AFGWC/TN-82/002

(1)

# SOURCE BOOK OF THE SOLAR-GEOPHYSICAL ENVIRONMENT

AD A138682

By

MAJ RAY E. TOWNSEND
CAPT RICHARD W. CANNATA
CAPT ROBERT D. PROCHASKA
CAPT GARY E. RATTRAY
1LT JOHN C. HOLBROOK

DTIC FILE COPY

APPROVED FOR PUBLIC RELEASE:DISTRIBUTION UNLIMITED

AUGUST 1982

UNITED STATES AIR FORCE
AIR WEATHER SERVICE (MAC)
AIR FORCE GLOBAL WEATHER CENTRAL
OFFUTT AFB NE 68113

DTIC
ELECTE
FEB 17 1984
E

84 02 24 043

REVIEW AND APPROVAL STATEMENT

This publication approved for public release. There is no objection to unlimited distribution of this document to the public at large, or by the Defense Technical Informatiion Center (DTIC) or to the National Technical Information Service (NTIS).

This technical publication has been reviewed and is approved for publication.

Charles W. Cook

CHARLES W. COOK, GM-13
Reviewing Official

FOR THE COMMANDER

James Kerlin

JAMES KERLIN, Colonel, USAF
Chief, Technical Services Division

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER: AFGWC/TN-82/002 | 2. GOVT ACCESSION NO.: AD-A138 682 |
| 3. RECIPIENT'S CATALOG NUMBER: | |
| 4. TITLE *(and Subtitle)*: Source Book of the Solar-Geophysical Environment | 5. TYPE OF REPORT & PERIOD COVERED: |
| | 6. PERFORMING ORG. REPORT NUMBER: |
| 7. AUTHOR(s): Maj Ray E. Townsend; Capt Richard W. Cannata; Capt Robert D. Prochaska; Capt Gary E. Rattray; 1Lt John C. Holbrook | 8. CONTRACT OR GRANT NUMBER(s): |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: HQ Air Force Global Weather Central (MAC), Offutt AFB, Nebraska 68113 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Global Weather Central (MAC), Offutt AFB, Nebraska 68113 | 12. REPORT DATE: August 1982 |
| | 13. NUMBER OF PAGES: 353 + xvi |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: HQ Air Force Global Weather Central (MAC), Offutt AFB, Nebraska 68113 | 15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A |

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

N/A

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

| | |
|---|---|
| Solar-Geophysical | Ionosphere |
| Solar System | Radiowave Propagation |
| Sunspot | Spacecraft Charging |
| Solar Flares | Space Radiation |
| Magnetosphere | Solar-Terrestrial |

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This technical note is intended as a text for the Space Environmental Support System (SESS) analyst and as a reference for users of SESS products. The early chapters on physics, coordinate systems, and astronomy are intended to provide a brief review of topics of importance. The bulk of the text deals with the solar-terrestrial system, with the choice of topics driven by operational considerations. This material is intended to complement rather than replace current textbooks and journals dealing with solar-geophysics and astronomy.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

## PREFACE

The Source Book has evolved from a Plan of Instruction published by Captain Robert D. Prochaska in 1979. It and the second edition of this text, which appeared in 1981, were intended as internal documents of the Space Environmental Support Branch, Air Force Global Weather Central. The third edition appeared early in 1982. The primary purpose of this text is to provide a single, composite reference for personnel assigned to the Space Environmental Support System. It also functions as the textbook for several formal courses taught by this unit for users and forecaster/analysts working in the field of solar-geophysics. In this mode, it is intended for use with the Space Science Workbook, AFGWC/FM-82/002.

This edition provides both refinement and expansion of the previous edition. These changes are the result of suggestions by analysts, students, and forecasters who reviewed the previous text and the not inconsiderable expansion of SESS operational involvement. A new section on ionospheric models is the most significant change. All the material has, however, undergone some revision. No attempt has been made to provide derivations or proofs of included theories. Similarly, the text is not intended as a complete review of existing theories. Operational utility was the primary basis for inclusion.

As the material included has expanded, so too have the number of people involved in its production. Captain Gary Rattray, Captain Richard W. Cannata, and Lieutenant John C. Holbrook wrote the sections on ionospheric models, the interplanetary medium, manned spacecraft operations, and long-term climatological effects. Previous students, team chiefs, and forecasters gave freely of their time to review and comment on various portions of the manuscript. Particular thanks are due Col Kenneth German and Lt Col George Wortham for their review of this text and many helpful suggestions. The ultimate challenge of converting the seemingly endless pages of writing into the finished, print-ready text was managed with an ever present smile by the Word Processing Center (AFGWC/DAW). To these and many others, I owe many thanks.

RAY E. TOWNSEND, Major, USAF
Offutt AFB, NE 68113
August 1982

Accession For
NTIS GRA&I [X]
DTIC TAB [ ]
Unannounced [ ]
Justification
By
Distribution/
Availability Codes
Avail and/or
Dist Special
A-1

DTIC COPY INSPECTED 2

SOURCE BOOK OF THE SOLAR-GEOPHYSICAL ENVIRONMENT

TABLE OF CONTENTS


| | PAGE |
|---|---|
| List of Illustrations | viii |
| List of Tables | xv |
| Introduction | xvi |
| CHAPTER 1 - History and Operations | 1 |
| 1.1 SESS Mission | 1 |
| 1.2 Historical Milestones | 1 |
| 1.3 The Early Years | 3 |
| 1.4 Transition to AFGWC | 4 |
| 1.5 Evaluation of SESS at AFGWC | 5 |
| 1.6 Development of SESS Software | 7 |
| 1.6.1 Data Base | 7 |
| 1.6.2 Display Software | 7 |
| 1.6.3 Forecast Software | 8 |
| 1.7 Ionospheric Models | 9 |
| 1.8 AWS/NOAA Interaction | 9 |
| 1.9 SESS Organization | 10 |
| CHAPTER 2 - Background Physics | 11 |
| 2.1 Atomic Structure | 11 |
| 2.1.1 Atomic Energy Storage | 11 |
| 2.1.2 Absorption and Emission | 12 |
| 2.1.3 Temperature | 13 |
| 2.2 Electromagnetic Radiation | 14 |
| 2.2.1 Fields | 15 |
| 2.2.2 Waves | 17 |
| 2.3 Radiation and Matter | 25 |
| 2.3.1 Plasma Frequency Effects | 27 |
| 2.3.2 Polarization Changes | 29 |
| 2.4 Radiation Analysis | 30 |
| 2.4.1 Blackbody Analysis | 30 |
| 2.4.2 Spectral Analysis | 31 |
| 2.4.3 The Zeeman Effect | 32 |
| 2.4.4 Doppler Effect | 32 |
| 2.5 Summary | 32 |
| CHAPTER 3 - Astronomical Instrumentation | 35 |
| 3.1 Characteristics of Telescopes | 35 |
| 3.2 Types of Telescopes and Seeing | 36 |
| 3.3 Optical Instrumentation | 36 |
| 3.4 SOON Instrumentation | 37 |
| 3.5 Radio Telescope and Resolution | 38 |
| 3.6 Summary | 39 |

CHAPTER 4 - Space ..... 40
4.1 The Solar System ..... 40
4.2 The Terrestrial Planets ..... 41
4.3 The Jovian Planets ..... 42
4.4 Interplanetary Matter ..... 43
4.5 Summary ..... 45

CHAPTER 5 - Stellar Evolution ..... 46
5.1 Stellar Properties ..... 46
5.2 The Hertzsprung-Russell Diagram ..... 47
5.3 Stellar Energy Sources ..... 48
5.4 Star Birth ..... 49
5.5 Main Sequence Evolution ..... 50
5.6 Post Main Sequence Evolution ..... 50
5.7 Star Death ..... 51
5.8 Summary ..... 52

CHAPTER 6 - Positioning the Solar System ..... 53
6.1 Kepler's Laws ..... 53
6.2 Earth Motions ..... 55
6.3 Solar Motions ..... 56
6.4 Solar Coordinates ..... 57
6.5 Celestial Coordinates ..... 60
6.6 Summary ..... 62

CHAPTER 7 - The Quiet Sun ..... 63
7.1 Energy Transport ..... 63
7.2 Internal Structure ..... 65
7.2.1 The Solar Core ..... 65
7.2.2 Radiation Zone ..... 65
7.2.3 The Convection Zone ..... 66
7.2.4 The Photosphere ..... 67
7.3 The Solar Atmosphere ..... 68
7.3.1 Chromosphere ..... 68
7.3.2 Corona ..... 69
7.3.3 Coronal Holes ..... 70
7.4 Surface Velocity Features ..... 71
7.5 Summary ..... 72

CHAPTER 8 - The Active Sun ..... 73
8.1 Plage and Facula ..... 73
8.2 Sunspots ..... 74
8.2.1 Sunspot Groups ..... 74
8.2.2 Sunspot Classification ..... 75
8.2.3 Sunspot Variations ..... 81
8.2.4 Sunspot Model ..... 82
8.3 Prominences ..... 83
8.4 The Solar Cycle ..... 87
8.5 Radio Astronomy ..... 88
8.5.1 Basic Component ..... 92

8.5.2 The Slowly Varying Component..... 94
8.5.3 The Active Radio Sun..... 96
8.5.3.1 Noise Storms..... 97
8.5.3.2 Sweep Radio Bursts..... 97
8.5.3.3 Fixed Frequency Bursts..... 102
8.6 Solar Flares..... 105
8.6.1 Optical Flare Classification..... 105
8.6.2 X-ray Flare Classification..... 107
8.6.3 Flare Production..... 108
8.6.4 A Typical Flare-Producing Region..... 110
8.7 Flare Origin and Emission..... 112
8.8 Summary..... 113

CHAPTER 9 - The Interplanetary Medium..... 115
9.1 The Solar Wind..... 115
9.1.1 Solar Wind Plasma..... 115
9.1.2 Influence of the Magnetic Field..... 116
9.2 The Interplanetary Magnetic Field (IMF)..... 117
9.2.1 Formation of the IMF..... 117
9.2.2 Sector Structure..... 119
9.3 Flare Modification of the IP Medium..... 121
9.4 Long-Term Variation in the Solar Wind..... 124
9.5 IMF Analysis..... 128
9.6 Interplanetary Spacecraft..... 131
9.7 Summary..... 133

CHAPTER 10 - The Magnetosphere..... 134
10.1 Structure..... 134
10.2 Interaction with Interplanetary Space..... 138
10.3 Trapped Radiation Belts..... 140
10.3.1 Trapping Mechanisms..... 141
10.3.2 Proton Belts..... 144
10.3.3 Electron Belts..... 144
10.4 Main Earth Field..... 145
10.5 Secondary Geomagnetic Field..... 149
10.6 Geomagnetic Disturbances..... 149
10.6.1 Storm Phases..... 151
10.6.2 Types of Magnetic Storms..... 157
10.7 Geomagnetic Substorms..... 158
10.8 Geomagnetic Activity Indices..... 162
10.9 Geomagnetic Activity Forecasting..... 163
10.10 Aurora..... 169
10.10.1 Auroral Location..... 169
10.10.2 Auroral Appearance..... 170
10.11 Summary..... 173

CHAPTER 11 - The Ionosphere: Formation and Variation..... 174
11.1 The Neutral Atmosphere..... 174
11.1.1 Temperature Regimes..... 174
11.1.2 Chemical Composition Regimes..... 175
11.1.3 Density Variations..... 179

11.2 Ionospheric Formation..... 187
11.2.1 Chapman Layer Theory..... 188
11.2.2 Ionospheric Layers..... 191
11.3 Quiet Ionospheric Climatology..... 193
11.3.1 Latitudinal Ionospheric Regimes..... 193
11.3.1.1 The High Latitude Ionosphere..... 193
11.3.1.2 The Middle Latitude Ionosphere..... 196
11.3.1.3 Low Latitude Ionosphere..... 198
11.3.2 Ionospheric Height Regimes..... 200
11.3.2.1 Topside Ionosphere..... 201
11.3.2.2 Bottomside Ionosphere..... 203
11.3.2.2.a D Layer..... 203
11.3.2.2.b E Layer..... 205
11.3.2.2.c F Layer..... 206
11.4 Disturbed Ionospheric Variations..... 211
11.5 Eclipses and Meteors..... 214
11.6 Summary..... 215

CHAPTER 12 - Ionospheric Observations..... 217
12.1 Vertical Incidence Ionosonde..... 217
12.1.1 Operational Theory..... 217
12.1.2 Ionogram Interpretation..... 219
12.2 Oblique Ionosondes..... 223
12.2.1 Theory of Operation..... 225
12.2.2 Oblique Ionograms..... 227
12.3 Riometer..... 229
12.3.1 Theory of Operation..... 229
12.3.2 Equipment Design..... 230
12.4 Polarimeter..... 230
12.4.1 Dual Frequency Polarimeter..... 231
12.4.2 Faraday Rotation Polarimeter..... 231
12.4.3 Faraday Rotation Analysis..... 232
12.4.4 Polarimeter Calibration..... 235
12.5 Space-borne Instruments..... 240
12.6 Summary..... 241

CHAPTER 13 - Ionospheric Radiowave Propagation..... 242
13.1 Bands and Modes..... 242
13.1.1 Propagation Modes..... 242
13.1.2 Sky Wave Propagation..... 244
13.1.3 Skip Zone..... 246
13.2 Radiowave Transmission..... 246
13.2.1 Bandwidth..... 247
13.2.2 Interference..... 247
13.2.3 Signal and Noise..... 249
13.2.4 Fading and Absorption..... 251
13.3 The Long Waves - ELF, VLF, LF..... 252
13.3.1 Employment..... 252
13.3.2 Propagation Variability..... 253
13.3.3 Ionospheric Disturbances..... 253

13.4 Medium Frequencies - MF..... 254
13.4.1 MF Propagation Modes..... 255
13.4.2 Environmental Impact on MF..... 256
13.5 Shortwave Radio - the HF Band..... 256
13.5.1 HF Circuitry..... 257
13.5.2 Ionospheric Scattering..... 258
13.5.3 HF Climatology..... 261
13.5.3.1 HF Parameter Definitions..... 261
13.5.3.2 Diurnal HF Variations..... 263
13.5.3.3 Seasonal Effects..... 265
13.5.4 HF Anomalies..... 267
13.5.4.1 Geographic Anomalies..... 267
13.5.4.2 Structural Anomalies..... 268
13.5.5 Sudden Ionospheric Disturbances (SIDs)..... 271
13.5.6 Ionospheric Storms..... 272
13.5.6.1 HF Effects of a PCA..... 272
13.5.6.2 Geomagnetic Storms and HF Effects..... 273
13.5.6.2.a Negative Storms..... 273
13.5.6.2.b Positive Storms..... 277
13.5.7 HF Summary..... 277
13.6 Transionospherics - VHF, UHF, and SHF..... 279
13.6.1 Sky-Wave Propagation of VHF..... 279
13.6.2 Disturbance Effects on VHF Skywave..... 280
13.6.3 SATCOM and Scintillation..... 281
13.6.3.1 Scintillation Origin and Specification..... 281
13.6.3.2 Scintillation Climatology..... 282
13.6.4 Forecasting Scintillation..... 283
13.7 Summary..... 284

CHAPTER 14 - Ionospheric Modification and Modeling..... 286
14.1 Tests and Trapped Radiation..... 286
14.2 Electromagnetic Pulse Effects..... 287
14.3 Atmospheric Ionization Phemomena..... 288
14.4 Non-Nuclear Variations..... 289
14.5 Ionospheric Modeling..... 290
14.5.1 ITS-78..... 291
14.5.2 Polar Ionospheric Model..... 292
14.5.3 Four - Dimensional Ionospheric Model (4D)..... 293
14.5.4 The Ionospheric Scintillation Model..... 294
14.6 Summary..... 295

CHAPTER 15 - Spacecraft Operations..... 296
15.1 Satellite Charging..... 296
15.1.1 The Problem..... 296
15.1.2 Spacecraft Charging..... 297
15.1.3 Spacecraft Discharging..... 302
15.1.4 Forecasting..... 303
15.2 Spacecraft Drag..... 303
15.2.1 Heating Mechanisms..... 305
15.2.2 Spacecraft Impact..... 306

