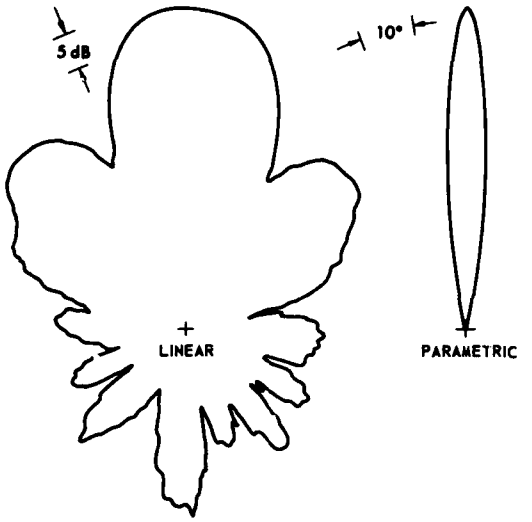


Figure 6—Comparison of beam patterns from projectors of the same size and frequency. Despite being much broader, the linear radiation pattern displays the usual diffraction lobes, while the parametric pattern is completely free of this undesirable effect.

In Europe, however, the practicality of the eventual applications of parametric arrays was, in the beginning, of no real consequence. The first papers on parametric arrays originating from elsewhere than Brown University were published by the Norwegian theorists Lauvstad, Naze, and Tjötta [53, 54]. In England, Professor Tucker [55] prompted Berktay [11-15, 56] to stem the tide of skepticism surrounding parametric arrays in a noteworthy series of papers on the engineering aspects of the problem. The year 1965 was a remarkable one for parametric array research in the European school; six papers from England and Norway were published, most on the theory but some with experimental results. The year 1967 was also impressive, as more model tank experiments were reported [57] and designs for parametric sonars were proposed [13-15]. Enough work had now been done to provide a basis for debate. Berktay [14] and Tjötta [58] agreed to disagree and Zverev, Kalachev, and Stepanov [59] spoke out from the Soviet Union, strongly criticizing some of Berktay's assumptions and predictions.

In the U.S., nothing further had been done on parametric arrays since the first model tank experiments of Bellin and Beyer [52], but the activity in Europe began to be interesting. Browning and Mellen decided to host a symposium on nonlinear acoustics at the Naval Underwater Sound Laboratory in New London, Conn., and Berktay was invited to speak. I was fortunate to attend this meeting, held in May 1968, where Mellen, Beyer, Cook, and Marsh also spoke on other finite-amplitude acoustics problems. Westervelt was there and joined in the discussion.

It was really a turning point for many of us because we were able to see that Berktay's approach circumvented the efficiency issue by considering the *total* problem, i.e., not only parametric effects but also the way they fit in with the environmental and engineering limitations of the application at hand. Berktay's [60] report was heavily laced with interesting model tank experiments and even included material on parametric reception. It helped us overcome the fundamental stumbling block of how much energy was going to be lost in parametric conversion by focusing attention on what could be done with what was left.

On the way back to Texas, I made up my mind to do some experiments on parametric arrays. By the summer of 1968, Joe Blue and I were well underway, encouraged and supported by the Office of Naval Research. Our work in a freshwater lake allowed us to get around the size limitations of laboratory tanks and permitted measurements to ranges in excess of 100 yd (91.4 m) [61]. By going to long ranges, it was possible to show that the parametric array had not minor lobes in its farfield radiation pattern (Figure 7).

About the same time, another Soviet paper on parametric arrays appeared [62], reporting work that rivaled the Norwegian and English investigations in the clever use of model tank facilities for

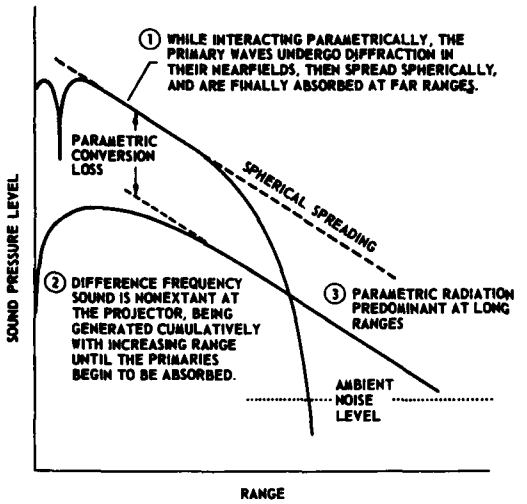


Figure 7—Case history of parametric array generation and propagation.

highly productive measurements. This paper was one of the last on parametric arrays to be published in the Soviet Union for the next 8 years.

In November 1969, another symposium was held under ONR sponsorship at the Applied Research Laboratories of the University of Texas at Austin. I had the pleasure of hosting this meeting, at which some truly exciting new work was discussed. Survey papers by Blackstock [16] and Berktay [60] put things in perspective, and the papers on parametric arrays treated beamwidths, source levels, the wide bandwidth capability, saturation effects, phase considerations, and transient effects [63]. Westervelt was encouraged to speak on nonlinear acoustics at this meeting and gave his first paper on parametric arrays in over a decade, in which he was heavily involved in general relativity.

Among other things, Westervelt touched on the parametric transients, which were first predicted by Berktay [11, 12] and were the subject of some beautiful experiments by Moffett et al. [64]. These transients are created when a short acoustic pulse is transmitted, and they can be explained from both a frequency and a time-domain argument as the "self-demodulation" of the primary transmission into lower frequency components (Figure 8). The parametric transients have the same effective directivity as the pure tone

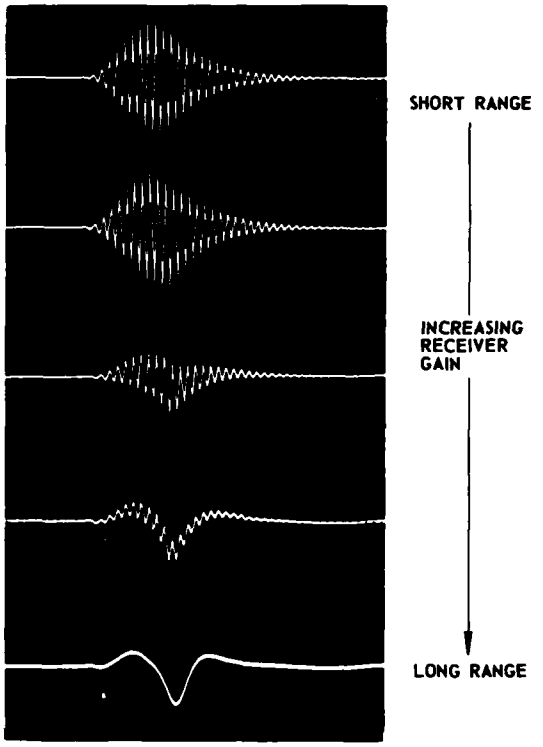


Figure 8—Parametric self-demodulation. A single, continuous wave pulse interacts with itself to form a parametric array transient. These measurements, taken from Ref. 64, shows the low-frequency signal being formed. At long ranges, the original pulse is absorbed, leaving only the transient. Self-demodulation occurs in every high-powered sonal transmission.

parametric array; thus they offer a means of transmitting superdirective impulses that can be used for some unique measurements.

By this time, the parametric array program at the Naval Underwater Systems Center (NUSC) was well under way [65] and has continued to expand to this day.