15.3 Radiation Effects on Manned Systems..... 310
15.3.1 Space Radiation..... 310
15.3.2 Space Radiation Environment..... 311
15.3.2.1 Galactic Cosmic Rays..... 311
15.3.2.2 Trapped Radiation Belt Structure..... 311
15.3.2.3 Solar Flare Particles..... 313
15.3.4 Radiation Hazards..... 314
15.3.4.1 Dosimetry..... 314
15.3.4.2 Biological Effects..... 316
15.3.5 Space Medicine..... 318
15.3.5.1 Radiation Dose Monitoring..... 318
15.3.5.2 Methods of Protection..... 319
15.4 Summary..... 321

CHAPTER 16 - Sun and Climate..... 322
16.1 The Early Search for a Solar-Terrestrial Connection..... 322
16.2 The Sun During Recorded History..... 324
16.2.1 Radiocarbon Dating..... 325
16.2.2 The Maunder Minimum..... 326
16.2.3 Long Period Cycles..... 326
16.3 Global Climatological Trends..... 328
16.4 Solar-Terrestrial Weather Connections..... 330
16.5 Other Possible Causes of Climate Variation..... 333
16.5.1 Volcanic Activity..... 333
16.5.2 Continental Drift..... 334
16.5.3 Pollution..... 334
16.5.4 Galactic Position..... 334
16.5.5 The Answer?..... 335
16.6 Summary..... 336

GLOSSARY..... 337

BIBLIOGRAPHY..... 348

LIST OF ILLUSTRATIONS


| FIGURE | | PAGE |
|---|---|---|
| 2.1 | Solid vector results from adding dashed vectors | 17 |
| 2.2 | Wave in Space | 19 |
| 2.3 | Phase Difference Between Two Waves of Equal Amplitude and ...velength | 21 |
| 2.4 | Constructive and Destructive Interference of Coherent Wa... | 21 |
| 2.5 | Combination of Two Waves of Differing Amplitude and Freq ... | 22 |
| 2.6 | Electromagnetic Wave in Space | 22 |
| 2.7 | Plane Polarized Signal (EM) Emitted by a Long Wire Antenna | 23 |
| 2.8 | Successive Views Along the Line of Propagation of the E Field Component of an EM Wave Taken from a Fixed Point Away from the Transmitter | 24 |
| 2.9 | Successive Views from a Fixed Point Along the Line of Propagation for a Circularly Polarized EM Wave Showing the Resultant E Field Component of the Wave. The B Field Vectors Would be Similar but Rotated by 90° | 25 |
| 2.10 | Ideal Blackbody Radiation Curves for Various Temperatures | 33 |
| 2.11 | Electromagnetic Spectrum | 34 |
| 5.1 | Color-Magnitude Diagram. Luminosity increases upwards. Temperature increases to the left | 47 |
| 6.1 | Equal areas swept out in equal times means a higher speed at closest approach | 53 |
| 6.2 | Relationship of Earth and Sun Rotational Axes to Ecliptic | 55 |
| 6.3 | Sample Page from Astronomical Almanac (1982) | 58 |
| 6.4 | Stoneyhurst Disk | 59 |
| 6.5 | Carrington Longitude | 60 |
| 7.1 | Photospheric Temperature Profile | 67 |
| 7.2 | Latitudinal Variability in Solar Rotation Rates | 71 |
| 8.1 | Zurich Classification System | 77 |

8.2 Mount Wilson magnetic classification, alpha and beta classification.... 78

8.3 Mount Wilson magnetic classification, beta-gamma and gamma groups..... 79

8.4 Mount Wilson magnetic classification, delta configuration..... 79

8.5 Sunspot Group Motions. (a) Proper motion in a newly formed group, and (b) The production of a spot group by splitting a larger spot..... 84

8.6 Differential rotation causes the leader spot to possess the hemispheric magnetic polarity..... 85

8.7 Plot of the Annual Zurich Sunspot Numbers..... 89

8.8 Long-term Forecast of Sunspot Number..... 90

8.9 Maunder Butterfly Diagram..... 90

8.10 Sunspot Latitude Migration and Leader Polarity Reversal..... 91

8.11 Change in plasma frequency with altitude above the photosphere..... 93

8.12 Quiet Sun Radio Variations..... 94

8.13 Slowly Varying Component..... 95

8.14 Active Sun Component of Solar Radio Emission..... 96

8.15 Comparison of Sweep Frequency Bursts..... 101

8.16 Evolution of a Large Flare and Associated Emissions..... 103

8.17 Classification of Discrete Frequency Bursts..... 104

8.18 Castelli U Criteria..... 106

9.1 Magnetic field line orientation near the sun as viewed from above the ecliptic plane. Field controls plasma in Region I. Plasma controls field, stretching it out in Region III. Region II is the effective source of the solar wind..... 118

9.2 Archimedian spiral structure of the IMF (a) in stationary reference plane and (b) rotating with the sun, showing plasma flow and magnetic field direction..... 119

9.3 Formation of a current sheet (meridional view). Current flows out of the plane of the paper along the dashed line. The earth is pictured above the current sheet..... 120

9.4 (a) Wavy Structure of the Heliomagnetic Current Sheet and (b) combined meridional and oblique views..... 122

9.5 Interaction of a CIR with the current sheet (a) from above the current sheet and (b) a meridional view..... 123

9.6 A CIR from above the ecliptic showing flare related plasma and IMF compression. A qualitative sketch, in equatorial cross section, of a flare-produced shock wave propagating into an ambient solar wind. The arrows indicate the plasma flow velocity, and the light lines indicate the magnetic field..... 125

9.7 Solar wind and density variations during a sector transition - the distribution (along the earth's orbit) of the solar wind velocity and density within a sector. The abscissa is reckoned from the time of the crossing of a sector boundary..... 126

9.8 Geomagnetic conditions during a sector crossing. Highest values of geomagnetic activity usually occur 1 to 3 days following SSB crossings..... 127

9.9 Speed of the Solar Wind, 1962-1970..... 127

9.10 Decomposition of the IMF Vector B..... 128

9.11 A current sheet warpage crossing central meridian causes a reversal in the dominant polarity (plus or minus) of the visible solar disk..... 130

9.12 ISEE-3 Coordinate System..... 131

10.1 Simple Dipole Geomagnetic Field..... 132

10.2 Magnetospheric Cross Section..... 136

10.3 Polar Cusps Looking East (from sunset meridian)..... 139

10.4 The Magnetosphere..... 139

10.5 Particle Access to the Magnetosphere..... 144

10.6 Magnetospheric Trapping Regions..... 145

10.7 Geomagnetic Axis..... 148

10.8 Geomagnetic Coordinates..... 150

10.9 Elements of Geomagnetic Field..... 151

10.10 Contour Map of Geomagnetic Field Strength (note tilt to the rotational equator)..... 152

10.11 Quiet Day Magnetometer Traces at Various Latitudes..... 154

10.12 Quiet Diurnal Magnetometer Variation for Equinoctual Months Near Solar Minimum..... 155

10.13 Magnetospheric Current Systems..... 156

10.14 Phases of Classical Geomagnetic Storm..... 156

10.15 Development of an SSC Storm at Several Latitudes..... 159

10.16 The average auroral oval position (dots) for varying levels of geomagnetic activity. Dashes identify the subauroral trough and lines the region of particle precipitation..... 160

10.17 Geomagnetic Substorm Development..... 161

10.18 Magnetospheric Connection into the Auroral Zone..... 172

10.19 Electron Energy Flux (KeV/cm$^2$ sr sec) Versus Magnetic Latitude for a South Polar Pass..... 173

11.1 Atmospheric Temperature Regimes..... 176

11.2 Effects of Solar Activity on Atmospheric Temperature Gradient..... 177

11.3 Elemental Density Variations with Height..... 178

11.4 Density Variations with Height for High Solar Activity..... 180

11.5 Effect of Solar Activity on Exospheric Temperature..... 182

11.6 Atmospheric Density Variations with Altitude..... 183

11.7 Density Variations Inferred from Spacecraft Drag Variations..... 184

11.8 Variation in Exospheric Temperature by Latitude, Local Solar Time, and Season..... 185

11.9 Semiannual Density Variations..... 186

11.10 Solar Activity and Atmospheric Density..... 187

11.11 Effects of Geomagnetic Activity on Atmospheric Density..... 189

11.12 Combined Effect of Density and Radiation, Variation with Altitude - A Chapman Layer..... 190

11.13 Composite Electron Density Profile for a Multicomponent Atmosphere..... 191

11.14 Variation in Plasmaspheric Morphology with Various Levels of Geomagnetic Activity..... 192

11.15 A meridional cross section of electron density with emphasis on the high latitude ionosphere. Values are inferred from spacecraft observations at 350 km..... 194

11.16 Contours of Log $_{10}N_{max}$ for Winter with $K_p$ = 1 (left) and Summer with $K_p$ = 3 (right)..... 196

11.17 Subauroral Trough Morphology from Spacecraft Measurements..... 197

11.18 Diurnal Variation of the F-layer Critical Frequency with Season at Two Sunspot Numbers..... 199

11.19 Distribution of the Equatorial Electrojet System (Units are $10^{-6}$ A $m^{-2}$)..... 200

11.20 Electron Concentration Above (left) and Below (right) $H_{max}$ Near the Geomagnetic Equator..... 201

11.21 Ionospheric Cross-Section Along the Local Noon-Midnight Meridian. Contours are of constant plasma frequency in MHz..... 202

11.22 Ionospheric Response to Varying Levels of Solar Emission as Measured by Solar Sunspot Number..... 208

11.23 Seasonal and Solar Cycle Variations in the Electron Density Profile.... 209

12.1 Sample Ionogram Showing the Ordinary (O) and Extraordinary (X) Traces, Cusps, and Ionospheric Layers..... 219

12.2 Theoretical Ionogram Showing Typical Parameters..... 220

12.3 Blanketing Sporadic E vs Qualifier 1..... 221

12.4 Two Types of Deviative Absorption..... 224

12.5 Lacuna due to Mid-level Irregularities..... 224

12.6 Range vs Frequency Spread F..... 225

12.7 Relationship between MFAC and $H_{max}$..... 225

12.8 Transmission Curves..... 226

12.9 Corresponding Vertical and Forward Scatter Oblique Ionograms..... 227

12.10 High and Low Ray Paths..... 228

12.11 Typical Crossed-Yagi Antenna..... 232

12.12 Sample Polarimeter Trace Showing Amplitude, Alternate, and Reference Channels During Calibration..... 234

12.13 Diurnal variation in polarization comparing a normal graph of a day's change (top) with the two out-of-phase strip chart recordings (bottom). Notice the complementary nature of the out-of-phase traces..... 235

12.14 A SITEC near center of the data record..... 236

12.15 Rapid polarization changes and scintillation in response to auroral precipitation..... 237

12.16 Slab Thickness Nomogram..... 238

13.1 Radio Wave Propagation Modes..... 243

13.2 Single and Multi-Hop Paths..... 245

13.3 Effect of Varying Takeoff Angle on Sky Wave Propagation..... 247

13.4 Skip Zone..... 248

13.5 A Sample 10 KHz Channel in the HF Band..... 249

13.6 Radio Noise Distribution, Local Winter Morning..... 250

13.7 Path Variation Resulting from Altering Control Point Altitude..... 258

13.8 Grazing Incidence on Transequatorial Paths..... 259

13.9 Forward Scatter System Using Backscatter Propagation to Avoid a Weaker Control Point..... 260

13.10 Horizontal Projection of East-West Ray Paths Showing Expected Deflections..... 261

13.11 HF Propagation Windows Showing Typical Diurnal Variation in Winter Mid-Latitudes..... 264

13.12 Multi-hop East-West Path Undergoing Sunrise from Right to Left..... 264

13.13 Comparison of Typical Winter and Summer "Window" Charts for an Upper-Middle Latitude HF System..... 268

13.14 Local time of maximum $f_oF2$ depression resulting from a negative ionospheric storm..... 275

13.15 Ionospheric Storm Effects on Usable HF Propagation Window..... 278

13.16 Nocturnal variations in S.I. for geomagnetically quiet (Q) and disturbed (D) days. E months are March, April, September, and October. D months are November through February, and J months include May through August..... 282

15.1 Changing Position of Magnetospheric Plasmas..... 298

15.2 Plasma Injection Signature from EP Data..... 300

15.3 Anomaly in Geomagnetic Field Strength..... 301

15.4 Variation in Position of the Plasmasphere with Time and Magnetic Activity..... 304

15.5 Impact of Geomagnetic Activity on Exospheric Temperature..... 307

15.6 Impact of Changing Solar Emissions on Atmospheric Density at 350km..... 308

15.7 Orbital Elements of Satellite Drag..... 309

15.8 Cyclic Variation in Galactic Cosmic Radiation..... 310

15.9 Trapped radiation belt structure showing trapped proton and electron distribution by energy..... 312

15.10 Electron and Bremsstrahlung Dose Estimates..... 320

16.1 Northern hemisphere annual mean temperatures for 1880-1968, compared to the eleven year mean of annual sunspot numbers plotted on center year. Years of ssmax are indicated by arrows at top, ssmin years by arrows at bottom..... 329

LIST OF TABLES


| TABLE | | PAGE |
|---|---|---|
| 10.1 | Geomagnetic Indices Conversion | 164 |
| 11.1 | Characteristics of the Major Ionospheric Layers | 192 |
| 11.2 | Seasonal and Local Time Variations in TEC for a Middle Latitude Station and $A_p$ = 30 | 213 |
| 11.3 | Major Recurrent Meteor Showers | 215 |
| 13.1 | Radio Wave Bands | 242 |
| 13.2 | Traveling Ionospheric Disturbances | 271 |
| 13.3 | $f_oF2$ Variations by latitude, season, and storm magnitude based on time in hours after the onset of the associated geomagnetic disturbance. Results are based on compiled statistics from a number of storms | 276 |
| 13.4 | Solar-Terrestrial and Temporal Dependence of Scintillation | 284 |
| 15.1 | Relative Biological Effectiveness of Various Radiation Sources | 315 |
| 15.2 | Gemini Orbital Parameters and Average Radiation Doses in Millirads | 315 |
| 15.3 | Predicted Radiation Dose (REM) for 1 year Exposure Assuming an RBE of 1 | 316 |
| 15.4 | Probable Radiation Dose Effects for a Sample Population | 317 |
| 15.5 | Established Dose Guidelines in REM | 318 |
| 15.6 | Expected Radiation Doses for Specific Orbits | 320 |
| 16.1 | Major Solar Excursions Over the Past 5000 Years | 327 |

# INTRODUCTION

Our inability to fully use the electromagnetic spectrum and the space environment is at least partially due to the lack of continuous, consistent measurements of this environment. Despite this constraint, the Space Environmental Support System (SESS) has long been charged with forecasting and analyzing the effects of this environment on Department of Defense operations. Success in this endeavor has been and will continue to be highly dependent on the individual's understanding of how the system interacts with the environment. This text is written to bridge the gap between the theorist and the operational user/analyst. It is intended as a text for the SESS analyst and as a reference for users of SESS products and others involved in their interpretation. The early chapters on physics, coordinate systems, and astronomy are intended to provide a brief review of topics of importance with emphasis on those areas of particular interest to SESS.

The bulk of the text deals with the solar-terrestrial system. The choice of topics has been driven by operational considerations. Additional material and theoretical discussions have been limited to that necessary to ensure understanding of operational concerns. This material is intended to complement rather than replace current textbooks and journals dealing with solar-geophysics and astronomy.

# CHAPTER 1

# HISTORY AND OPERATIONS

## 1.1 SESS Mission

The mission of the Space Environmental Support System, SESS, is to observe and forecast environmental conditions and system effects on Department of Defense (DOD) systems operating more than 50 kilometers above the earth's surface. Examples of affected operations include long range communications, radars, and systems using or tracking satellites in earth orbit. The major regions of concern are the sun, interplanetary space, the earth's magnetosphere, the earth's ionosphere, and the upward extension of the neutral atmosphere of the earth. For all practical purposes, SESS provides all DOD weather (environmental) support above the stratosphere.