It is appropriate to close this chronology by describing a remarkable development that has indelibly changed the climate for nonlinear acoustics research—the successful development of the first practical device employing a parametric array. At the next symposium on nonlinear acoustics, hosted by Berktay at the University of Birmingham in England during April 1971, the Raytheon Company discussed their tests on parametric echo sounding [66] (Figure 9). By developing a narrow, 12-kHz difference frequency



Figure 9—Echo ranging: a comparison, made at NUSC, of A-scan echo data for systems working against a target in a cluttered harbor [74].

beam from primary radiations in the 200-kHz band, they were able to acquire bathymetric data with 10 times the resolution of an ordinary 12-kHz depth sounder. This development ended an era of so-called "pure" research by opening a new era in which research was more strongly coupled to the search for new applications.

Simultaneously with this development, C. E. Fox of the Naval Ship Systems Command and R. F. Obrochta of the Office of Naval Research initiated a dialogue on the use of parametric depth sounders in the U.S. Navy. ARL/UT was asked to develop a prototype depth sounder for technical evaluation. This unit was successfully tested, and Raytheon then went into production, providing the Navy with parametric echo sounders for use on the newest, most advanced warships of the fleet.

The parametric array could now be truly called one of ONR's most remarkable discoveries. Not only was it a scientific achievement, but it was also destined to have an impact on the technology of naval operations.

PARAMETRIC ARRAYS—STATUS AND FUTURE

The dust has not yet settled on the past 5 years of parametric array research; we are still sorting out the multitude of papers, reports, and developments that have appeared. For this reason, I will be content to point out some highlights from this era and to offer some speculations on the future, as seen through the eyes of an experimentalist.

It is important to realize that most of current work on nonlinear underwater acoustics, at least in the United States and Europe, is being done in the area of parametric arrays. Since parametric arrays have been around since 1960, it is not surprising that the great emphasis today is on their practical application to a myriad of problems in naval science and technology.

Like it or not, basic research on parametric interaction phenomena is not currently of high priority at ONR or at the other sponsoring agencies in the naval community. Thus, the immediate future in this area seems to involve what we can do with what we already know. Hopefully, the climate for basic research can become more agreeable in the years to come so that the pump can once again be primed for the free flow of truly new fundamental concepts.

The pursuit of practical applications of parametric arrays is a demanding task, for it requires a broad, up-to-date knowledge of just about every subject in underwater sound. For example, in order to know what is practical and what is not, the scientist must know everything about the relationship of acoustics to the oceanic environment, including the sediments, solar heating and thermal structures, atmospheric phenomena, bottom roughness, etc. He also needs a good knowledge of the assets and limitations of existing systems and hardware.

PARAMETRIC RECEIVERS

It is important to begin with the concept of the parametric receiving array. First mentioned by Westervelt in his original paper [10], this idea was followed up in the basic laboratory tank studies of Berktay and Al-Temimi [56] and of Zverev and Kalachev [67]. These experiments served to confirm the concept and have also prompted several ONR field experiments on this topic [68, 69]. This work has been of interest to technologists involved in the passive surveillance of submarine traffic.

The parametric receiver develops a relatively high directivity at low frequencies by using a powerful high-frequency pump beam to interact nonlinearly with a low-frequency signal wave that is to be detected. An end-fire array is formed over the path from the pump to a hydrophone in much the same way as for the parametric transmitter. By choosing the pump frequency to be in the

"noise window," a minimum in the sea noise spectrum between 30 and 100 kHz, it is possible to operate such a device in an optimally quiet region. The end result is the creation of a directive end-fire array whose resolution characteristics are equivalent to those of a continuous end-fire array of the same length (Figure 10).

When most initiates to parametric reception learn of this equivalence, the eternal question inevitably arises: "Why not just use a (more efficient) linear array?" The answer lies in the economy of the problem. With the parametric device, only two small transducers are required rather than a more expensive array having a continuous distribution of elements. Furthermore, the parametric receiver has the advantage of good discrimination against noises arising in the backward direction, and it has an extremely facile capability for receiving sounds over wide frequency ranges. Figure 11 compares the beam patterns of linear and parametric receivers.

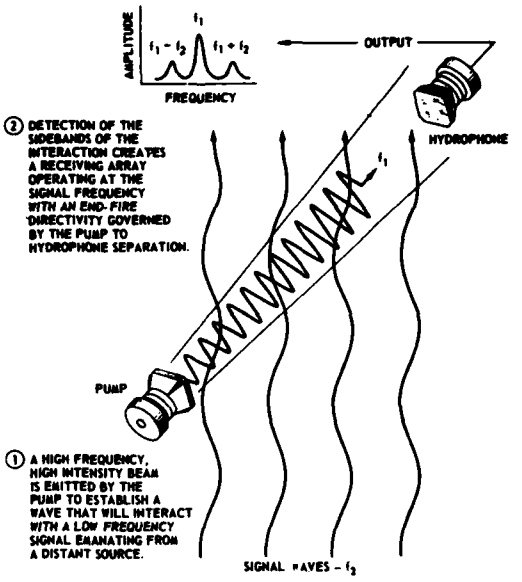


Figure 10—Concept of the parametric receiver.

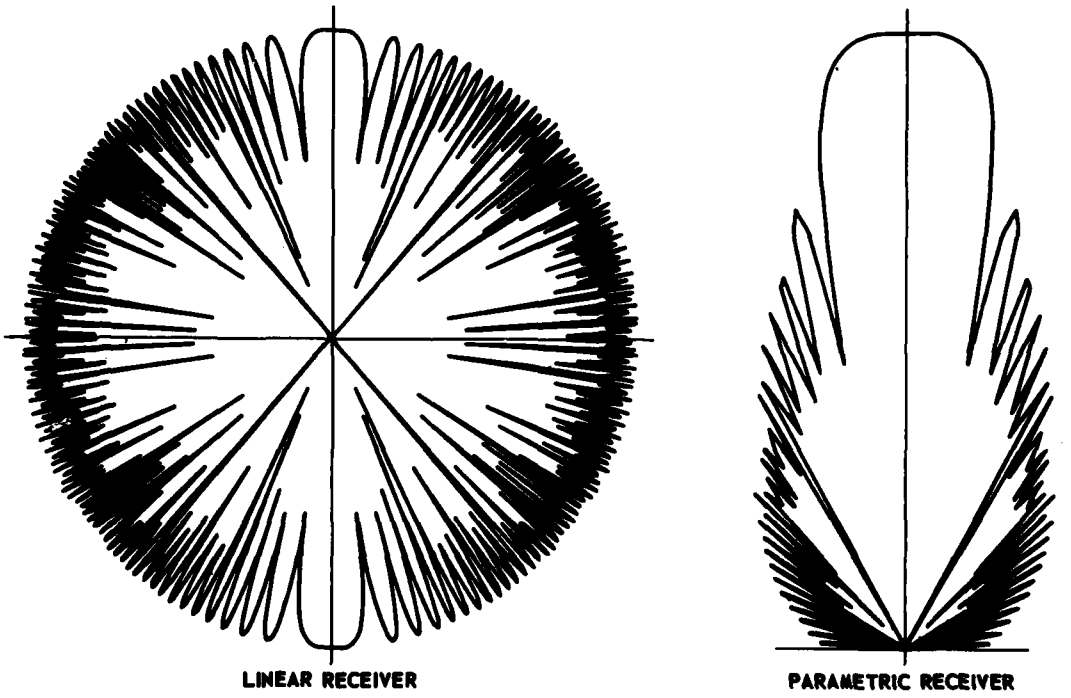


Figure 11—Comparison of beam patterns for linear and nonlinear operation of two transducers separated by 50 wavelengths.