SESS evolved in response to expanding technology as DOD, primarily the Air Force, began to reach into the upper atmosphere and interplanetary space. The common thread throughout the history of SESS is the continual expansion of support in response to increasing system complexity and expanding needs of DOD agencies.

## 1.2 Historical Milestones

1948 Dr. Mengel (Harvard University) proposed that the Air Force build a solar observatory. Air Weather Service (AWS) helped with the briefings and, later, in the site surveys for the solar observatory at Sacramento Peak, New Mexico.

1958 Just after the first Sputnik launch, the Air Force asked AWS to consider extending its mission. AWS made a comprehensive proposal, but the only part approved was training of AWS officers in geophysics and astrophysics.

1960 The AWS mission statement was changed to include "Special support to the Air Force in closely related scientific fields, including geophysics and astrophysics, as directed by Chief of Staff USAF." (Air Force Regulation 23-1). AWS established a one man Space Physics Branch at Headquarters AWS.

1962 The first solar forecast was issued (1 October 62) as a test. Addressees included Detachment 1, 4th Weather Wing (4WW), Ent AFB; Det 11, 4th Weather Group, Patrick AFB; Space Systems Division (SSD), Los Angeles California; and 6549 Aerospace Test Wing Sunnyvale, California. In November, AWS decided to continue the test indefinitely. The successors to those units are still SESS customers.

1963-64 Increased interest was expressed in solar data. SSD requested $A_p$ and F10 observations. Sunnyvale expressed concern that solar flare reporting was much too slow; reports were needed within 1 hour. Several customers were added to the daily forecasts, and the solar forecasting effort moved to 4WW at Ent AFB (2 people).

1965 The first SESS training courses were conducted. AWS people were sent to augment existing observatories to provide more real-time reporting of solar activity. A Solar Forecast Facility was organized at Ent AFB (4 people).

1966 More training courses were conducted and additional people sent to augment observatories. The Solar Forecast Facility was moved to Cheyenne Mt. as Det 7, 4WW, and Operating Locations (OLs) were established:

OL 10 - Cheyenne Mt, Colorado (forecast unit)
OL 1 Sacramento Peak, New Mexico
2 Sag Hill, Massachusetts
3 Maui, Hawaii
4 Athens, Greece
5 Manila, Phillippines
6 El Segundo, California
7 Ramey, Puerto Rico
8 Fort Davis, Texas (never manned)
9 Tehran, Iran

Continuous space environmental operations began at the Cheyenne Mountain forecast facility.

1968 Forecasting and analysis of the ionosphere began via 440L (an Over the Horizon, forward scatter radar) forecasts (ionospheric MUF/LUF forecasts). There were 46 officers and enlisted personnel at 11 different locations worldwide.

1969 Administrative control of overseas sites was given to 1st and 2nd Weather Wings. More emphasis was placed on the ionosphere, expanding it to a 24 hours/day operation.

1970 The first 24 hour high frequency radio wave propagation forecast was prepared for public release. Det 7, 4WW was deactivated, and 4WW became the SESS manager.

1972 A conference was held to produce cooperative forecasts between NOAA and AWS. NOAA's unit at Boulder received the solar forecast mission, while AWS retained solar event notification, geomagnetic activity and ionospheric analysis and forecasting. 4WW was disbanded, and 12th WS became the parent unit for the SESS forecasting unit and the Continental United States (CONUS) (and Puerto Rico) solar observatories.

1973 SESS Forecast Center moved to Air Force Global Weather Central (AFGWC) to use increased computer power available there.

1974-77 Many new customers were added. Data Monitors (weather observers) were added to handle increased data flow during solar maximum. About 85 AWS and 15 Air Force Communications Service (AFCS) personnel were assigned to SESS activities. Several staff meteorologists were identified for SESS training. The first SOON (Solar Observing Optical Network) and RSTN (Radio Solar Telescope Network) systems were deployed. The SPAN (Solar Proton Alert Network operated by NOAA and using observations in Australia, the Canary Islands, and at Boulder Colorado) operation was curtailed, and Holloman

Observatory was opened. Athen's observatory operation was scaled down in anticipation of new Mideast SOON/RSTN site. Satellite data were acquired from GOES and EP spacecraft.

1979 Tehran Observatory was closed, and Learmonth Observatory opened.

1980 Interplanetary magnetic field measurements were first routinely available to SESS by way of the ISEE-3 spacecraft.

## 1.3 The Early Years

The Space Environmental Support System was conceived prior to 1962 to support operations in space. Air Weather Service responded to the new requirement by forming a small working group at Scott AFB, Illinois. This working group was the formal start of what is now known as SESS. AWS supported the SESS effort, and plans were made for a full-time, operational network of solar optical telescopes to monitor the sun. AWS recognized it did not have the technical expertise to undertake such an endeavor alone; so the Air Force Cambridge Research Laboratory (AFCRL) took the scientific (and to some extent the operational) lead. AFCRL developed basic optical capabilities at its Sacramento Peak Observatory, Sunspot, New Mexico, and the radio work was done at its Sagamore Hill, Massachusetts, radio observatory. As the system began to grow, AFCRL assisted in establishing the world-wide network by instrumenting Manila, Tehran, Ramey, and Athens during 1965-1968.

The operational cadre for SESS developed from the AWS working group and moved to Ent AFB, Colorado in 1964. The operation at this time was not 24-hour, but did provide space environmental data on a daily basis to NORAD and other agencies. The NORAD Cheyenne Mountain Complex (NCMC) was opened in 1965, and the center was moved from Ent to NCMC. The 24-hour operation began in 1965. At this time, the primary customer was NORAD, and most operational products were provided to NCMC. Agencies supported within the NCMC included the Space Defense Center (SDC) and Missle Warning sections. Both of these operations were affected by long- and short-term solar variations. Probably the first significant impact was the 27 May 1967 solar event. NORAD was notified in real-time; however, outside agencies were not aware of the solar flare and made decisions ignoring significant environmental effects. The environmental effects negated the effectiveness of detection systems which had a direct bearing on decisions at the highest levels of government. The impact of this flare provided the necessary force to incorporate space environmental data into the Air Force decision making process. Several additional customers were added solely because of the effect of this event. By 1968, the solar optical and radio networks were generally complete. AWS now had a full 24-hour (minus a short period of no radio coverage in the winter) system observing the sun at optical and radio wavelengths. In the short wavelengths (X-rays), operational support was provided from the VELA satellite system. The data were read out at the Air Force Satellite Control Facility (AFSCF), Sunnyvale AFS, California and provided to SESS at NCMC by phone and teletype. VELA coverage was about fifty percent, so the optical and radio networks were paramount. This data system and operational setup remained basically unchanged until the 1973 move to AFGWC.

In late 1967, an ionospheric section was formed to provide support to ionospheric-dependent systems. The first operational system was the 440L over-the-horizon (OTH) forward scatter radar operating over the Eurasian continent. Day-to-day forecasts of maximum usable frequencies (MUF) and lowest usable frequencies (LUF)were provided to assist frequency managers. In February 1970, the ionospheric section went to a 24-hour operation. Additional ionospheric support was provided to systems such as Cobra Talon and the FPS-85 phased array radar at Eglin AFB, Florida. Automation was required to provide the ionospheric modeling needed by the new systems. The solar function was totally manual, so the ionospheric function was the driver for on-line automation. A teletype terminal was obtained from a semi-operational data base on a computer at Boulder, Colorado. This initial, real-time automated support assisted in providing ionospheric products and enhanced solar operations.

## 1.4 Transition to AFGWC

Planning began in 1969 to move the SESS operational center to AFGWC. Personnel were assigned to AFGWC to begin software work for the move. Somewhere in the mid-1960s, AFCRL contracted for an elaborate software package for SESS data processing at AFGWC. This project should have been designed to operate on the AFGWC computers. The overall design was well-conceived for a scientifically oriented environment, but unfortunately incompatible with the operational orientation of the AFGWC computer system.

Fortunately, other endeavors were not so ill-fated. An example was the conversion of the ITS-78 (Institute of Telecommunication Services) High Frequency (HF) Radio Propagation Program. This was successfully converted to AFGWC's computers and became a workhorse for AWS and Air Force Communications Service (AFCS). In the early years, it produced the base-lines needed to make the 440L forecasts at NCMC. Another major project while preparing for the move was the ingest and real-time processing of VELA data at AFGWC. This was basically completed by the summer of 1972, but still had bugs for some time. One of the conditions for the processing of VELA at AFGWC was additional hardware capability. This maximized core on available computers and required additional mass storage.

The actual move of SESS operations to AFGWC began in 1972 with the responsibility for automated products. One of the first of these was the 27-day solar radio flux prediction. The program was converted from NCMC and was transmitted via the Automated Weather Network (AWN) from AFGWC. The Proton Prediction Study (PPS) was developed by AFCRL and implemented at AFGWC. The real-time alerts and geophysical reports, were transferred to AFGWC in the summer of 1973. The HF Propagation Reports and Reports of Geomagnetic Activity were among the last functions assumed by AFGWC. The geomagnetic activity products were done in an automated mode; they had been manual at NCMC. Use of the computer to transmit event messages also had a difficult start. Card decks were punched, then turned into the System Duty Officer (SDO) for loading. Occasionally, this took an inordinate amount of time.

In late 1973, a UNIVAC 1004 was connected on-line into System I to allow SESS a real-time loading capability to meet critical timelines. Two additional hardware changes improved the overall reliability of event notification. The first was an upgrade of internal devices to Whisper-swifts, which improved display of the data coming from the Automated Weather Network (AWN). The second was a drop on the Astrogeophysical Teletype Network (ATN). This ATN drop allowed SESS operations both receive and transmit capabilities even when AFGWC computers were not available. Another added capability was the NOAA/AFGWC Data Link. This line coupled AFGWC to a minicomputer at Boulder and provided access to real-time x-ray data from the Geosynchronous Operational Environmental Satellite (GOES) system. GOES, combined with VELA, gave AFGWC about 75-80% x-ray coverage. Essentially, the operation looked much like it had at NCMC, with limited automation.

## 1.5 Evolution of SESS at AFGWC

During the period from late 1973 through 1975, requirements continued to increase. The 440L, which ceased operation in March 1975, and Cobra Talon which ceased operation in 1976 are the only exceptions. In 1974, AFGWC assumed the main processing function for AFCS frequency management studies. A direct line was installed from Richards-Gebaur AFB, Missouri to AFGWC. In order to translate the data, a UNIVAC DCT 1000 was rented by AFCS. This allowed the SESS work center to translate 8-level paper tape to cards and vice-versa. This system was very cumbersome, and an additional manpower slot was authorized for this function. The capability to support classified AFCS requirements was also developed.

In 1975, the acquisition of an on-line DCT 2000 changed the work center mode of operation. Until that time, solar and VELA batch jobs were run every 90 minutes, and print was provided to the work center from System I-II-IV with delivery time from 10 minutes to over an hour. The DCT 2000 allowed the batch print to be sent directly to the work center and assisted in data quality control. At the same time, the solar batch was changed from 90 minutes to hourly to allow better use of the automated system for processing and display of ionospheric data. Considerable software effort was expended to maximize the capabilities of the DCT 2000. A new data base design was starting to be seriously considered. With considerable interaction with the Data Base Management Branch, a functional description for the new data base was developed. Two basic options emerged: a vendor supported data base or a data base generally designed along the lines of the AFGWC conventional meteorological data base. The latter design was chosen and was intended to solve the solar event association void created by the failure of the earlier sofware package. It must be emphasized that the Astrogeophysical Data Base (AGDB) was only intended and designed for processing solar data.

New hardware again dominated SESS activities in 1976. Early in the year, a Cathode Ray Tube (CRT) was installed in demand mode on System I. This was the first demand CRT into System I and proved to be a giant step forward in SESS operations. This allowed direct access to System I data bases and a great deal of software was modified or created to use this capability. One such program was the updated Proton Prediction Study (PPS76). This program could normally receive output in less than a minute and increased the work

center's capability to make timely high energy proton predictions. It also assisted in producing more timely geomagnetic analyses and alert messages sent on the AWN. A major new data source also became available during 1976. The Energetic Particle (EP) data line was installed on a piggy-back circuit from SAC and provided 56 channels of energetic proton and electron data from geostationary altitude. Part of the software for use of these data was provided by the Air Force Geophysics Laboratory (AFGL, formerly AFCRL).

In 1977, two additional CRT's were installed, one replacing the ASR into System V and an additional one into System I. These, combined with a new Dataspeed 40 Printer, increased response capability of the work center. An additional Dataspeed 40 was installed in the classified work area in Room 40. In the operational procedures area, a new Proton Prediction Study (PPS77) was installed. A new HF Propagation Report (HFUS1 and HFUS2) was developed and implemented. This involved analyzing and forecasting HF propagation for 20 sectors of the northern hemisphere by a quality factor. The report frequency was changed from 12 hours to 6 hours. Support also began to the Cobra Dane phased-array radar at Shemya, Alaska. This included daily updates and monthly baseline total electron content (TEC) forecasts. The first operational products from the newly developed Four Dimensional Ionospheric Modeling System (4D) were produced in July, 1977. The SAC-Solar Conference, which connected the work center to several SAC agencies by voice, was put into back-up status by using a direct line from AFGWC to SAC computers. Custom tailored messages developed to replace the voice conference produced a considerable time savings for the work center and a hard copy for the users. It also increased SESS dependency on computer systems.

Major milestones in 1978 included addition of new products and increased processing capabilities. The Mystic Star program was transferred from the Pentagon to AFGWC and provided HF radio prediction for selected, high-priority, overwater flights. The input and output data are provided by/to Presidential Airways at Andrews AFB by AUTODIN. Additional high frequency forecasts for AFCS and MAC Command Posts for overwater flights were started. The BEERCAN message, incorporating the same information as the HF Propagation Report in a format for voice broadcast from AFCS Aeronautical Airways Stations, was added. Effort was also expended to improve the high energy proton observations from the GOES satellites. An event program was added to take real-time data from the AWN/ATN and display it on the CRT for forecaster assimilation.

The BEERCAN message was terminated during 1980, and a second solar forecast message was added at 1200Z (complementing the 2200Z product). This was made possible by the conversion of the NOAA forecast center to a 24-hour operation. Developmental work continued on the solar data base, with new software being completed to associate solar events and corresponding geophysical activity. Courses taught at AFGWC to train SESS forecasters gained full Air Training Command (ATC) recognition and approval, and all duty personnel were recertified. Interplanetary field observations from the ISEE-3 spacecraft became available in real-time from Boulder. This provided the capability for about a one hour advance notice of an impending geomagnetic disturbance.

Responsibility for AFCC jobs was transferred to Electronic Communications Applications Center (Annapolis, Maryland) in 1981. The onset of Space Shuttle operations during this year marked a slight expansion of SESS responsibilities. Late in the year, the permanent civilian forecaster position at AFGWC was first filled. By mid-1982, the responsibility for AFCC production had been completely removed from SESS.

## 1.6 Development of SESS Software.

AFGWC's extensive dependence on computer systems is a natural outgrowth of the expansion of the AWS SESS mission and its increasing complexity. Computer systems permit a considerable expansion in support capability and sophistication without a corresponding growth in manpower. The success of this venture is critically dependent on available software.

### 1.6.1 Data Base

When the SESS centralized function moved from the NCMC to AFGWC, it became apparent that the SOLSRT/DATSRT "data base" was grossly inadequate. The transfer of the program VIONSS from the NCMC and the development of the GEMAG2 processor to replace GMAIN defined the data base philosophy for the next 3-4 years. This philosophy was to incorporate all functions dealing with a given data type (error check the data; maintain a data file; produce all displays and analyses required for this data type; and produce any products required from this data) within a single program with its own unique data file. Each data processor was written as the requirement for that data rose to the top of the queue. This led to the "old" data base with ten or so very different handler/display structures. In 1975, an effort was made to develop a real-time SESS data base to ultimately replace the SOLSRT/DATSRT/Separate program data file system.

From July 1975 through December 1975, a working group met to define the needs of such a data base and determine how best to satisfy these needs. The details and results of the working group can be found in the group's final report and the functional description of the Astrogeophysical Data Base (AGDB). The AGDB was designed to allow easy access to all SESS solar data and to allow association and linking of solar event and region data within the data base. The decision was made to design an in-house handler rather than use UNIVAC's DMS-1100 data base, primarily due to the overhead of the DMS-1100 handler. The overall system was to include real-time and batch decoders, a data base handler, a query language to allow the data monitor to interact with the data base, and event analysis software to associate and display the data stored in the data base. Serious effort to replace SOLSRT and its attendent files finally began in late 1981.

### 1.6.2 Display Software

The display software for SESS data was also severely affected by the failure of the initial software package. The main reason for the extensive data base built was to provide input to a fairly comprehensive Solar Display Package. This was designed to give SESS forecasters the necessary tools to monitor the state of the space environment, their primary job. Without the

data base, the Solar Display Package was just so many stacks of computer listings.

As with the data base, the display software grew as requirements increased. However, requirements grew faster than programming resources, so the SPACEWATCH display software was never written. The resulting display system covered only a few data types and consisted of a collection of over 30 programs that range from fairly sophisticated analysis/display packages (VIONSS, TECDIS, EPDIS) to programs that simply list data (HLMS, SOLAR "displays").

The AGDB display and analysis software was to alleviate some of the problems of the "old data base" display software. It would provide a SPACEWATCH software package which would give the forecaster a single, front-end driver to interact with the data base to get the display needed; in essence, a SESS Selective Display Model (SDM). Due to problems mentioned earlier, completion of this display package has proceeded slowly. The real-time event analysis portion was removed from the package, and a completely redesigned version was implemented in early 1978. This gives the forecasters a partial, "quick and dirty" look at incoming data during a solar flare, but is only a fraction of the display software required for just solar data. Perhaps the largest problem in eliminating the program unique files system has been the VIONSS system. The large number and complexity of software which interact with this file have markedly slowed its conversion to the AGDB.

### 1.6.3 Forecast Software

Until fairly recently, most of the efforts in the prediction area have been in forecasting solar protons. The problem of forecasting solar flares was passed back to Air Force Geophysics Laboratory (AFGL) many years ago along with that of forecasting geomagnetic disturbances. A 5 day running mean is still one of the best ionospheric forecasting techniques. A few prediction programs do exist (FLUX for Solar 10.7cm radio flux, QMPCT for ionospheric storms), all initially developed prior to 1972 at NCMC and imported to AFGWC. However, the main thrust has been in solar proton prediction.

The programs PROTON (Smart-Shea) and PROTNK (Kuck) were the first PROTON predictors automated in 1972. PROTON was developed under the AFCRL Space Forecasting Plan; PROTNK came from NCMC. Both used only limited data and were fairly crude. The first upgrade was the delivery of the 1976 Proton Prediction Study (PPS76) which incorporated all available proton prediction schemes into one package. This program, developed by Don Smart at AFGL, was modified to improve forecaster interface and upgrade some of the prediction schemes through 1977, when the next version (PPS77) was put into production. This model was limited by considering input parameters separately, resulting in considerable forecaster subjectivity. Currently, PPS requires that all data used in the prediction come from the forecaster. When the AGDB is completed, the prediction software should run directly from the data base as certain thresholds are exceeded.

## 1.7 Ionospheric Models

SESS ionospheric models have been designed to accurately specify the real-time ionosphere. They include both mission-specific and general purpose programs. Most use a combination of observations and climatology to provide a smoothed data field for forecaster analysis. Some models permit only minor or crude adjustments, while others possess extensive sophistication and capability. In either case, the accuracy of the final product is dependent upon the validity of available climatology and forecaster expertise. The primary models now in use include ITS-7B, Polar, and the Four-Dimensional. New models dealing with scintillation analysis and auroral boundary location are being incorporated. The limited size of the available real time ionospheric data base and the customer-specific nature of several of these models are significant limitations.

## 1.8 AWS/NOAA Interaction

Another segment of the overall AFGWC SESS operation is OL-B, AFGWC, Boulder, Colorado. U.S. Air Force/Air Weather Service (AWS) personnel first arrived at the National Oceanic and Atmospheric Administration (NOAA), Space Environment Services Center (SESC), Boulder, Colorado in 1970. In 1972, AWS and NOAA/SESC began joint operations of the forecasting center. OL-B has served as liaison between the civilian government agency and the military in order to reduce or avoid duplication of services and products, and as liaison between AWS and the civilian scientific community in the field of solar-geophysics. Cooperative technique development, knowledge of improvements in the "state-of-the-art" for forecasting and evaluation of solar activity, and an increased pool of manpower to provide real-time solar-geophysical support have all been real advantages gained from collocating OL-B, AFGWC with SESC.

OL-B personnel have directly contributed to numerous support services for the USAF during its existence at Boulder. Real-time support has included the production of high-quality joint forecasts and the installation of a high speed, computer-to-computer data link system between Boulder and Offutt AFB. Among the data provided over this link are one and five-minute averages for x-ray flux measured by NOAA/GOES satellites, five-minute averaged proton data for events, hourly summaries of proton, electron, and geomagnetic data, and measurements of 90-minute gamma variations of the geomagnetic field taken from the Boulder magnetometer. Other real-time support from OL-B has been furnished via direct "Hotline" from NOAA/SESC to AFGWC; by relaying measurements, such as fifteen-minute geophysical data summaries from the jointly operated High Latitude Monitoring Station; and by assisting in the development of other products, such as the Joint USAF/NOAA Region and Activity Summaries, which are used extensively by USAF/AWS observatories.

Software development for transmissions via the real-time link has been achieved through close cooperation between OL-B and Real-Time Data Service (NOAA/RTDS) personnel. For example, OL-B personnel were instrumental in developing algorithms and software for calculating differential particle flux values at various energy levels and integral particle flux based on satellite measurements. Some further products from OL-B technique development projects include correlations between the planetary geomagnetic A-index and the Fredericksburg A-index; a computer program for estimating polar cap absorption

from proton flux measurements at satellite levels; adaptation of the Smart/Shea Proton Predictions Program for NOAA system use; computer techniques for scaling the Boulder magnetometer; plotting H-alpha synoptic charts; plotting solar x-ray fluxes; and monthly/quarterly solar prediction verification programs.

OL-B personnel have attended and reported on seminars and workshops held by the scientific community, provided special data to military units, and coordinated on special projects. They have ensured the continuous flow of high-quality solar-geophysical and ionospheric data to AFGWC through coordination and quality control. OL-B personnel have worked extensively with NOAA and AFGL to establish better data requirements for systems, such as the new Solar Optical Observing Network (SOON), designed to satisfy the joint flare forecasting mission. OL-B has also aided in establishing contingency procedures so that NOAA/SESC may back-up AFGWC/Space Environment Support Branch (WSE) for events that occur during periods of outage.

## 1.9 SESS Organization

SESS operations at AFGWC are functionally divided into five areas. These include DO (requirements, data sources, and plans), ADSV (decoders), TSIS (development and software modification), OL-B (liaison to Boulder), and WSE (Operational Support). WSE, also known as the Space Environmental Support Branch (SESB), consists of several forecast teams, training personnel, and an administrative section.

SESS operations outside AFGWC are divided primarily into geographic regions. 1WW controls the Pacific solar observatories (Learmonth and Palehua) and the contract for Manila; 2WW has the European solar observatory (Athens); 3WW has CONUS solar observatories (Holloman, Sagamore Hill, and Ramey, Puerto Rico); and 2WS (AFGWC) provides Staff Weather Officers to Air Force Systems Command, where much SESS related work is under study.

# CHAPTER 2

# BACKGROUND PHYSICS

The solar-terrestrial environment involves many elements of science. Before discussing the environment, we turn to a brief discussion of these elements of basic science. They include atomic structure, electromagnetic radiation, radiation analysis, and plasma physics.

## 2.1 Atomic Structure

Matter may be sub-divided into molecules. Each molecule is composed of atoms of one or more elements. For example, a glass of water is composed of water molecules, and each water molecule is composed of two hydrogen atoms and one oxygen atom. The smallest particle which can still be associated with macroscopic material is the atom. Its diameter is about $10^{-8}$cm, defined as 1 angstrom. The atom has a positively charged nucleus composed of various sub-atomic particles. The positive charge results from protons contained within the nucleus. Neutrons, electrically neutral particles with a mass similar to that of the protons, fill out the atmoic nucleus. The nucleus has a net positive charge equal to its atomic number multiplied by the fundamental unit of electrical charge, and a diameter of about $10^{-12}$cm. Electrons, each with one unit of negative electrical charge, circulate about the nucleus. They are more than a thousand times less massive than protons or neutrons. In an electrically neutral atom, the number of electrons surrounding the nucleus is the same as the number of the atom, balancing exactly the positive charge of the nucleus.

The difference between one type atom and another is the number of neutrons and protons in the nucleus. Hydrogen is the simplest atom and is composed of one proton and one electron. Helium, the second-most simple atom, has a nucleus of two protons and two neutrons and two electrons in orbit about the nucleus. By adding a certain number of protons (and possibly some neutrons) to the nucleus and putting additional electrons in orbit about it, we can change one type atom to another. Thus protons, neutrons, and electrons are the building blocks of atoms.

Two or more atoms can be combined to form a molecule. The number and type of atoms in the molecule determine the type of molecule. Two (or more) atoms of the same type can be joined to form a molecule of that element. For example, two hydrogen atoms join to form a hydrogen molecule. Two or more atoms of different elements always form a molecule of something different. For example, two oxygen atoms with a carbon atom form carbon dioxide, or one nitrogen and one oxygen atom form nitrous oxide. The structure of atoms and molecules is not static. Rather, they are constantly changing.

### 2.1.1 Atomic Energy Storage

Atoms can store energy in the orbits of their electrons. The electron is assumed to have a series of concentric orbits, called shells, that it can occupy. The lowest energy orbit, called the ground state, is the most stable.

Each higher orbit represents a discrete amount of extra energy stored. This additional energy makes the atom a bit more susceptible to change. Given too much energy, the electron is no longer constrained by the nucleus. Since the orbits are separated (like stair steps) by discrete quanta of energy, only a certain number may exist. This implies that atoms (or electrons attached to atomic nuclei) can store energy only in certain size packets (quanta). Different quanta of energy must be stored in different atoms, or may not be capable of storage by any atom.

Adding the proper amount of energy to an atom moves an electron from the ground state to a higher orbit (or shell). The more energy added the farther the electron moves from the nucleus. With more than one electron in the atom, the picture is only slightly more complex. Each shell may contain a maximum number of electrons, and the ground state has all electrons as close to the nucleus as possible. Each shell is said to be filled when it contains its maximum number of electrons. The atom can absorb those quanta of energy necessary to move an electron to an unfilled shell. Similarly, electrons which are part of the atoms of a molecule can store energy by occupying the appropriate (atomic) energy orbits.

If we permit an atom to absorb more energy than is required to move its most weakly held electron beyond the highest available orbit, the electron may escape the atom. Ionization is the process of adding or subtracting electrons from an atom or molecule. The positive charge of a proton is the same size (magnitude) as the negative charge on an electron. Since we saw that an atom normally has the same number of protons and electrons, the net charge on the atom is usually zero. The atom is normally electrically neutral. An atom or molecule which is not electrically neutral is called an ion. If we add an electron (negative charge) to a neutral atom or molecule, we make it a negative ion. If we remove an electron from the atom or molecule we make it a positive ion. In space environmental work, we usually deal with positive (electron deficient) ions.

It is possible to ionize an atom or molecule by removing an electron from any orbit. The least energy required to remove an electron is called the ionization potential. This is the amount of energy required to remove an electron from the highest energy (most weakly held) orbit. More energy is required to remove additional electrons or electrons from lower energy (more stable) orbits.

### 2.1.2 Absorption and Emission

Absorption can be defined as the process whereby some electromagnetic (EM) radiation encounters an atom (or molecule), and the radiation gives up its energy to the atom (or molecule). Light is one form of EM radiation. It exists in packets (quanta) of energy called photons. Light is absorbed by an atom when a photon of light ceases to exist, and the atom gains the energy previously contained in the photon. Not all light which passes through a group of atoms is absorbed. Some is reflected, or bounces off the atoms unchanged; some is refracted, or turned in direction; and some is transmitted, or passes through as if the atoms were not there. Only certain quanta can be absorbed by a particular atom. Which photons are absorbed depends on the

arrangement (excited or ground state) of the atom and its electrons when it encounters the photon. If we measure which photons are absorbed, we can determine the state of the atom--hence, its temperature, pressure, and the density of the medium in which it exists.

The energy which the atom gains may be used to ionize the atom or to raise an electron to a higher (excited) energy level (orbit). It may also be stored in the atom (or molecule) as kinetic energy (i.e. speed). If an atom absorbs an amount of energy slightly greater than the ionization potential an electron would normally be removed, and the remaining energy used to accelerate the electron or the remaining ion.

Emission is the inverse of absorption. An atom (or molecule) gives up some energy in the form of electromagnetic radiation. The energy given up may come from energy stored in electron orbits, various motions of the atoms in the molecule, or in various movements of the molecule. Since energy stored in an electron orbit is in discrete amounts, the energy of the emitted photons occurs in precise amounts. Photons of different energy have different frequencies, or colors. If we reverse the process of ionization and let an ion "capture" an electron, we release an amount of energy at least as large as the energy required to ionize the atom. Thus we have two types of emission. One, called line emission, is due to energy released by a "bound" electron, an electron in orbit which falls to a lower energy orbit. This emission results in a particular frequency of EM radiation (pure color of light). The second type of emission is continuum radiation, radiation produced by capture of an unbound, or free electron. This process does not yield a particular color, but typically produces a broad band of radiation (many mixed colors). It sounds like a simple, two state system. You start with a ground state atom, add radiation which is absorbed by the atom to excite electrons into higher orbits or to ionize the atom, and the atom returns to the ground state by emitting a photon. The symmetry of the system is broken by collisions.

Collisions between atoms or molecules of a system redistribute the energy available in the system. Most astrogeophysical systems are gaseous systems, and collisions are an important part of gas interactions. When two particles (atoms, molecules, ions, or electrons) collide, they exchange part of their energy. This trade tends to equalize the energy of the particles. (Notice that this will result in low mass particles having the greatest velocities, since energy is proportional to mass times velocity squared.) In addition, stored energy in one form may be changed to energy stored by other means. Motion is such a form. Motion is just another way of describing the temperature of a gas. A particle may absorb some radiation and collide with another particle, thereby using the radiation energy to heat the gas rather than reemitting the energy as radiation.

### 2.1.3 Temperature

In space physics, two different means of specifying the temperature of a gas are used. Kinetic temperature is a measure of the speed of the gas particles. By contrast, the temperature which we feel (or sense) is called the sensible temperature.

With kinetic temperature, we equate the thermal energy of a gas particle to its kinetic energy. We do this with the equation

$$k\,T = 1/2\,mv^2,$$

where

$k\,T$ = thermal energy of particle,
$1/2\,mv^2$ = kinetic energy of particle,
$k$ = Boltzmann's constant,
$T$ = temperature of the gas particle,
$m$ = mass of gas particle, and
$v$ = speed of gas particle.

We can use the temperature of the particle as a measure of its speed. This speed is a type of energy storage by the gas particle.

Sensible temperature is determined by measuring the energy given by a group of gas particles to a measuring device, a thermometer, for example. If the gas density is low, a thermometer may not receive enough energy from the gas to raise its temperature to that of the gas particles. Low particle density can cause the sensible temperature to differ from the kinetic temperature. This is particularly true in the upper atmosphere and in interplanetary space.

The energy of a single particle is sometimes specified in terms of electron-volts (eV) rather than degrees. One electron-volt is the energy gained by an electron when it is accelerated through a one volt electrical potential drop. A typical proton at the solar "surface" has an energy of about one-half electron-volt, for a velocity of approximately ten kilometers a second.