At present, the rigorous requirements on the special electronics required for parametric reception are limiting its immediate wide-scale implementation. There is no doubt, however, that these engineering problems can be overcome, probably permitting the circuitry to be microminiaturized and contained in the hydrophone itself. Truchard's work [70], which even includes theory and experiment on techniques for processing the signal in the time domain, has provided considerable insight into the delicate mechanisms of parametric reception.

Despite the considerable skepticism surrounding parametric reception, many of us remain convinced of its ultimate utility, both as a physical model for further research in the promising field of parametric amplification [71] and as a tool for acoustic measurements in oceanography. How else, for example, are we ever going to develop techniques for surveying the directive properties of low-frequency noise in the ocean, an important project yet undone, whose execution by conventional linear techniques would require resources no country would commit (Figure 12)? Parametric receiver techniques, on the other hand, could yield not only the directional spectra but also the spatial correlation, as well as many other directivity-related factors of fundamental statistical importance.

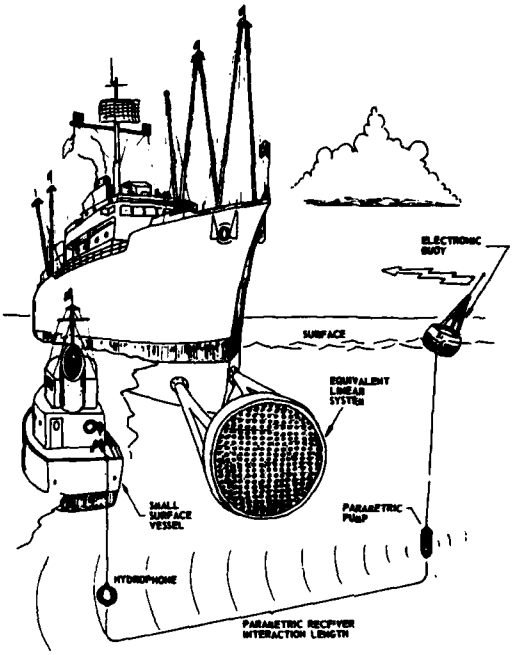


Figure 12—Comparison of experiments for ambient noise measurements. This sketch illustrates the point that a parametric receiver can be deployed from a small oceanographic research vessel to make directional measurements of ambient noise fields at low frequencies. The equivalent linear system involves a much greater investment in apparatus and logistics.

BATHYMETRIC PROFILERS

Since the Raytheon Company's first tests with parametric bottom profilers [72], additional tests have been undertaken at Honeywell [73], NUSC [74], and elsewhere. All of these investigators have focused on the advantages made possible by the high-resolution parametric beam that also has a low enough frequency to penetrate the sediments. Each new project has gone a bit deeper into the problem, and this trend will undoubtedly continue in the future, enabling more to be learned about marine sedimentation as well as acoustic instrumentation.

One of the most recent developments is now in progress in Norway where the Simrad firm has teamed up with the Universities of Trondheim and Bergen to design and test one of the most advanced parametric sonars yet developed. Their unit uses a towed array to receive the parametric echo, which is a sophisticated phaseshifted signal derived from the Barker code. This provides additional noise suppression without sacrificing time resolution [75].

Yet another advanced development is underway at Raytheon, where the second generation of parametric profilers is being developed for use in offshore mining for rare minerals (Walsh, private communication, 1976).

I have long been a strong supporter of parametric applications in this area and have had the opportunity to discuss this problem in detail [43]. The development of a narrow parametric beam at frequencies low enough to penetrate the sediments seems a natural combination of the right things for the right job (Figure 13).

It further seems that the technique has not even begun to be exploited, due to the fact that the parameters chosen for the existing prototypes are really limited to sediment penetrations measured

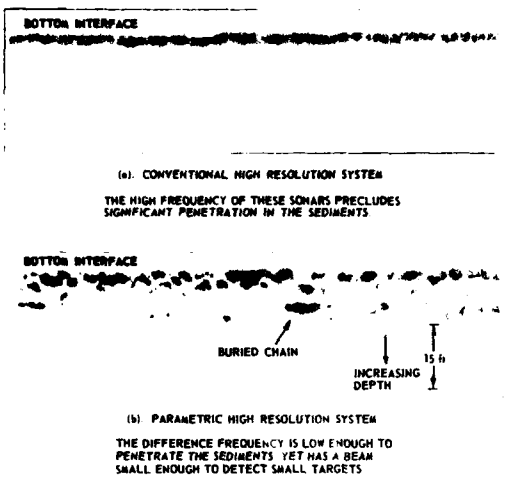


Figure 13–Parametric subbottom sonar. A comparison of oceanographic recorder outputs demonstrate the detection of an anchor chain buried in an alluvial mud and sand sediment in the harbor at Portobello, Panama.

in the tens of meters. It remains for some geophysical research agency to fully realize this potential by sponsoring a deep-penetration experiment in the seismic frequency band (about 50 to 150 Hz).

The list of potential payoffs for such an endeavor is impressive and includes such possibilities as using a high-resolution parametric transient to measure the phase shift of each successive sedimentary layer. This list could help classify each formation and may have economic significance in an energy-dependent society. For example, echoes from gas-bearing strata (called "hot spots" from their relatively high target strength) might then be classified from considerations of more than just amplitude.

But why stop at the energy bearing formations? Why not go for the deep crust, the mantle, and even the core of the earth by providing the geologist with a tool capable of answering the incredibly fundamental questions we have about our own planet?

Mantle reflections are usually carried out between transducers situated at the critical angles necessary for maximizing the energy arriving at the hydrophone. At the low frequencies required, the radiations are of poor angular resolution. This difficulty raises critical questions as to what benefits would accrue with a narrow parametric beam and whether it would be strong enough to overcome the noise at the receiver. These issues are crucial, no doubt, and remain to be settled by future research and cooperation between nonlinear acousticians and seismologists.

BURIED OBJECT DETECTION

In coastal waters, the hydrodynamic environment is usually quite active. Legions of poets have characterized the shifting sands and their appetite for devouring everything falling upon them. This activity poses special problems in both military and civilian operations because things like mines, ancient ships, and pipelines simply get lost in the sediment. Magnetics can sometimes be used for locating lost objects if they are large (like shipwrecks) and contain large quantities of iron. However, the magnetometer is essentially a point sensor with little or no angular resolution, and this factor poses special problems in the remote delineation of suspected targets. Although computer-aided techniques can be used to alleviate this problem, a better approach is the use of high-resolution sonar.

Sonars have problems too, and the major difficulty is to develop a system having good enough angular resolution while having at the same time a low enough frequency to penetrate the sediments at low grazing angles. This represents yet another natural setting for parametric sonar. We have been looking at this problem for several years from both scientific and engineering vantage points.

By burying an array of hydrophones, it has been possible to beam a parametric array through the sediment to measure its susceptibility to acoustic penetration [76]. Those experiments produced an unexpected result; it was determined that energy was transmitted into the bottom at grazing angles below the classical critical angle for total internal reflection, below which the energy is usually reflected and contained in the overlying water column. The dialogue on this problem has recently been joined by two theorists at the Naval Research Laboratory [77]. Their analysis shows that parametric generation in the region of the beam directly overlying the sediments enables the parametric sonar to literally "drop its waves" into

the bottom at higher, more efficient grazing angles. Thus, a purely scientific exercise, valuable in its own right, has no small impact on the feasibility of future concepts for subbottom sonar.