## 2.2 Electromagnetic Radiation

Electromagnetic radiation is one of the most important elements in the earth-sun interaction. Traveling at light speed, it carries both energy and information. Although light is the most familiar form of electromagnetic radiation, the sun emits many other forms, or frequencies, of electromagnetic radiation. These include radio, x-rays, infrared, gamma rays, and ultraviolet, to name just a few. This radiation may interact with matter in a variety of ways. Depending on the interaction, the radiation may act like a particle or like a wave. EM radiation is thought to result from variations in an existing electric or magnetic field.

Several processes can produce electromagnetic radiation. Any emission of an EM wave involves changing energy which was stored in matter into a photon. The energy storage mechanism may be kinetic (speed of particles), atomic (electrons stored in energetic orbits), or molecular (ionized particles).

Cerenkov radiation occurs when a particle enters a region in which its speed would be greater than the speed of light. As it enters, it slows to

below the speed of light in that material and gives up the extra energy as radiation. Cerenkov radiation is evidence of relativistic particles (highly energetic particles moving at near light speed). Some solar radio bursts involve Cerenkov emission.

Bremsstrahlung (German for braking radiation) results from a rapid change in the speed and/or direction of a charged particle. This acceleration is due to the attractive or repulsive force fields of atomic nuclei near which the particle moves (note that acceleration is variation in speed or direction). The energy required to accelerate the particle is given off, often as a gamma or x-ray photon. This is thought to be the origin of at least a portion of the x-ray flux associated with a solar flare.

Synchrotron radiation is emitted by high-energy charged particles which spiral about a magnetic field line. As the particles change directions (accelerate) they give up energy. Any charged particle trapped by a magnetic field will eventually give up its kinetic energy as synchrotron radiation, if no other processes are at work. Synchrotron radiation shows the existence of high velocity charged particles in a magnetic field, and is important in producing solar radio bursts. The name comes from the name of large particle accelerators which produce similar radiation on earth.

Plasma radiation is the emission of oscillating charged particles. When a plasma particle changes direction in its oscillation (it is accelerated), it gives up energy. If no other forces exist, an oscillating plasma particle will eventually give up all its energy as a plasma wave. The photon produced has characteristics determined by the oscillating particle which emitted it. Some solar radio bursts are plasma emissions.

Thermal radiation results from the random, or thermal motions of atoms constituting the emitter. If the emitter is a plasma, both thermal and plasma emission may occur. Collisions occur, each releasing and/or absorbing photons of varying energy. Consequently, thermal radiation covers a wide range of colors. An important class of thermal radiators are termed blackbodies.

### 2.2.1 Fields

The concept of a field is crucial to any discussion of solar-terrestrial physics. Two like electric charges, such as two electrons, repel each other. How does each know that the other is there? Each electron is said to create an electric field around it, and each electron senses the electric field of another electron and "pushes" against it. An electric field is said to be composed of imaginary "lines of force" which show the direction of acceleration of a positive particle at any point in the field. The field of an electron would have lines of force pointing towards the electron. We detect the presence of the electric field by observing the force it exerts on a charged particle. The strength of the electric field (E) is directly proportional to the strength of the measured force, and its direction is in the direction of motion of positively charged particles. Thus, the electric lines of force are vectors, with magnitude and direction. If we could take an unlimited number of measurements of the strength and direction of the field very close together in time, we could determine how the field varies with time at all points in space, and we could define the field exactly.

We can detect a stationary magnetic field by moving a charged particle through the field at some speed and measuring the force the field exerts on the particle. If the particle is moved parallel to the magnetic field, it would experience no force. If it is moved in any other direction, the force on the particle will be perpendicular to both the motion of the charge and the magnetic field direction at that point.

Combinations of electric fields or magnetic fields may be added, using the rules of vector addition, to show the resultant field. If two fields are measured separately, and then applied simultaneously to a charged particle, their effect is the vector sum of their individual effects. This is clarified in Figure 2.1 for electric fields. The same figure is true for magnetic fields, if E is replaced by B (E represents the strength and direction of the electric field - it is a vector. B does the same for the magnetic field.) Just as a charged particle moving in a field experiences a force, so too, does a charged particle moving in a combined, electromagnetic field. This force, called the Lorentz Force, is described by the vector equation

$$F = qE + qv \times B$$

where F, E, v, and B are all vectors and the x is a vector cross product (similar but not identical to multiplication of scalars). The force due to the electric field is in the direction of motion of the protons or positive ions. It is combined with the force due to the magnetic field, the result of which is perpendicular to both the existing magnetic field and the initial velocity of the particle. If the particle's initial motion (v) is towards the left of this page, and the magnetic field is directed toward the top of this page, the particle will experience a force into the page (assuming the particle has a positive charge). Reversing the direction of motion (to the right) or the magnetic field (towards bottom of page) or charge (to negative) would result in a force out of the page.

As soon as the particle moves slightly (into the page, say) in response to this force, the direction of v is changed. Now v is into the page. If B is still towards the top of the page the particle will feel a force towards the right. The particle's motion to the right in response to the force again changes v. This, in turn, alters the direction of F. F will now be out of the page. The charged particle describes a circle (actually, a spiral) about the magnetic field line. Opposite charges will spiral in opposite directions.

If the electric field is due to some external source it will cause the charged particles to move in one direction or another (depending on charge) as they spiral about the magnetic field. It is possible to determine the direction of F at any time by what is known as the "right hand rule." Rotate the vector v so as to bring its point into alignment with the point of B (by the shortest route). F will lie in the direction of advance of a right-handed screw turned in the same direction that v was turned. Try this thought experiment on the above discussion until you feel confident with it. It is important in understanding how particles move within the earth's magnetosphere and ionosphere.

Electric and magnetic fields are not independent except for the special case of static fields. Any time either an electric or a magnetic field

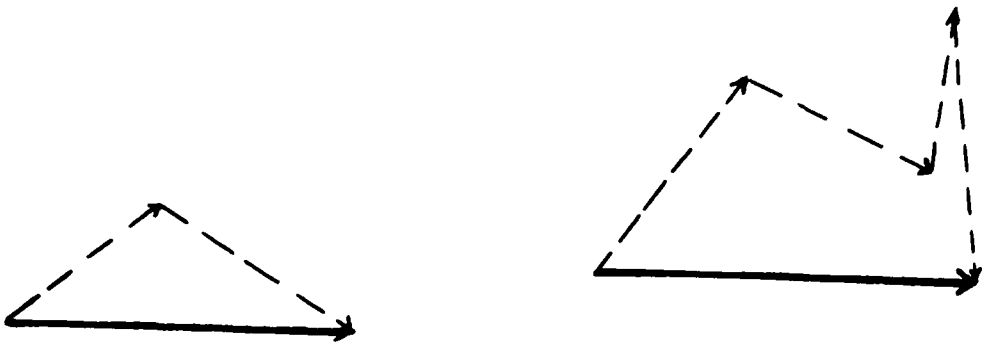


Figure 2.1 Solid Vector Results from Adding Dashed Vectors.

changes its magnitude or direction a field of the other type is produced. Michael Faraday showed that a changing magnetic field would produce an electric field, the basis for an electric generator. Ampere's law shows that movement of an electric charge (and the resultant change in electric field) produces a magnetic field, a concept used in driving an electric motor. (A moving charge is, by definition, a current.)

Electric fields are produced either by a collection of electrical charges or by magnetic fields which change in time. A single positive charge in space will produce a simple radial electrical field around itself pointing outward (i.e., a positive test charge will feel a force directed radially away from the positive charge we are examining). A negative charge will have the same field only directed radially inward. If a postive and a negative charge are brought close to one another, their fields combine in such a way as to give what is called a dipole field.

Magnetic fields are produced by either movement of charges singly or in currents, or by electric fields which are changing in time. Even on the atomic scale, tiny currents set up by the orbital spin of electrons or by their intrinsic spin produce atomic scale magnetic fields. The equivalent of the electric dipole, the magnetic dipole has a similar utility in discussing matter/field interactions. In the case of the electric dipole, any superimposed uniform electric field which is not parallel to the axis joining the charges will repel one of the charges and attract the other until the axis of the field is parallel to the axis of the charges. The same will happen to the magnetic dipole placed in a uniform magnetic field, just as if "magnetic charges" had existed in the positions which the electric charges occupy in the electric dipole. The latter effect explains the operation of a magnetic compass.

### 2.2.2 Waves

Maxwell is really the originator of the concept of electromagnetic radiation. He sifted through the experimental work performed in electricity

and magnetism up to his day (1865) and, by adding several missing pieces, was able to provide a concise, theoretical basis for all classical electromagnetic phenomena. The equations which bear his name remain the starting place for the study of the interaction between electromagnetic radiation and microscopic matter. Perhaps the most startling new idea to be derived from Maxwell's work was that disturbances occurring in an electric or magnetic field should propagate away from their origin as an EM wave. The electric and magnetic fields should be coupled and detectable at a point remote from the place of the original disturbance. His calculations showed that such a wave should travel in a vacuum at a particular speed known as the speed of light (c). Hertz, in 1887, confirmed that EM radiation did indeed exist.

EM radiation can be treated as a stream of photons or as wave. The wave-particle duality of EM radiation requires that we turn, briefly, to a discussion of waves. Wave properties can be illustrated using a smoothly varying electric field as the source of the wave. The maximum strength of the electric field ($E_o$) determines the wave's amplitude. One complete oscillation of the field strength, i.e., from any point on the variation to the point where the pattern begins to repeat itself, is one cycle. The distance covered by the wave during one cycle is its wavelength. The time (T) required for one cycle is the period. The number of cycles which can occur during a given time is called the frequency. Note that frequency is the inverse of the period. The argument of the sine function describing the disturbance is called the phase. The phase of a wave tells us where along the wave we are at the instant.

Any disturbance in an electric or magnetic field will cause energy to be radiated. Only a few disturbances produce pure sinusoidal electromagnetic waves. Sine waves, however, are extremely important, because many natural (and artificial) sources of radiation produce close approximations to pure sinusoidal oscillations. Many real disturbances can be represented to any degree of accuracy by superimposing pure sine wave oscillations. Therefore, the concepts just introduced to describe such oscillations, i.e., frequency, amplitude, and phase, are basic to further work.

The spatial properties of the wave are illustrated in Figure 2.2. The variable electric field at the origin is assumed to be constant over the Y-Z plane which passes through the origin. This ensures that the amplitude of the fields which move down the X-axis will remain constant, simplifying the analysis. The wavecrests of the generating electric field propagate in a vacuum at the speed of light, and either faster or slower in material media. Their speed is called the phase speed of the disturbance. (Phase speeds, but not energy can exceed the speed of light. Einstein's prediction that the speed of light should be the maximum attainable in the universe refers to energy carrying signals. In a dispersive medium, i.e., one in which the speed of propagation varies with the frequency, energy is carried at the group velocity. The group velocity is always equal to or less than the speed of light. Group velocity is the speed of a "group" of waves, and a group of waves is required to carry information.) Conceptually, a photon is just a single wave packet. For now, we will disregard the magnetic field which would result from the varying electric field. It would be at right angles to the electric field and the direction of wave propagation.

One complete oscillation will occur in the time period T. The radiation associated with the beginning of the oscillation will have traveled (in a vacuum) a distance of (c) times (T) or what is normally designated $\lambda$ , the wavelength. The sinusoidal generating field will be reflected as a sinusoidal wave along the X-axis.

In Figure 2.3, the concept of phase difference is illustrated. If two electric fields begin varying at two times separated by 1/4 of their oscillation period (T/4) their separate wave forms would appear as in Figure 2.3. One wave would "lead" the other in phase angle by 1/4 period or pi/2 radians in the argument of the describing sine function (the argument will vary by 2pi radians = 360° in one period). Both would be identical in frequency, maximum amplitude, and speed of propagation, but they would be "out of phase" by 90° (90° is pi/2 radians). Phase differences can occur, because two sources of similar frequency radiations are out of phase; or they can occur at a distance from a single source of radiation when one part of the signal must travel a longer path to a point of detection than the other parts. Two fields oscillating in different directions (say along the Y and Z axes) may be propagated at different speeds through a medium, changing, therefore, their phase relationship to one another with distance from the emitter. Two signals which show a steady or easily described relationship between their phases are said to be coherent. This will be the case for almost all sources of EM radiation used in communication systems. Coherent waves may combine in two ways. If they are in phase, constructive interference will result. If the two waves combine out of phase, destructive interference results. Examples of both situations are shown in Figure 2.4.

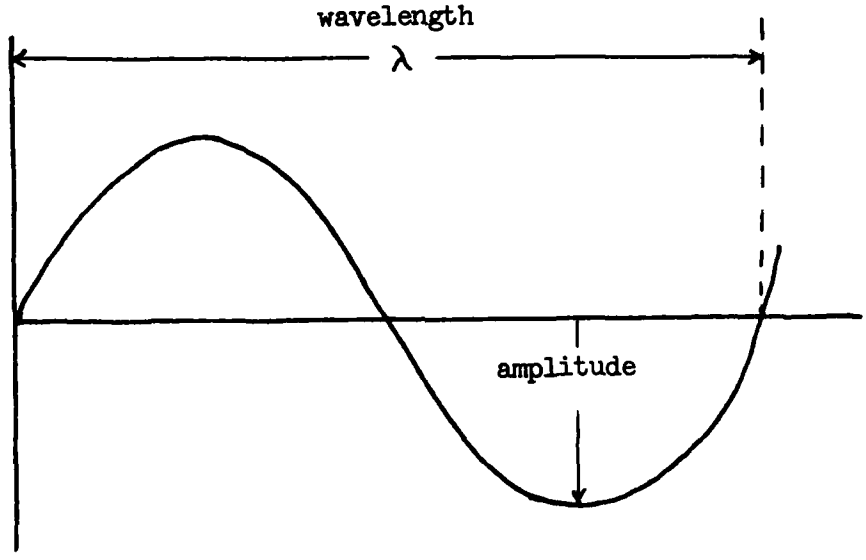



Figure 2.2 A Wave in Space.

If several waves have randomly changing phase differences, they are said to be incoherent, and they will not combine as straight-forwardly. We will not discuss this latter possibility; most effects of interest are related to the coherent condition.

The possibility for synthesis of several signals allows us to discuss the effects on a signal in terms of frequency, with the understanding that those effects apply only to the components of the signal oscillating at or near that frequency. It is possible that there will be alteration of the waveform of a non-sinusoidal signal which passes through a medium whose effects depend on frequency (a dispersive medium), since the combined oscillations have different frequencies.

The speed of a pure sinusoidal wave in a vacuum or a material is its phase speed. Figure 2.5 shows what occurs when two signals, different in frequency, are combined. The waves are in phase (zero phase difference) in some places and show as much as a 180° phase difference in other places. The resultant waveform shows an amplitude which is at some places the sum of the amplitudes of the combining signals and at other places their difference. The envelope of the combined wave is now a periodic function broken into "beats" or "groups". If the medium in which these groups were formed shows a constant phase speed for every frequency, these groups travel at this speed. If, however, the medium is dispersive, the groups show a distinct speed of their own. In all cases where a group speed is meaningful, energy is propagated at the group speed. Group speed is the speed of any "crest", "trough", or other feature on the "envelope" enclosing the group of waves. When the phase speed exceeds the group speed, waves conceptually move through the envelope.

Electromagnetic radiation can be thought of as an electric wave moving through space in phase with a magnetic wave. (One produces the other, so the wave travels through space by continuously regenerating itself.) The waves oscillate in planes which are perpendicular to one another and to the direction of propagation. Such a wave is shown in Figure 2.6. Notice that the electric and magnetic vectors which describe the wave at a given instant lie in the same plane (the Ey-B plane). Moreover, this plane is perpendicular to the direction of propagation. Such a wave is called a transverse wave. EM waves are transverse waves.