MODE SELECTION IN SHALLOW WATER

Since the absorption of sound in seawater increases with frequency, long-range systems naturally operate at low frequencies. On the continental shelves and in such critical shallow water areas as the North Sea, the Baltic, and the Mediterranean, the water depth is often no more than 10 to 100 acoustic wavelengths. When this situation obtains, the water column becomes an acoustic waveguide, much like the antenna feeds on radar systems.

Waveguide propagation can be very complicated, especially in underwater acoustics where the rough ocean boundaries introduce scattering losses and where solar heating establishes thermal gradients that refract the sound beams. Perhaps the most confusing aspect of shallow-water acoustics is the simultaneous excitation of several interfering modes of propagation at the transmitter. Multimode excitation is unavoidable with most conventional, low-frequency sound sources. Figure 14 illustrates some of the principles of shallow water wound propagation.

Parametric arrays offer a means of simplifying the propagation picture by selectively exciting discrete modes in the waveguide [78]. At the same time, directivity is gained to greatly reduce the reverberation. Other possibilities exist and are being addressed in current ONR research in this area (Figure 15).

It is safe to say that the parametric approach to shallow-water acoustics looks very promising but that research on this problem is truly in its infancy. Future ONR projects to take the current work out of the scale model stage and into full-scale research on the continental shelf should be expected to address such topics as wideband excitation and propagation, target resonances at low frequencies, modal target classification, mode-locked undersea communications, and improved capabilities for conducting a wide class of Doppler measurements. Each of these topics has important implications for naval science as well as oceanography.

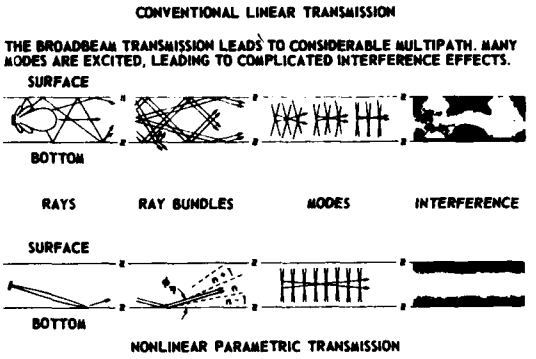


Figure 14—Concepts in shallow-water propagation.

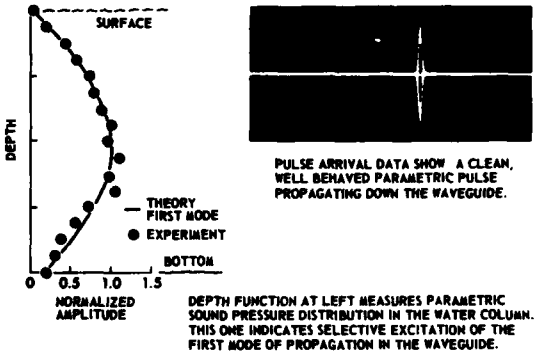


Figure 15—Model experiments in a shallow-water lagoon. The relationship of nonlinear acoustics to normal mode propagation is being examined for ONR on a preliminary basis in a saltwater estuary before full-scale measurements at sea.

DOPPLER MEASUREMENTS

The apparent change in the frequency of sound due to the relative motion of source and receiver has been a classical measurement problem since the time of is discoverer, Christian Doppler. In underwater acoustics, Doppler techniques are of prime importance in oceanography, where they are used to measure the dynamics of the sea surface, as well as in naval tactics, where they are used for navigation and for determining the presence and movement of targets under surveillance.

Parametric arrays offer a breakthrough in Doppler measurements because of their narrow sound

beams that are completely devoid of undesirable minor lobes. The superdirective parametric array is immune to the Doppler frequency spread generated by moving objects (such as the sea surface) that are insonified by minor lobes in the near vicinity of the system. This immunity is extremely helpful when it is desired to measure the Doppler of some remote process that may subtend a relatively low grazing angle with respect to the horizontal plane. With a conventional linear system, the envelope of minor lobes invariably insonifies the sea surface directly above the device. The surface movement then generates a Doppler signal that appears as a masking noise in the receiver and severely limits the system's remote-sensing capability. These difficulties appear to be greatly alleviated with the parametric beam since it has no side lobes (Figure 16).

Furthermore, the transit of a Doppler sonar platform in reverberant environments creates another noise signal, proportional to both the extent of the system's minor lobe distribution and to the width of its major beam. This noise, called own ship's Doppler, has long been a serious problem, one to which electronic signal processing is often applied in an effort to nullify the induced noise. Such schemes are not particularly successful, however, because each lobe sees a different Doppler frequency component. Thus the induced noise spectrum and isually quite broad and frequently swamps the Doppler component of the target. The sharp, unidirectional beam of the parametric array gets around this problem because it sees only a narrowband of own Doppler noise, which is easy to nullify with electronic signal processing.

Finally, the parametric Doppler technique has the unique advantage of being able to transmit while it is receiving. Since the difference frequency radiation is actually generated by two high-frequency sounds that interact fairly far away from the electroacoustic source, the acoustic "sing around" between source and receiver is greatly reduced. For many measurement problems, the parametric signals may be transmitted continuously, with no fear of overloading the receiver with a high-intensity transmission at the carrier frequency.

Although it may seem strange at first sight, wideband Doppler techniques show promise for applications in environments subject to high Doppler reverberation [79]. A theoretical paper by Fenlon [80] on the spectra of parametric arrays appears useful in this regard.

The application of parametric techniques to Doppler measurements is just starting to be developed. Analyses (Bucker, private communications, 1976) and experiments [81] are beginning to appear. The future will undoubtedly see greater emphasis on this important aspect of underwater sound.

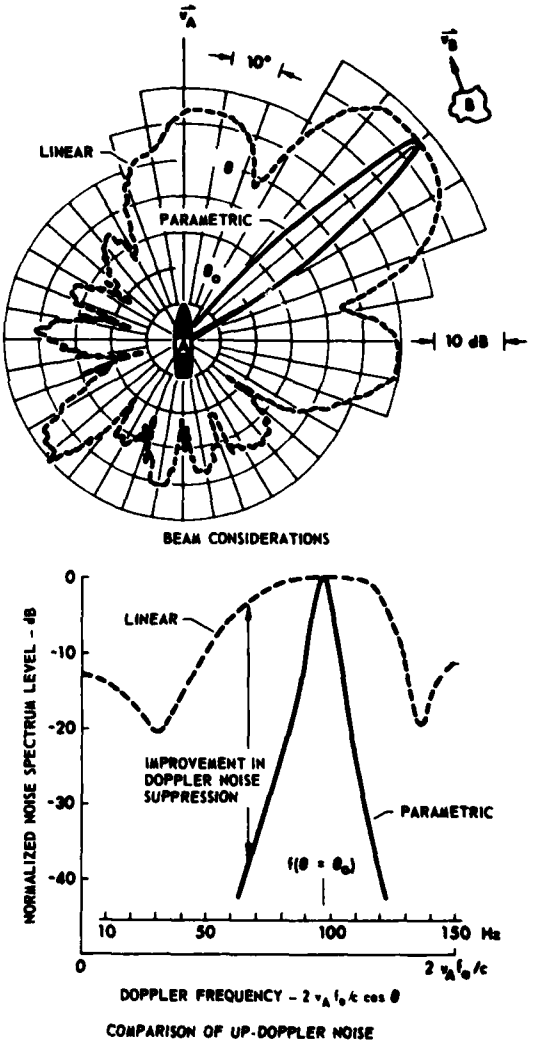


Figure 16—Doppler measurements. Parametric systems, with their narrow, sidelobe-free beams, are much more immune to noise generated by relative motion of clutter in the field of view.

EFFICIENCY AND FINITE AMPLITUDE ATTENUATION IN PARAMETRIC ARRAYS

The loss of energy in the primary-to-difference frequency conversion process has long been a major stumbling block in parametric array applications. Most observers sooner or later give some consideration to increasing the efficiency through the insertion of a more nonlinear fluid in the interaction volume. Experiments have actually been done along these lines with remarkable success [82]. However, the practical implementation of this approach is currently subject to a few shortcomings that need some clarification.

First, the efficiency of an ordinary, unsaturated parametric array is proportional to the power input to the interaction volume. By increasing the power, one increases the efficiency, until shocks begin to form in the multifrequency primary wave. At this point, the shock front dissipation effectively reduces the amount of power available for parametric interaction. The primary wave eventually becomes limited by the saturation phenomenon. During this process, the difference frequency amplitude becomes dependent on the square root of the input power, and the parametric beam begins to broaden (See Fig. 17).

In this region a curious situation develops; the nonlinearity of the medium drops out of the picture to the extent that the difference frequency pressure no longer depends on the specific value of this parameter. What really happens is that two

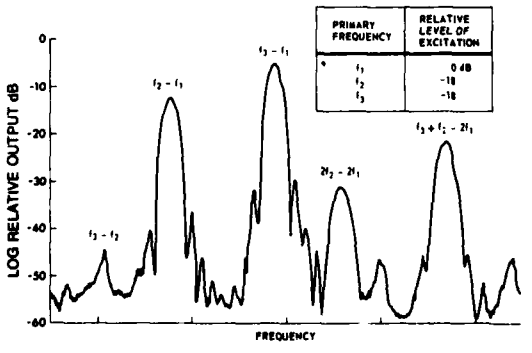


Figure 17—Difference frequency spectrum for multitone excitation. The manifest of frequency components gets complicated when several primary radiations are transmitted. This plot, from NUSC, shows that the harmonics of the primaries are active in creating harmonics of the difference frequency. [83]

competing nonlinear mechanisms (parametric interaction and acoustic saturation) become diametrically opposed to each other, allowing the parameter of nonlinearity to cancel itself out of the mathematical solution to the problem. Some nonlinearity is still required, of course, or we would not have developed either of these two effects in the first place; however, in this regime, it no longer matters whether the nonlinearity is high or low.

What then can be done in the way of optimizing parametric array interaction from the standpoint of medium characteristics? Bartram's [84] work shows that the only medium parameter having any consequence in the problem is sound velocity constant c_o, which appears in the denominator of his solution. It may be quite possible then to increase the parametric efficiency by going to a fluid having a slow sound velocity.

Of the relatively few liquids whose sound velocity and nonlinearity have been tabulated, liquid nitrogen has the slowest velocity (869 m/s). However, this would probably be a rather cantankerous fluid to use in underwater acoustics, not only because the long interaction volume would have to be encased in a Dewar flask at sea but also because of the tendency of this fluid to change state. In their measurements of the nonlinearity of liquid nitrogen, Hsiu-fen et al. [85] reported that ". . . bubbles of gaseous nitrogen tended to accumulate in the medium and settle on the receiver, causing fluctuations in the amplitude of the second harmonic pulses."

In the final analysis, no fluid has yet been found to increase the efficiency of the parametric array at high intensities. Even if such a fluid were located, the problem remains of positioning it in a long tube encompassing the interaction volume. The use of such a tube runs counter to the compact size advantage of the original concept of parametric arrays. Certain configurations may be feasible, however, depending on the outcome of future research.

ABSORPTION OF SOUND BY SOUND

A unique configuration of the parametric array involves the use of one wave to eliminate another. In this problem, a high-frequency wave of low

amplitude is acted on by a low-frequency wave inserted in the medium by an intense source. At the origin, the two waves (which are still linear entities) undergo a linear combination, with the result that the high-frequency wave appears as a modulation superimposed on the low-frequency waveform. With propagation, however, the combined waveform goes into shock due to the large intensity of the low-frequency component. When this happens, the high-frequency oscillations are forced to "crawl up" the sawtooth in the compressional phase and "slide down" it in the rarefactional phase. They are eventually compacted toward the shock fronts, where they are converted to heat by the nonlinearly induced dissipation in those regions. Figure 18 illustrates this effect.

Westervelt [87] is credited with originating the idea of absorbing sound with sound in another ONR-sponsored work, conducted by Schaffer and Blackstock [88]. Their experiments, as well as those of Moffett et al. [89] show that a high-frequency sound can be attenuated by a low-frequency sound, but not vice versa. This phenomenon apparently eliminates a broad category of applications for this effect in the quieting of ships, machinery, and industrial processes because the low-frequency absorber wave appears to be even more obnoxious than the one it is desired to eliminate.

On the other hand, the fundamental mechanism of sound absorbed by sound is undoubtedly at work, though unrecognized, in many important acoustical processes already confronting us today. What role does this mechanism play, for example, in jet and screw noise abatement? Could this effect be significant in weighting the frequency distributions of acoustic absorption data acquired with the explosive shot technique? Does the low-frequency ambient noise in the ocean (which increases in intensity with decrease in frequency) play a perceptible role in damping the upper regions of the noise spectrum? These and other basic questions remain to be answered with future research in this area. The theoretical tools for these investigations are beginning to be developed. Besides Westervelt's work, the studies of Pridham [90] and Krasil'nikov et al. [91] should be expected to be useful in future analyses.

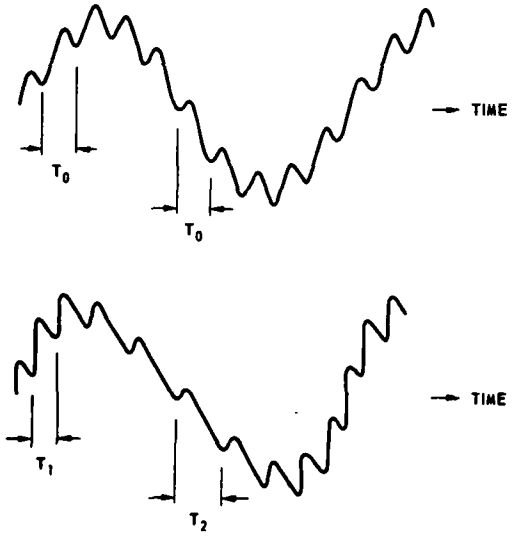


Figure 18—Absorbtion of sound by sound. The superposition of an intense low-frequency wave on a high-frequency beam causes the latter to be attenuated. Finite-amplitude effects force the short wavelengths "over the top and under the bottom" as the long wavelength goes into shock. [86]

BUBBLE ENHANCED PARAMETRIC SOURCES

One of the newest, most interesting problems in parametric arrays involves the use of microbubbles in the interaction volume. This use greatly increases the nonlinearity in a small region and enables some interesting nonlinear mechanisms to be studied. Almost all of these occur in a state of extreme nonlinearity, cavitation and saturation as well as dispersion.

At least three mechanisms are effective in generating sound with bubbles. The first involves the collapse of a cavitation bubble under pressure and the wideband noise pulse that accompanies that occurrence [92]. A second mechanism is the nonlinear oscillation of a bubble in a sound field [93], which involves both the nonlinearity of the gas in the bubble and the dynamic nonlinearity of the bubble structure. A third mechanism is the periodic generation and depletion of bubbles, with the attendant sound-producing expansion and contraction of the bubble volume. See Fig. 19.