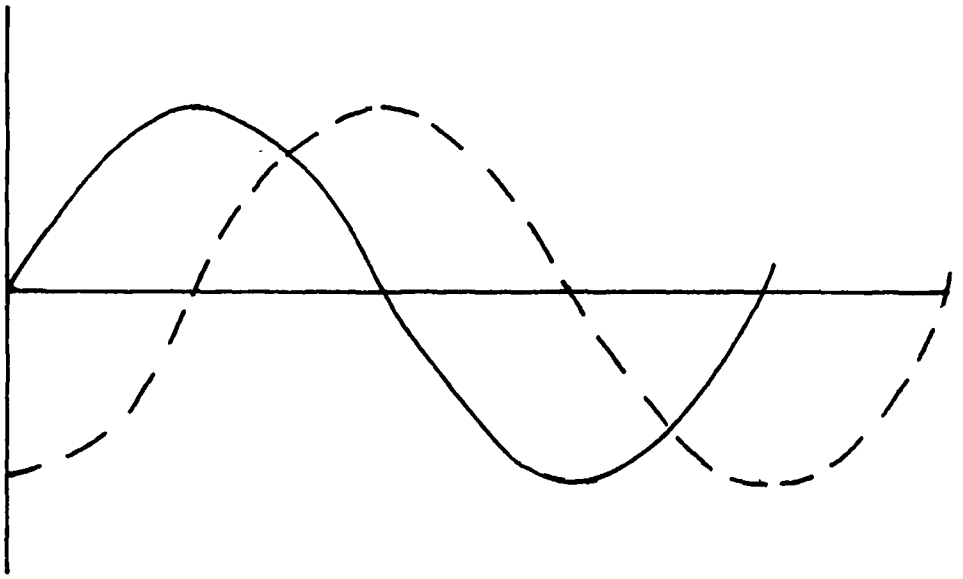


Figure 2.3 Phase Difference Between Two Waves of Equal Amplitude and Wavelength.

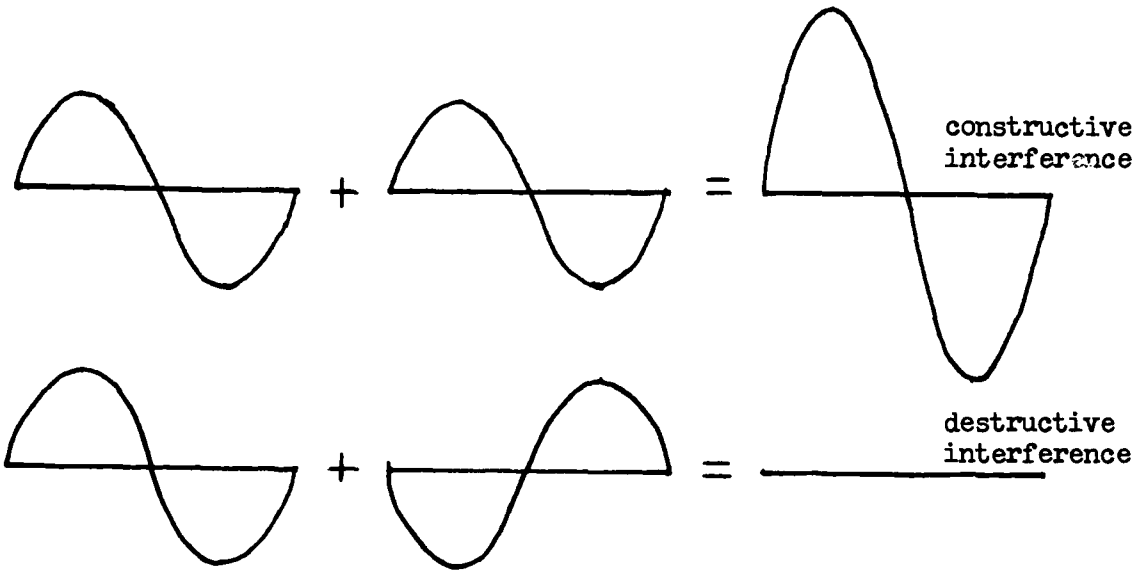



Figure 2.4 Constructive and Destructive Interference of Coherent Waves.

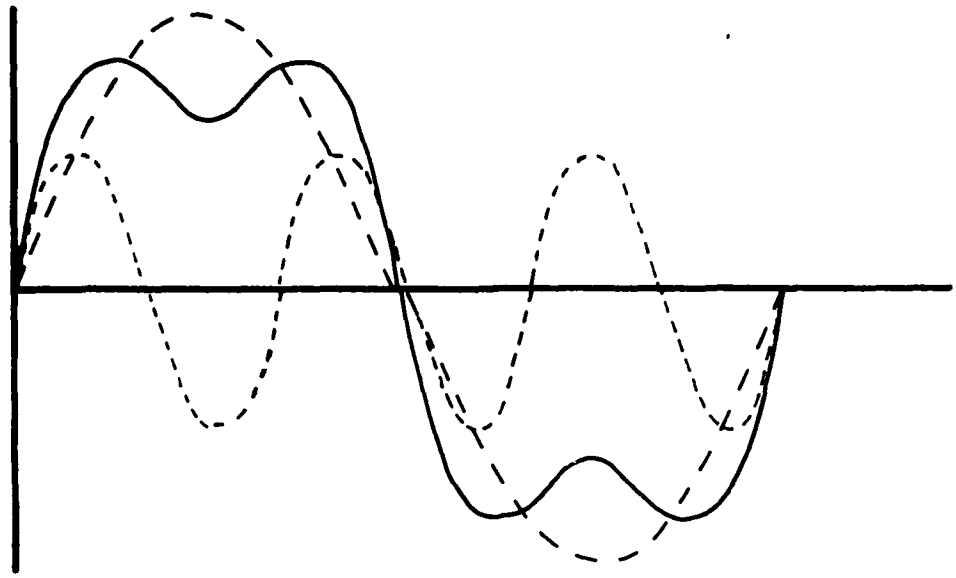


Figure 2.5 Combination of Two Waves of Differing Amplitude and Frequency.

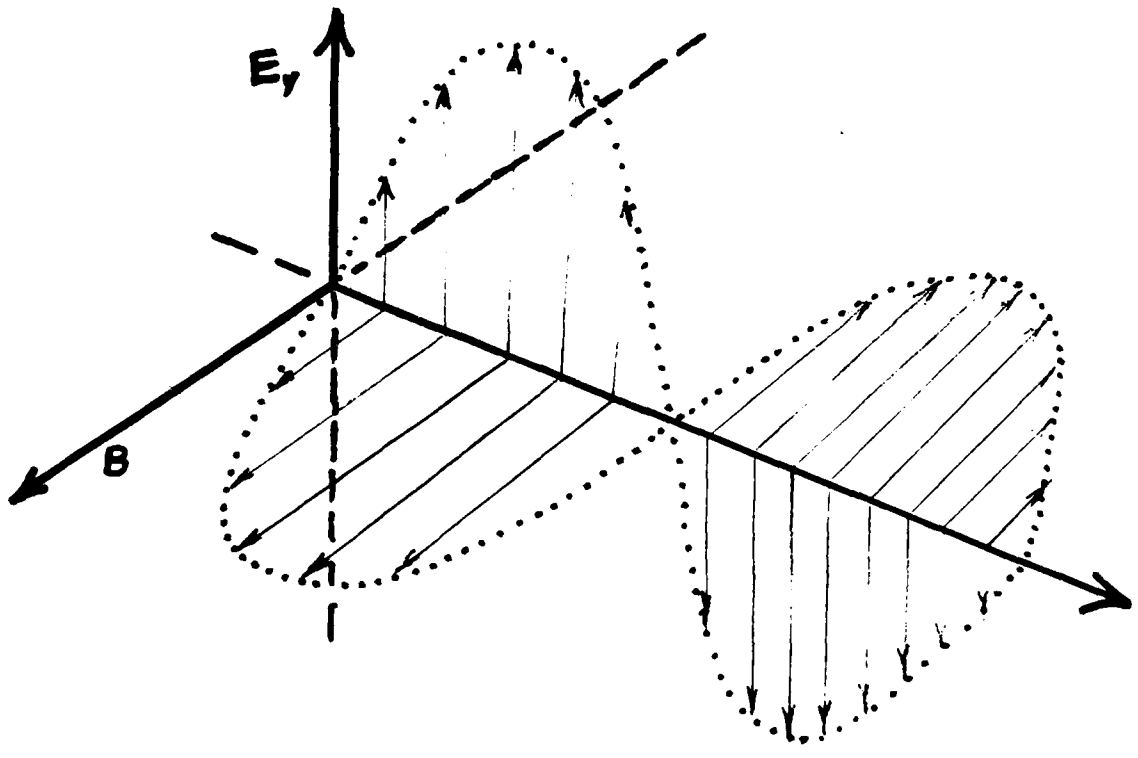



Figure 2.6 Electromagnetic Wave in Space.

A moving electron will produce a time-varying electric field. This will produce an EM wave. If the electron is now confined to move up and down a long wire (an aerial) the electric field component of the EM wave produced will be confined to a plane. The resulting magnetic field component will oscillate in a plane perpendicular to the aerial (also called an antenna). Since the EM wave is a transverse wave, both the E and B wave oscillations are perpendicular to the direction of propagation. In what follows, the B component will be neglected for simplicity.

Notice that the line of propagation and the "line" of oscillation of the E field define a plane (see Figure 2.7). A wave for which this is true is said to be plane-polarized. Polarization exists when transverse vibrations occur in some regular manner. The EM radiation emitted by a plasma of randomly situated particles would be unpolarized. Unpolarized light can be polarized by reflection or use of a polarizer. Plane polarized radiation is a fairly restrictive form; other forms exist.

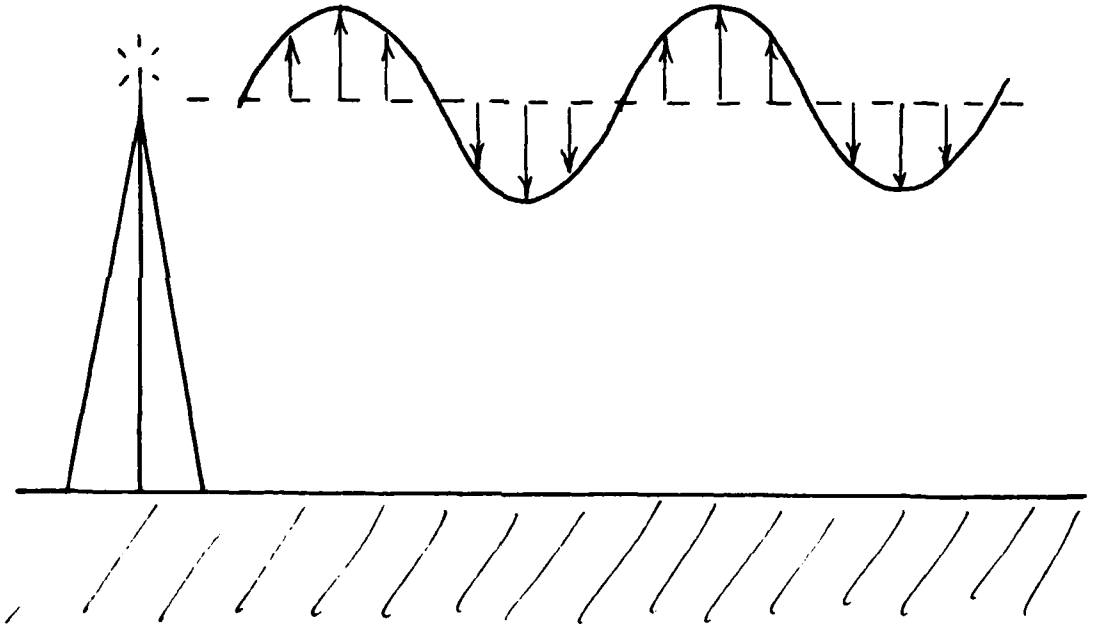


Figure 2.7 Plane Polarized Signal (EM) Emitted by a Long Wire Antenna.

We typically define the polarization of an EM signal in terms of the electric field component of the wave. If we were to look along the direction of propagation at the electric field vector in Figure 2.7 it would be found to vary in time (at a given point a fixed distance from the antenna) as shown in Figure 2.8. Such a wave is termed linearly polarized. A plane polarized wave is also linearly polarized. A similar pattern would appear for the magnetic (B) field component, though it would be rotated $90^{o}$ from the E field line.

Now imagine two antennas like that of Figure 2.7 aligned 90° to each other. A time-varying signal is induced in each antenna, so two electric fields are generated. If this unit is operated as a single antenna, the E component of the generated EM wave will be the vector sum of the two electric fields. If the inputs to each component of the antenna are varied so that the amplitude of the resultant E vector is constant but changes in direction, we obtain a circularly polarized wave. Figure 2.9 is the applicable version of

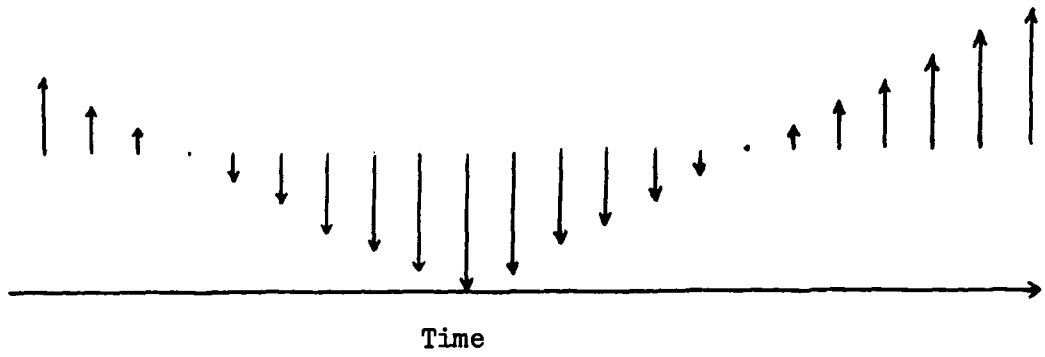



Figure 2.8 Successive views along the line of propagation of the E field component of an EM wave taken from a fixed point away from the transmitter.

Figure 2.8. The locus of all points described by the tip of the resultant E field component is a circle. Notice that the time required for the E component (or the B component) to describe a circle at a given point along the propagation path is equal to the period (T) of the wave. (The same will be true for all other polarization patterns, i.e., linear and elliptical.)

Elliptical polarization results from varying the input to either the X or Y antenna (or both) so that both the direction and the amplitude of the E field (and, consequently, the B field) change in time at a given point. The resultant locus is an ellipse.

The polarization of an EM wave can be helpful in studying either the transmitter or the medium through which the wave passed. The instrument used for this analysis is a polarimeter. The transmitter antenna need not be modified to vary the wave polarization (the method used to create circular polarization above). Such modifications can sometimes be accomplished electronically. The polarization can also be altered by the intervening medium. The change is in the alignment of the polarization vector--not the type of polarization. It is this phenomenon which permits SESS to measure the ionospheric electron content at various locations.

## 2.3 Radiation and Matter

The effects of matter on radiation (energy like heat or light) are in large measure due to the fact that it is possible for electrons to become separated from their nuclei, or in some way reorient themselves with respect to those nuclei. The first condition applies to the general class of matter known as conductors. In fluids and solids, atomic electrons farthest from the nucleus are sometimes easily detached. In particular, if many atoms of a material which exhibits this property are packed together, these "free" electrons circulate throughout the entire assemblage. An example is a copper wire. The average drift of the free electrons of the material in some direction is called an electrical current (although the direction of current flow is taken to be the direction of motion of the protons/positive ions).

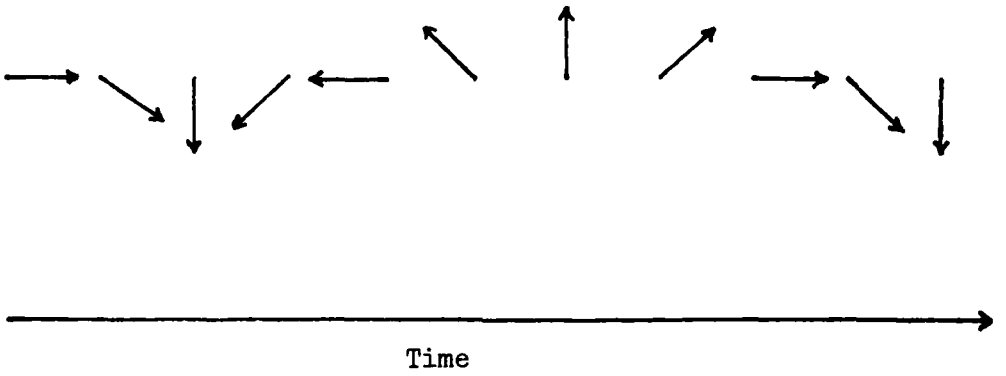



Figure 2.9 Successive views from a fixed point along the line of propagation for a circularly polarized EM wave showing the resultant E field component of the wave. The B field vectors would be similar but rotated by 90°.

All substances exhibit conductivity to some extent, but the range is enormous. Copper will conduct electricity $10^7$ times more easily than will sea water. Sea water will conduct $10^{12}$ times more easily than ordinary glass. The division between good and bad conductors, therefore, is a matter of definition. Sea water, as a passing note, is generally considered to be a good conductor.

It is known from experiment and theory that the conductivity of a material plays a very large role in determining how well and in what manner an electromagnetic wave will propagate through it. A material's magnetic properties can also be important, but at many frequencies, for many media, the magnetic effects are so small that the electrical properties alone determine the propagation characteristics.

In the highest parts of the atmosphere, solar radiation is so powerful that electrons are separated from their nuclei. The result is a plasma

(ionized gas) of free, light, negative electrons and heavy, positive nuclei (ions). The free electrons interact very strongly with radiation propagating through the plasma. In the highest regions of the earth's atmosphere, in the solar atmosphere, and in interplanetary space, this interaction is both complicated and important.

The properties of an EM wave in a vacuum provide a baseline for assessing the impact of other media. Since there are no charges, no dielectrics, and no magnetic materials in a vacuum, the speed of the wave is thought to be the maximum attainable in the universe, the speed of light. All frequencies travel at this speed, so no distortion from dispersive effects occurs. No energy is lost to absorption (because there is no matter to absorb energy). The wave is unchanging in time and space. These characteristics will be contrasted with those found in matter.

A non-dispersive medium is a material which affects all frequencies equally. The phase speeds of all frequencies in a non-dispersive medium are equal. In general, this condition is met only in a vacuum and a perfect dielectric (non-conductor). In fact, a perfect dielectric, in which the electric and magnetic properties are the same everywhere (homogeneous), differs from free space (a vacuum) only in that the speed of the wave (v) is a constant less than c. As a consequence, the wavelength ($\lambda$) in the dielectric is also less than that in free space for a wave of the same frequency (f), since $f = \frac{v}{\lambda}$.

If the dielectric is non-uniform in space, i.e., its electric and magnetic properties change along the wave path, then refraction can occur (the direction of propagation can be altered). Reflection and refraction can also occur at the boundary of two perfect dielectrics of differing properties. However, in these cases, no loss of electromagnetic energy is experienced.

All media are dispersive to some degree. A dispersive medium affects various frequencies differently. The phase velocity of a wave in such a medium can vary with frequency: (1) because the ease with which the atomic and molecular dipoles can be aligned changes with frequency or (2) because there is finite conductivity. Generally, the polarizability of the atomic dipoles does change with frequency. These effects may be so slight, however, that some media approach a perfect dielectric very closely over a certain frequency range.

Finite conductivity introduces another factor not found in propagation in a vacuum. The conducting electrons, in their movement throughout the material, can be thought to collide with the fixed, heavy atoms making up the material and impart some of their energy to these atoms. This energy loss is detected as heat and is called ohmic or resistive loss. The attenuation of the penetrating wave is exponential with distance in a conducting medium, so a medium with finite conductivity can effectively block an electromagnetic wave.

### 2.3.1 Plasma Frequency Effects

A plasma is a "gas" composed of charged particles. In theory, the plasma could be composed of all negatively (or all positively) charged particles and would be called a negative (positive) plasma. In practice, a negatively charged plasma would strongly attract all the positive particles around it. To keep the plasma purely negative would require such massive amounts of energy (to stop the inflow of positive particles) that, in practice, we never get negative or positive plasmas. Instead, we get neutral plasmas, in which the total positive charge balances the total negative charge. The equilibrium solution for the location of the plasma particle is a symmetric, interspersed grid of positive and negative particles. Since each charged particle repels all of the similarly charged particles, the positively charged particles spread out to fill the available space as evenly as possible. The negatively charged particles are also symmetrically distributed in space. Each electron is attracted to the nearby positive particles, but since the positive ions are spread uniformly about each electron, it feels no net force. The two plasmas (one negative, one positive) coexist largely as if each were the only plasma in the space. The neutral plasma does not attract extra charge, positive or negative, so no energy is expended to keep out "extra" charges.

Any disturbance of the location of one charged particle will produce a force to return the system to equilibrium. If I push one electron to the right, it is now closer to the other electrons on the right than on the left. It feels the repulsion of the electrons on the right more strongly than those on the left and so is accelerated back towards its equilibrium position. Like the pendulum on a clock which tries to reach the lowest point of its swing, the electron overshoots its equilibrium position. It then feels a stronger repulsion on the left and is pushed back to the right. The disturbance produces an oscillation of the electron about its equilibrium position. The frequency of the oscillation is closely approximated by the electron plasma frequency, given by the formula:

$$f_{pe} = \frac{(N_e\ e^2)^{1/2}}{M_e} \qquad 9\sqrt{N_e}\ \text{kHz},$$

where $N_e$ is electron density per cubic centimeter, $M_e$ the electron mass, and e the charge on each electron. Since the plasma frequency is inversely porportional to the mass of the particle, and since the mass of a proton is about one thousand times that of an electron, we assume the protons don't move. Since any other positive ion is much heavier than a proton, its plasma frequency is even lower. Thus, we normally assume only the electrons oscillate. The electron plasma frequency is commonly thought of as the natural oscillation frequency of the plasma.

Charged particles are accelerated by electric fields, the electrons in one direction and the positive ions in the opposite direction. If the electric field is constant, the result is a shift of the negative particles relative to positive particles. Magnetic fields also influence moving charged particles by forcing them to spiral around nearby magnetic field lines. The frequency of this motion, known as the gyro frequency, $f_g$, is given by

$$f_g = \frac{eB}{2M_e\pi} ,$$

where B is the magnetic field strength. The particles effectively attach themselves to the magnetic field lines. If the charged particles move other than along the field line, the magnetic field must follow. If the magnetic field moves, the charged particles must follow; they are frozen together. This plasma-field coupling is described by frozen field theory.

Plasmas and electromagnetic waves interact in very particular ways. Since an EM wave has an electric field which first points in one direction and then in the other, it first attempts to move the electrons one direction and then the opposite. The frequency of the changing electric field direction (the frequency of the wave) is compared to the electron plasma frequency to determine what happens. If the wave frequency is much larger than electron plasma frequency, the electric field changes too rapidly for the electrons to react to it, and the slower ions don't even know the wave is present. The wave is transmitted unchanged by the plasma. If the frequencies are equal, the plasma can absorb some energy from the wave and set the electrons oscillating at the electron plasma frequency. If the plasma frequency is much greater than the wave frequency the electrons shift so rapidly that the wave cannot get into the plasma and is reflected back without loss of energy.

When the plasma is less than the wave frequencies but changing along the path of the wave, the wave experiences retardation and refraction. The wave bends and slows as it passes through the plasma. If the plasma frequency is large enough, the wave can be refracted, or bent back, to its original medium without being reflected, or bounced back. Even if the maximum plasma frequency is too low to reverse the wave, it can slow and bend the wave enough to cause system effects. This slowing and refraction is often accompanied by a loss of wave energy to heating.

When the plasma and wave frequencies match, much of the wave energy is used to heat the plasma. Some of the wave energy is used to drive oscillations of the electrons. These electrons hold that energy for a fraction of a second before giving it up as a new EM wave. If, however, the oscillating electron collides with another particle before it gives off a plasma wave, the energy is used to speed up the gas particles (heat the gas).

A magnetic field line in a plasma complicates the interaction of an electromagnetic wave with the plasma. The angle between the magnetic field line and the direction of travel of the wave controls the interaction through the Lorentz force equation. When the wave attempts to put a plasma particle into oscillation and the attempted motion is exactly along the magnetic field line, it is as if the magnetic field were absent. In all other cases, the magnetic field controls the motion of the plasma and changes the electromagnetic wave in the process. Specifically, the speed of the wave in the plasma is slower with the magnetic field present, and the wave polarization may be altered.

### 2.3.2 Polarization Changes

If a linearly polarized EM wave enters a plasma embedded in a magnetic field, the polarization angle will be changed. The amount of change depends on the magnetic field strength along the line of propagation and the electron density. This effect is known as Faraday Rotation and is given in radians by the equation

$$\Omega = \frac{K}{f^2} \int B \cos\theta \, N_e dl,$$

where $K = 2.35 \times 10^{-5}$ ( a constant),

$\theta$ = angle between the direction of propagation and the ambient magnetic field direction (<u>not</u> the B component of the EM wave),

f = wave frequency (in Hertz),

B = magnetic field intensity in gamma, and

$\int N_e dl$ = total electron content (electrons/$m^2$) along the line of propagation in the plasma.

As the angle between the EM wave direction of propagation and the ambient magnetic field approaches 90°, the above equation becomes undefined. (This situation exists for a geostationary spacecraft transmitting an EM wave to its subpoint when both are also on the gemagnetic equator.)

*Conceptually, Faraday Rotation results from the modification of* the EM wave by the magnetized plasma. The E component of the EM wave sets the plasma electrons (and protons) in motion parallel to the E component (and perpendicular to the line of propagation). With a magnetized plasma, the Lorentz Force will act on charged particles which are placed in motion. This will occur unless the particle motion (not to be confused with the wave propagation) is parallel to the ambient magnetic field. The E component incites particle motion; the ambient B field (not the wave B component) deflects the particles from their original path (due to Lorentz Force); and the particles retransmit the E component of the wave in a slightly different (rotated) orientation from that of the original wave. These actions occur within the magnetized plasma and produce a rotation of the polarization vector during the wave's transit of the plasma. They result from the line of propagation of the wave being inclined to the ambient magnetic field lines of force at an angle other then 90°.

An additional effect results if the ambient magnetic field is not parallel to the B field component of the EM wavefront at the point where the wave is incident on the plasma. When this occurs, the wave is split (magnetoionic splitting) into two components: one (called the ordinary wave) with its B field component parallel to the ambient magnetic field, and the other (extraordinary wave) with its B field component perpendicular to the ambient magnetic field. In essence, the magnetic component of the EM wave is decomposed into components parallel and perpendicular to the ambient field in the wavefront. This decomposition results in the two separate waves mentioned.

These two waves travel slightly different paths at slightly different speeds through the plasma. On exiting the plasma, the two components recombine. The different path lengths and different speeds may (in general do) result in a slight phase difference between the two waves at recombination. Since their recombined phase is slightly different, the polarization of the recombined wave may also vary slightly from the original. The degree of phase slippage and polarization change will ultimately depend on the ambient magnetic field, plasma electron density, and path length in the plasma. This effect (B not parallel to ambient magnetic field lines) is often small by comparison with the former (line of propagation inclined less than $90^{o}$ with respect to ambient magnetic field line). Nonetheless, it is significant because of the wave splitting which occurs. These effects are directly applicable to ionospheric analysis of both terrestrial and other planetary atmospheres.

Faraday Rotation and magnetoionic splitting do not alter the type of polarization; only the orientation of the polarization vector. A linearly polarized signal will remain linearly polarized, but its E and B components (fixed with respect to each other) will be rotated through some angle with respect to their original orientation in space. Faraday Rotation would also alter the orientation of circularly and elliptically polarized signals, but the effect would not be readily discernable on the circularly polarized signal.

## 2.4 Radiation Analysis

Electromagnetic radiation is our only source of information on distant objects such as the sun and stars. We can derive three general types of information from an EM wave: direction, quantity, and quality. Directional information is of considerable importance for coordinate systems, galactic structure, and system motions. Quantity of radiation received depends on the distance and brilliance of the source plus the density of the intervening medium. Quality information addresses such features as the temperature, density, and size of the source and is derived from a detailed analysis of the variation in intensity with wavelength of the received radiation. Since stars like the sun are nearly ideal radiators (and absorbers), it is often convenient to use blackbody approximations in their analysis.

### 2.4.1 Blackbody Analysis

An ideal blackbody absorbs all incident radiation regardless of wavelength. Moreover, it emits at all wavelengths. The intensity of radiation at each wavelength is a unique function of the temperature of the emitter. The wavelength at which a blackbody emits the most radiation is inversely related to its temperature. This wavelength determines the "color" of the blackbody.

Hotter objects are generally bluer, and cooler objects redder (since red is a longer wavelength than blue light).

$$\lambda_{max} = \frac{\text{constant}}{T}$$

Higher temperature bodies radiate more total energy, centered on a higher frequency, than do low temperature bodies. This means that hotter objects will be considerably brighter than slightly cooler ones of similar size.

$$E = (\text{const})\ T^4$$

By measuring the spectrum of radiated power versus frequency of a blackbody we can determine the radiator's effective temperature. Of course, the total brightness of an object (luminosity) is also a function of the object's surface area. Luminosity is given by

$$L = 4\pi R^2\ (\text{const})\ T^4,$$

where R is the radius of the body.

The apparent brightness falls off as $1/r^2$ (where r is the distance from the observer to the object), since the object's brightness spreads out in all directions equally. Quiet sun emission is blackbody radiation. See Figure 2.10. Electromagnetic radiation from the quiet sun roughly approximates a 5280°K blackbody radiator. The blackbody curve is well approximated in the visible range, where 41% of the energy is emitted, and in infrared, where 52% of the energy is emitted. A review of the blackbody curves in Figure 2.10 reveals the differences which result from changing the temperature of a blackbody (note: the temperature of importance for a star is its surface temperature).

### 2.4.2 Spectral Analysis

The emission or absorption of a particular wavelength of light (color of photon) is ultimately due to an electron changing energy levels in an atom in the emitter or absorber. A particular wavelength (frequency) photon corresponds to a particular quanta of energy. If no atoms exist (in a particular object) in which the energy difference between two electron orbits equals x, then this object can neither absorb nor emit photons of wavelength corresponding to energy x. Moreover, if such atoms do exist they must have an electron appropriately situated (i.e., in one of the two orbits involved) in order for emission or absorption of a photon of energy x to occur. The number of photons of energy x absorbed or emitted by a particular medium depends on the number of atoms having (1) orbits separated by energy x and (2) electrons in one of the two particular orbits. The former requirement (orbit spacing) depends on the magnetic field and the elements (hydrogen, iron, carbon) present in the object. The latter requirement (electrons appropriately positioned) depends on the temperature and density (pressure) in the material. These dependencies mean that studying the number and energy (wavelength or frequency) of photons emitted (or absorbed) by an object reveals a lot about the object. (Note: the number of emitted or absorbed photons determines the intensity of radiation or absorption at the frequency.)

Spectroscopy is the study of the intensity/wavelength variations in the radiation of a given object. Light of a single frequency is called a "line." The "spectrum" of a particular object consists of all the lines of radiation which it emits and all of the dark lines at which it has absorbed light (or

not emitted any). Since blackbodies emit and absorb at all wavelengths, their spectra will have an infinite number of lines -- some in emission and some in absorption. A study of these spectra reveal much about the composition, temperature, pressure, size, and even shape and motion of the object. A spectrograph is used for this analysis.

### 2.4.3 The Zeeman Effect

If the emitter has a magnetic field (such as those associated with sunspots) additional analysis is possible using a magnetograph to isolate a particular spectral line. In the presence of a magnetic field, the atomic orbits split to yield three separate orbits. One is in the normal position, but the other two are slightly higher and lower in energy, respectively, than the original line. The stronger the magnetic field, the greater the split in the resulting triplet. Electron transitions will then yield photons of slightly higher and lower energies in addition to those of expected energy. Measuring the amount of this displacement provides a measure of the direction and intensity of the magnetic field. This splitting is known as the Zeeman Effect.