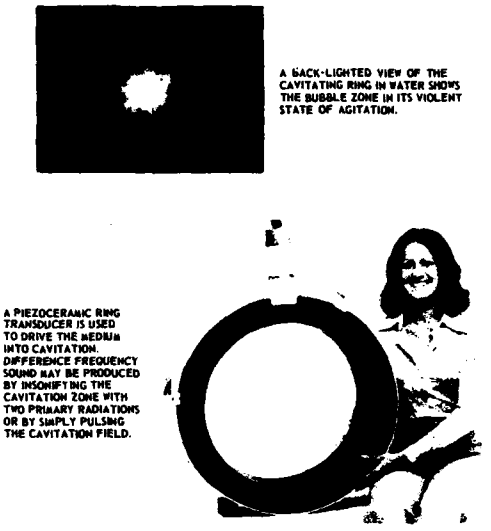


Figure 19—Cavitation-enhanced sound source. Although usually detrimental, cavitation bubbles greatly increase the medium nonlinearity, enabling parametric interaction to produce intense low-frequency sound at high efficiencies.

The second mechanism was used in the experiments of Dunn et al. [94] at the University of Birmingham. Their work served to intrigue many modern researchers by demonstrating enhanced generation of sum, difference, and harmonic components as a result of parametric interaction.

The presence of the bubbles also increases the attenuation and scattering of waves passing through them, so that the cumulative effects usually associated with parametric interaction over a long, end-fire array are no longer prevalent. This attenuation and scattering of course destroys the high directivity of the resultant radiation.

On the other hand, these radiations are generated at high efficiencies. In somewhat of an understatement, Zabolotskaya and Soluyan [71] observed that "this effect has practical advantages for the emission of a low frequency wave." The ability to generate intense, low frequency radiations is indeed a prime motivation behind basic research in this area.

Several interesting variations on this theme have also been reported. By insonifying a thin plane of bubbles, Lockwood [95] has developed a geometry for interaction that retains some directivity. Here, the insonified patch acts as a planar source of difference frequency sound, which can have an aperture large enough to develop a narrow sound beam.

The third mechanism is used in yet another approach, taken by Clynch and Thompson (private communications, 1976). Here a ring transducer is used to insonify a focal region with a train of large amplitude cw pulses. These pulses create their own bubble field, which periodically expands and contracts the entire focal region at the pulse repetition frequency. Extremely low frequency sounds are produced at high intensities.

The myriad of physical effects that occur in any of these processes, coupled with the relative newness of these problems, have limited their scope to purely basic investigations. Although much research remains to be done, it is not unreasonable to imagine that bubble interactions may some day provide us with efficient wideband sound sources at those difficult frequencies in the infrasonic to low audio range (i.e., 1 to 100 Hz). Before such developments can occur, however, many more experiments must be done with supporting theory to better delineate all the crucial mechanisms and phenome... resulting from their combination.

INVESTIGATIONS IN OTHER MEDIA

Although underwater acoustics has been a major focal point for nonlinear investigations, many of the phenomena there evolved have been applied to other media.

The parametric array, for example, works in air, as has been verified by Bennett and Blackstock [96]. Finite amplitude effects are also being used to study and improve acoustic techniques for removing particulate pollutants from industrial smokestacks [97]. Another emerging possibility involves the use of nonlinear surface waves in solids and bulk waves in liquids to perform replica correlation for high-speed signal processing in sonar, radar, and radio communications. This technique uses a finite-amplitude acoustic replica of the transmitted signal interacting with a time-reversed transformation of the received signal traveling in the opposite direction. Crosscorrelation is expected on some segment of the medium where the two pulses overlap [98, 99]. These and many other interesting efforts may benefit from

and complement nonlinear research in underwater acoustics (Figs. 20, 21).



Figure 20—Parametric "TOPS" array installed on U.S.S. Dolphin *(AGSS 555), an R & D submarine. Here configured as a side-scanning sonar, the TOPS is a high-powered (80-kW) research tool.*

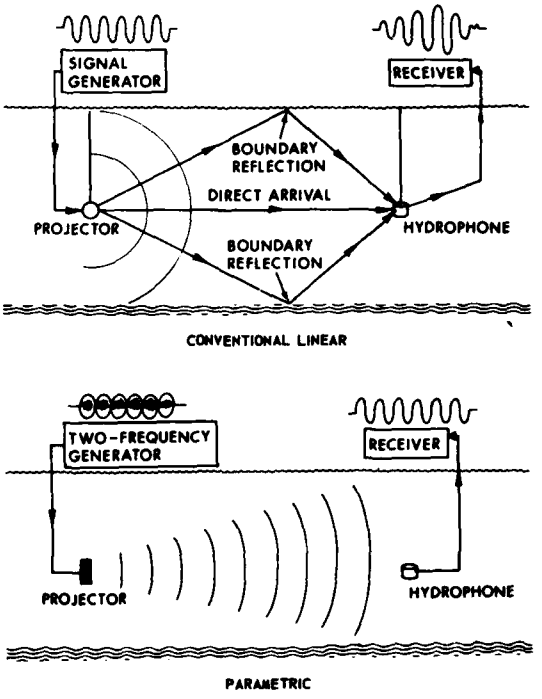


Figure 21—Sonar calibration. The sketches compare calibration techniques in confined waters by showing how the parametric system circumvents multipath reflection problems. [74]

PERSPECTIVE

All of the topics considered in these pages have two aspects in common: (a) they have been researched at least enough for us to talk about them, and (b) they probably have a future, some as naval applications and others as beneficial exercises for understanding new physics and for the conduct of related research.

What can be said about the expected topics of the future, those for which we do not yet have an adequate understanding? This question is of course very difficult to answer since the descriptions have also not been developed. In an effort to categorize the types of developments and discoveries one might expect from future research we can, however, examine the achievements, the methods, and the trends of research in progress.

Since the truly significant discoveries really cannot be planned, programed, or scheduled, they are the most difficult to anticipate. About all that can be done with respect to the breakthroughs is to encourage them. The Office of Naval Research has operated on this premise for many years, believing that the choice of a particular scientist or laboratory is often more important than the initial research topic. It is nonetheless likely that discoveries in nonlinear acoustics comparable to that of the parametric array may again be made. These discoveries, of course, are more likely if they are encouraged.

Historically, it has been the theoretical community that has made the memorable breakthrough. However, theory today includes the disciplines of computer modeling and data reduction, where the likelihood of a discovery is somewhat more remote. Ironically, the truly theoretical segment of the nonlinear acoustics community (at least in the West) has in recent years received a decreasing share of the encouragement. Emphasis is now on experimentation and the search for applications. These endeavors often uncover potential discoveries, especially in acoustics where one does not need a giant cyclotron or a huge telescope to make fundamental measurements. But in the long run, it is always left to the theorist to explain the puzzles uncovered by inquisitive experimentation.

With regard to the trends in future nonlinear acoustics research, one can extrapolate from the

existing pools of knowledge with various degrees of assurance.

In the near future, we must surely see the extension of a research theme, perhaps best characterized as the environmental aspect. The nonlinear acoustic entities laid down by Stokes, Earnshaw, Fay, Fubini, and Westervelt are adequately understood only for ideal, well-behaved media. A significant portion of the research effort is therefore being directed at the influence of real media effects. Although there is quite a bit of overlap, several topics can be associated with each of the major centers or schools of research, as inferred from their open literature publications. Thus, the story pertaining to random inhomogeneities, turbulence, and their effect on parametric arrays should be expected to be written in England. The Soviet school is heavily involved with the theory of dispersion and diffraction, which is pertinent to parametric interactions in bubbles and solids. Besides the bubble problem, the interest in America includes propagation effects influenced by boundaries, including the surface and bottom of the sea and the thermal layers within it.

Basic nonlinear acoustics problems of general interest in most of these centers appear to include the interaction of noise with itself and with other radiations, interactions in crystals and other solids, better theoretical descriptions in existing and sometimes poorly explained phenomena, the extension of problems treated in underwater sound to air acoustics, and the reflection, refraction, and scattering of finite-amplitude waves.

Perhaps the largest single research trend in nonlinear acoustics is the race for practical applications. Although many of these applications are mentioned in the text, their complete description is far beyond the scope of this work. This race is exciting, not only because of its military and economic impact but also because of its utility in producing tools for research and development in other fields. Subbottom profiling of the sediments and structures of the Earth with parametric sonar will undoubtedly continue to be developed. The ultimate questions are to what depth of penetration it will be limited and how it will best be used in mining and exploration. Also emerging are applications to fundamental biomedical measurements on the acoustics properties of tissue. Communication with divers, relocation of equipment, and other aspects of offshore petroleum operations appear in many cases to be well suited to nonlinear sonar techniques. Marine archeologists involved in historical as well as prehistoric site surveys could well use high-resolution parametric systems in the search for small artifacts and ancient habitats. Finally, a wide category of naval applications of nonlinear acoustic devices show great potential for realization. These applications include Doppler navigation, communications, submarine detection and surveillance, mine countermeasures, homing systems, calibration procedures, and many others.

Over the past decade, nonlinear acoustics has expanded to include a remarkable and respectable number of interesting and important problems. The next 10 years will undoubtedly see many of these problems brought to fruition, and many others will surely be discovered and delineated.

ACKNOWLEDGMENTS

The author is indebted to D. T. Blackstock and F. H. Fenlon for their comments on the manuscript. M. B. Moffett is thanked for his assistance in providing illustrations from the Naval Underwater Systems Center.

REFERENCES

1. L. Euler, *Mém. Acad. Sci. Berlin* **II**, 274–315. (1755).
2. G. G. Stokes, *Philos. Mag* **33**, 349–356 (1848).
3. R. D. Fay, *J. Acoust. Soc. Amer.* **3**, 222–241 (1931).
4. E. Fubini, *Alta Frequenza* **4**, 530–581 (1935).
5. A. L. Thuras, R. T. Jenkins, and H. T. O'Neill, *J. Acoust. Soc. Amer.* **6**, 173–180 (1935).
6. C. Eckart, *Phys. Rev.* **73**, 68 (1948).
7. M. O. Lighthill, *Proc. R. Soc. Lond.*, **A222**, 1–32 (1954); **A211**, 564–587 (1952).

8. R. T. Beyer, *J. Acoust. Soc. Amer.* **32**, 719-721 (1960).
9. F. E. Fox and W. A. Wallace, *J. Acoust. Soc. Amer.* **26**, 994–1006 (1954).
10. P. J. Westervelt, *J. Acoust. Soc. Amer.* **35**, 535–537 (1963). Presented at the 59th A.S.A. Meeting, 1960.
11. H. O. Berktay, *J. Sound Vib.* **2**, 435–461 (1965).
12. H. O. Berktay, *J. Sound Vib.* **2**, 462–470 (1965).
13. H. O. Berktay, *J. Sound Vib.* **5**, 155–163 (1967).
14. H. O. Berktay, *J. Sound Vib.* **6**, 244–254 (1967).
15. H. O. Berktay, *J. Sound Vib.* **6**, 268–269 (1967).
16. D. T. Blackstock, in *Nonlinear Acoustics, Proceedings of the 1969 ARL Symposium*, T. G. Muir, ed., University of Texas at Austin, Applied Research Laboratories, 1970.
17. L. Bjørnø, in *Ultrasonics International 1975 Symposium Proceedings*, Imperial College, London (IPC House, London, 1975).
18. Lagrange, *Oeuvres de Lagrange*, Vol. 1, Ganthier-Villars, Paris, 1867.
19. S. D. Poisson, *J. L'Ecole Polytech.* (Paris) **1**, 364–370 (1808).
20. S. Earnshaw, *Trans. R. Soc. Lond.* **150**, 133–148 (1860).
21. B. Riemann, *Abhandl. Ges. Wiss., Göttingen, Math-Physik-Kl.* **8**, 43–65 (1859-59).
22. Lord Rayleigh, *Proc. R. Soc. Lond.* A **84**, 247–284 (1910).
23. P. Biquard, *Ann. Phys.* (Paris), ser. 11, **6**, 195–304 (1936).
24. L. K. Zarembo and V. A. Krasil'nikov, *Sov. Phys. Acoust.* **2**(68), 580–559 (1959).
25. R. T. Beyer, *J. Acoust. Soc. Amer.* **32**, 719–721 (1960).
26. I. G. Mikhailov and V. A. Shutilov, *Sov. Phys. Acoust.* **10**, 385–389 (1965).
27. I. Rudnick, *J. Acoust. Soc. Amer.* **30**, 564–567 (1958).
28. C. E. Hargrove and K. Achyuthan, in *Physical Acoustics*, Vol. IIB, W. P. Mason, ed., Academic Press, New York, 1965.
29. W. J. M. Rankine, *Philos. Trans. R. Soc. Lond.* **160**, 277–288 (1870).
30. G. I. Taylor, *Proc. R. Soc. Lond. A* **84**, 371–377 (1910).
31. J. S. Mendouese, *J. Acoust. Soc. Amer.* **25**, 51–54 (1953).
32. D. M. Towle and R. B. Lindsay, *J. Acoust. Soc. Amer.* **27**, 530–533 (1955).
33. V. Narasimhan and R. T. Beyer, *J. Acoust. Soc. Amer.* **28**, 1233–1236 (1956).
34. B. D. Cook, *J. Acoust. Soc. Amer.* **34**, 941–946 (1962).
35. D. T. Blackstock, *J. Acoust. Soc. Amer.* **36**, 534–542 (1964).
36. R. P. Barnes and R. T. Beyer, *J. Acoust. Soc. Amer.* **36**, 1371–1377 (1964).
37. W. W. Lester, *J. Acoust. Soc. Amer.* **34**, 1991(A) (1962); **40**, 847–851 (1966).
38. J. A. Shooter, T. G. Muir, and D. T. Blackstock, *J. Acoust. Soc. Amer.* **55**, 54–62 (1974).
39. E. V. Romanenko, *Sov. Phys. Acoust.* **5**, 100–104 (1959).
40. D. T. Blackstock, *J. Acoust. Soc. Amer.* **39**, 1019–1026 (1966).
41. R. P. Ryan, Q. G. Lutsch, and R. T. Beyer, *J. Acoust. Soc. Amer.* **34**, 31–35 (1962).
42. H. W. Marsh, in *Application of Finite-Amplitude Acoustics to Underwater Sounds, Proceedings of the 1968 Symposium*, a seminar at Navy Underwater Sound Laboratory, May 1968, R. H. Mellen, ed., Naval Underwater Sound Laboratory Rep. No. 1084, 1970.
43. T. G. Muir, *Physics in Sound in Maritime Sediments*, L. D. Hampton, ed., Plenum Press, New York, 1974.
44. R. K. Gould et al., *J. Acoust. Soc. Amer.* **40**, 421–427 (1966).
45. J. C. Lockwood, T. G. Muir, and D. T. Blackstock, *J. Acoust. Soc. Amer.* **53**, 1148–1153 (1973).
46. L. E. Hargrove and E. A. Hiedemann, *J. Acoust. Soc. Amer.* **33**, 1747–1749 (1961).
47. M. A. Breazale and E. A. Hiedemann, *J. Acoust. Soc. Amer.* **33**, 700–701 (1961).
48. A. L. Van Buren and M. A. Breazeale, *J. Acoust. Soc. Amer.* **44**, 1014–1020 (1968).
49. A. L. Van Buren and M. A. Breazeale, *J. Acoust. Soc. Amer.* **44**, 1021–1027 (1968).
50. V. M. Albers, *Underwater Sound, Benchmark Papers in Acoustics*, p. 415, Dowden, Hutchinson, and Ross, Inc. Stroudsburg, Pa., 1972.
51. H. Lamb, *The Dynamical Theory of Sound*, 2d ed., p. 183, Dover Publishers, Inc., New York, 1925.
52. O. L. S. Bellin and R. T. Beyer, *J. Acoust. Soc. Amer.* **34**, 1051–1054 (1962). Presented at the 59th A.S.A. Meeting, 1960.
53. V. Lauvstad and S. Tjötta, *J. Acoust. Soc. Amer.* **35**, 929–930 (1963).
54. V. Lauvstad, J. Naze, and S. Tjötta, *Acta Universitatis Bergensis Series Mathematica*, No. 12, 1–24 (1964).
55. D. G. Tucker, *J. Sound Vib.* **2**, 429–434 (1965).
56. H. O. Berktay and C. A. Al-Temimi, *J. Sound Vib.* **9**, 295–307 (1969).
57. H. Hobaek, *J. Sound Vib.* **6**, 460–463 (1967).
58. S. Tjötta, *J. Sound Vib.* **6**, 270 (1967).
59. V. A. Zverev, A. I. Kalachev, and N. S. Stepanov, *Sov. Phys. Acoust.* **13**, 324–326 (1968).
60. H. O. Berktay, in *Application of Finite-Amplitude Acoustics to Underwater Sound*, Proceedings of

the 1968 Symposium, R. H. Mellen, ed.; Naval Underwater Sound Laboratory Rep. No. 1084, 1970.
61. T. G. Muir and J. E. Blue, *J. Acoust. Soc. Amer.* **46**, 227–232 (1969).
62. V. A. Zverev and A. I. Kalachev, *Sov. Phys. Acoust*, **14**, 173 (1968).
63. T. G. Muir, ed., *Nonlinear Acoustics, Proceedings of the 1969 Symposium at Applied Research Laboratories, The University of Texas at Austin* (1970).
64. M. B. Moffett, P. J. Westervelt, and R. T. Beyer, *J. Acoust. Soc. Amer.* **49**, 339–343 (1970).
65. R. H. Mellen and D. G. Browning, *J. Acoust. Soc. Amer.* **49**(3), 932–935 (1970).
66. G. M. Walsh, in *Proceedings of the British Acoustics Society Specialist Meeting on Nonlinear Acoustics* (1971), British Acoustical Society, London, 1971.
67. V. A. Zverev and A. I. Kalachev, *Sov. Phys. Acoust.* **16**, 204–208 (1970).
68. Barnard et al., *J. Acoust. Soc. Amer.* **52**, 1437–1441 (1972).
69. H. O. Berktay and T. G. Muir, *J. Acoust. Soc. Amer.* **53**, 1377–1383 (1973).
70. J. J. Truchard, *J. Acoust. Soc. Amer.* **58**, 1141–1150 (1975); **59**, 528 (1976).
71. E. A. Zabolotskaya and S. I. Soluyan, *Sov. Phys. Acoust.* **18**, 396–398 (1973).
72. R. H. Nichols, Jr., *J. Acoust. Soc. Amer.* **50**, 1086–1087 (1971).
73. C. C. Fox and O. L. Akervold, *J. Acoust. Soc. Amer.* **53**, 382(A) (1973).
74. W. L. Konrad and J. G. Navin, Naval Underwater Systems Center Tech. Rep. No. 4645, 1974.
75. P. Pettersen et al., in *Proceedings of the Joint Conference on "Instrumentation in Oceanography,"* University of North Wales, Bangor (1975).
76. L. A. Thompson and T. G. Muir, *J. Acoust. Soc. Amer.* **55**, 429(A) (1974).
77. J. Jarzynski and L. Flax, *J. Acoust. Soc. Amer.* **59**, S29 (1976).
78. T. G. Muir and J. R. Clynch, in *Recent Developments in Underwater Acoustics*, British Institute of Acoustics Conference Proceedings, British Institute of Acoustics, London, 1976.
79. R. Laval, in *Sound Propagation in Shallow Water*, Vol. II, pp. 235–246, O. L. Hastrup and O. V. Olesen, eds., SACLANTCEN Conference Proceedings CP-14, SACLANT ASW Research Centre, La Spezia, Italy, 1974.
80. F. H. Fenlon, *J. Acoust. Soc. Amer.* **53**, 1752–1754 (1973).
81. W. I. Roderick, in *Recent Developments in Underwater Acoustics*, British Institute of Acoustics Conference Proceedings, British Institute of Acoustics, London, 1976.
82. L. Bjørnø, B. Chrisoffersen, and M. P. Schreiber, *Acustica*, **35**, 99–106 (1976).
83. W. L. Konrad, Naval Underwater Systems Center, Tech. Rep. No. 5227, 1975.
84. J. F. Bartram, *J. Acoust. Soc. Amer.* **52**, 1042–1044(L) (1972).
85. K. Hsiu-fen, L. K. Zarembo, and V. A. Krasil'nikov, *Sov. Phys. Acoust.* **9**, 306–307 (1963).
86. W. L. Konrad, M. B. Moffett, and L. F. Carlton, Naval Underwater Systems Center Tech. Memo. No. TD124-92-75, 1975.
87. P. J. Westervelt, *J. Acoust. Soc. Amer.* **59**, 760–764 (1976).
88. M. E. Schaffer and D. T. Blackstock, *J. Acoust. Soc. Amer.* **57**, S73 (1975).
89. M. B. Moffett, W. L. Konrad, and A. T. Corcella, NUSC Tech. Memo. TDIX-C16-73, 1973.
90. R. G. Pridham, *J. Acoust. Soc. Amer.* **55**, 550 (1974).
91. V A. Krasil'nikov, O. V. Rudenko, and A. S. Chirkin, *Sov. Phys. Acoust.* **21**, 80–81 (1975).
92. H. G. Flynn, *J. Acoust. Soc. Amer.* **58**, 1160–1170 (1976).
93. W. Lauterborn, *J. Acoust. Soc. Amer.* **59**, 283–293 (1976).
94. D. J. Dunn, M. Kuljis, and V. G. Welsby, *J. Sound Vib.* **2**, 471–476 (1965).
95. J. C. Lockwood and D. P. Smith, *J. Acoust. Soc. Amer.* **57**, S73 (1975).
96. M. B. Bennett and D. T. Blackstock, *J. Acoust. Soc. Amer.* **57**, 562–568 (1975).
97. D. S. Scott, *J. Sound Vib.* **43**, 607–619 (1975).
98. P. Das and D. T. Blackstock, *J. Acoust. Soc. Amer.* **54**, 134–135 (1973).
99. V. A. Krasil'nikov and V. E. Lyamor, *Sov. Phys. Acoust.* **19**, 516–517 (1974).

* U. S. GOVERNMENT PRINTING OFFICE : 1977 O - 242-122