The Zeeman Effect is used by the Mount Wilson Observatory (and others) to measure the field strength, gradient, and polarity of sunspot magnetic fields. This data is routinely measured in units of gauss for field strength, gamma ($10^{-5}$ gauss) per kilometer for field gradient, and colors (red is plus, or away from the sun, and blue is minus, or towards the sun) for field polarity. Be careful not to confuse field polarity with radiowave (or EM wave) polarization.

### 2.4.4 Doppler Effect

If an emitter is in motion with respect to the observer, the electromagnetic waves received will have different frequencies (and wavelengths) from that which they would have were they at rest with respect to the observer. Known as the Doppler Effect, this is commonly observed as the change in pitch of a siren as an emergency vehicle approaches, passes, and moves away. The observed wavelength is related to the velocity of the object, v; the speed of light, c; and the rest wavelength. For velocities significantly less than the speed of light

$$v/c = \frac{\lambda - \lambda_o}{\lambda_o}$$

This effect can be used to measure the motion of the emitter. This permits us to measure the motion of solar flares and the sun's rotation rate, among the other things. An object moving away from the observer produces radiation shifted to longer wavelengths and is said to be "red shifted." An approaching emitter's radiation is of shorter than rest wavelength and is said to be "blue shifted." The apparent color of the object changes only in extreme cases. What is shifted are the positions of the spectral lines. Their shift is often no more than a fraction of an angstrom from their rest position.

## 2.5 Summary

The electromagnetic spectrum can be broken down into a number of discrete frequency ranges. These ranges, identified in Figure 2.11, have different

uses and propagation modes. An understanding of why and how various types of emission are produced permits us to infer much about conditions (temperature, pressure, rotation rate, etc.) inside the emitter. This, in turn, helps us develop tools with which to forecast future emissions.

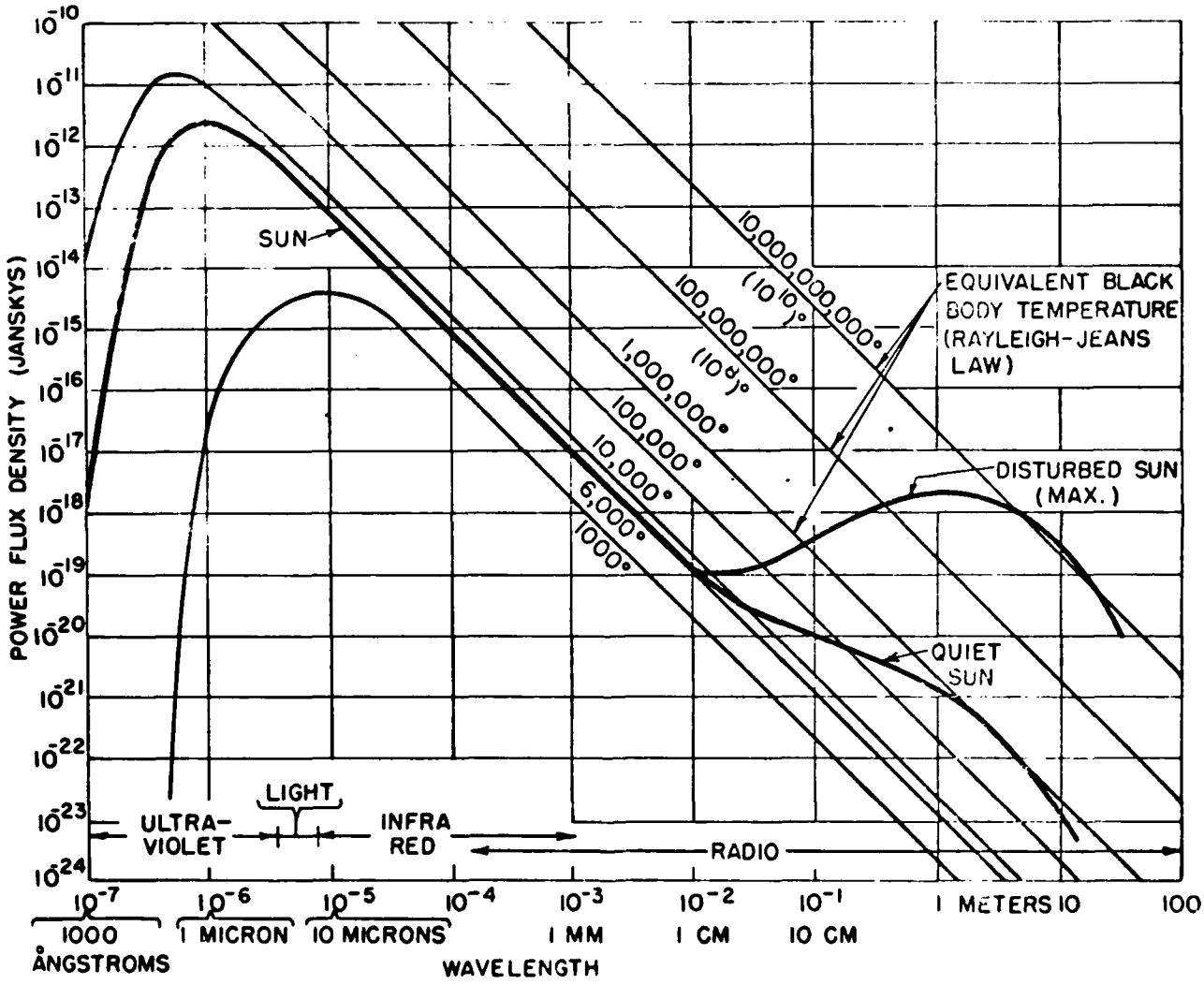



Figure 2.10 Ideal Blackbody Radiation Curves for Various Temperatures. (Valley, Ed., Handbook of Geophysics and Space Environments, 1965).

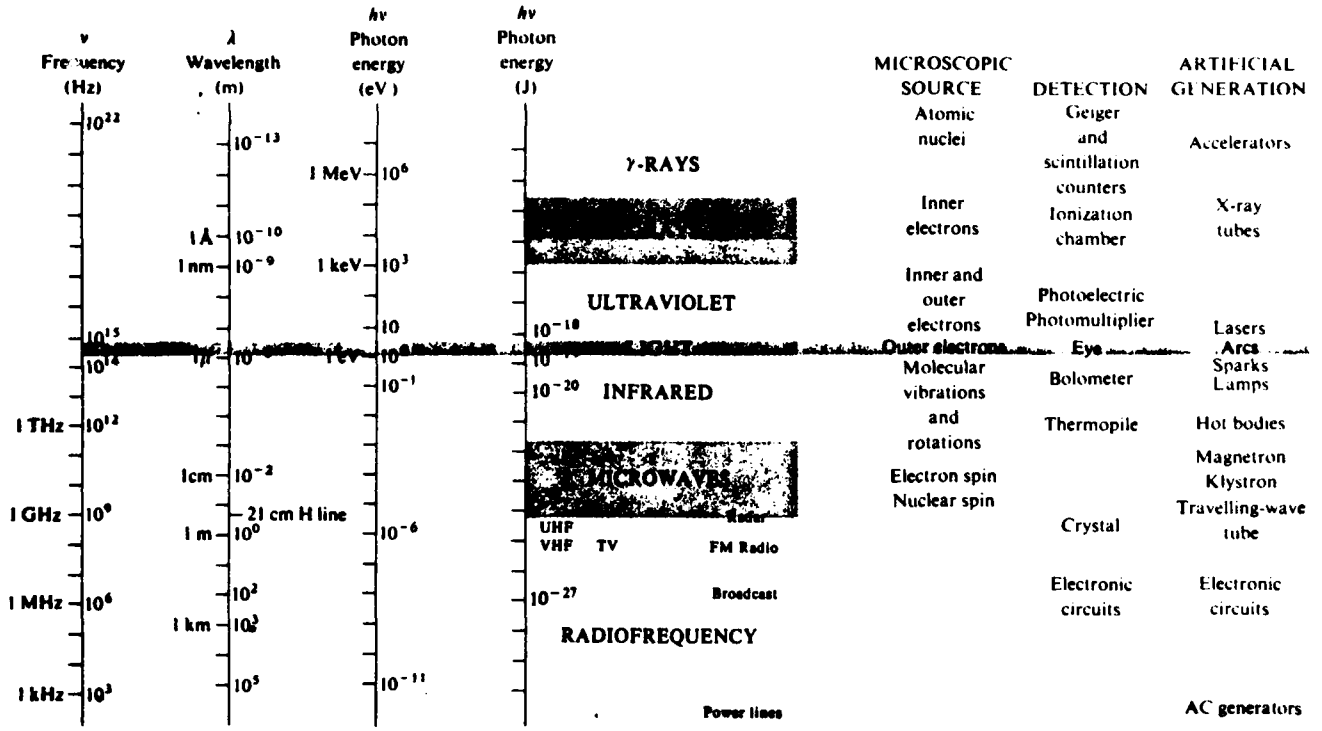



**Figure 2.11 Electromagnetic Spectrum (from Hecht and Zajac, Optics, 1974).**

# CHAPTER 3

# ASTRONOMICAL INSTRUMENTATION

Most astronomical instruments are designed to extend the observer's senses to cover the entire electromagnetic (EM) spectrum. Light, x-rays, and radio waves are but three portions of the EM spectrum which we observe routinely. Although the wavelengths differ, the general principles of operation for observing at these wavelengths are similar. The basic instrument is, in each case, a telescope.

## 3.1 Characteristics of Telescopes

Telescopes come in many shapes and sizes, but they all have the same primary purpose: to increase (and, in some cases, extend to other frequency ranges) the light-gathering power of the eye. Different shapes and sizes are required to efficiently observe some frequencies. X-rays, for example, are best observed from space, since they don't penetrate the earth's atmosphere to the surface. In order to increase the light-gathering power, we must increase the telescope's size.

The size of a telescope is its most significant light-gathering feature. The primary difference between the eye and a large, optical telescope is the size of the "light bundle" that each will accept (and focus at some point). The element which limits a telescope's light acceptance is the diameter or aperture (a) of its primary lens or mirror (in general terms). The light-gathering power of a telescope is proportional to the square of the aperture (since the area of a circle is dependent on its radius squared). Hence, a 10-inch telescope has 100 times the light-gathering power of a 1-inch aperture instrument. The maximum aperture of the human eye is about 1/5 inch. The aperture of a SOON solar telescope is about 10 inches.

The "power" listed for many small telescopes is not light-gathering power, but their magnification. Magnification is not a particulary important feature of a telescope, since it can be changed easily by changing eyepieces. The effective magnification can be determined by dividing the focal length of the primary lens (or mirror), $f_o$, by the focal length of the eyepiece, $f_e$. The light-gathering power of a telescope limits the maximum magnification (M) which can be used effectively with a given telescope. Higher magnification can, of course, be achieved with any telescope simply by inserting an eyepiece of shorter focal length. Magnification greater than M achieves only an increase in blur (i.e. the object will appear larger but much less distinct as magnification is increased). The actual value of M for a given instrument depends on many things, but is approximately equal to 50a for a in inches at optical wavelengths.

Several other terms are often used in describing the performance of a telescope. The focal length (f) is the distance from the lens (or mirror) to the point at which an image is formed for an object located at infinity. (For all practical purposes, the sun is at infinity with respect to an earthbound telescope.) The focal length divided by the aperture is termed the focal ratio. The focal ratio is normally written as f/10, where the focal ratio is

10, and f is the focal length. In reality, f/10 = a, since, by definition f/a = 10, where 10 is the focal ratio. The term "focal ratio" is not commonly associated with a symbol (as f represents focal length, etc.). This accounts for the rather awkward manner in which it is normally written.

## 3.2 Types of Telescopes and Seeing

As a telescope is ranked by its size, it is characterized by the manner in which it forms an image. Two methods are commonly employed - refraction and reflection. A refracting telescope uses lenses to bend (refract) light into an image. A reflecting telescope employs mirrors to converge light into an image. Radio telescopes are commonly reflecting telescopes. Image formation is in front of the primary collector. Large stellar telescopes also use mirrors and are called reflectors. Smaller optical instruments like the SOON telescope are typically refractors.

Some refractors use an evacuated tube to improve their image quality. Just as the water in a stream distorts our view of the stream bottom, so, too does the earth's atmosphere distort our view of the sun and other astronomical objects. Like a lens, the atmosphere bends or refracts the sunlight. Winds and turbulence cause the image to bounce or scintillate (compare to the effect of ripples in the stream). The atmosphere also refracts light into and out of the beam. We can't eliminate the atmosphere without putting the telescope in space, but some large instruments are placed on mountains to get above a large portion of the earth's atmosphere. Still others are flown on aircraft or balloons. The SOON system (and some large research telescopes like those at Sacramento Peak and Kitt Peak) evacuate the tube in which the lenses are mounted. This at least eliminates the deleterious effects of the atmosphere in the area where the image is being formed.

The effects of the atmosphere on a given observation are estimated by the observer. This helps the user compare observations between sites and may account for one site being able to see more sunspots or identify a particular feature while another cannot. Seeing conditions are usually evaluated subjectively on a numerical scale of 1-5. Seeing of 1 is very poor, with plage areas appearing as blobs and small spots sometimes invisible. Fair seeing, when most features are discernable, is classed as 3. Excellent seeing conditions (5) are not often attained in many parts of the world but afford sharp viewing of extremely fine detail.

## 3.3 Optical Instrumentation

Although direct viewing of various objects at the eyepiece can be enjoyable, much of today's research is dependent on the types of instrumentation used with the telescope. Such instruments are intended to make objective measurements of the radiation received. Two general categories of instruments are in routine use today. They measure the quantity of radiation and its variation in intensity with wavelength. In reality, many instruments measure both quantities to some degree. Such hybrid instruments have greatly increased the scientific capability of available telescopes.

The two basic instruments are the photometer and the spectrograph. The photometer is intended to count the number of photons incident on a given detector surface. This is usually accomplished by a photomultiplier tube. Such a device generates perhaps a million photoelectrons for every photon incident on it. By altering the detector surface or inserting filters ahead of the detectors, the brightness of a given source can be measured in various wavelength ranges. A spectrograph spreads a portion of the electromagnetic spectrum out to permit wavelength by wavelength analysis. A rainbow is just a spectrum produced by water droplets replacing the spectrograph. Early spectrographs used prisms, but today's instruments generally use diffraction gratings. Unfortunately, considerable light is lost in spreading out the light, by wavelength, from a given emitter. The intensity at a given wavelength is much less than the integrated intensity of the source (the total brightness is greater than the brightness of any single component color).

Image tubes provided a beginning in the hybrid instrument field. The basic idea evolved using a spectrograph to make individual wavelengths available and then employing a photomultiplier to measure the intensity at each wavelength. Such devices are in wide use today. Image tubes themselves provided an initially crude replacement for the astronomical photograph. The intent was to produce a permanent, measurable image of fainter sources in less time than was possible with film. Likewise, the image tube generated an electronic signature readily available for computer processing. Computer analysis, particularly in real time, has greatly increased the effectiveness of modern instruments.

## 3.4 SOON Instrumentation

The Solar Observing Optical Network embodies current technology to provide a very powerful optical tool. Designed to eliminate subjectivity in flare observing, the SOON is paving the way for improved flare forecasting. The videometer is the heart of the SOON system. It combines a filtered photomultiplier-type system with computer analysis for automated flare patrol and region analysis.

An image of the sun in the light of hydrogen-alpha (one line generated by a hydrogen atom) is scanned by a video system similar to a television camera. The image is divided into a number of tiny squares or pixels (like the matrix on a TV screen). A sensing system calibrated to discern 64 separate brightness levels then determines the brightness level of each pixel. The analyst defines for the computer a certain block of pixels as a sunspot group. Within each sunspot group, the computer separates pixels into four brightness categories - filament (very dark), quiet sun, plage, and flare (including faint, normal, and brilliant categories). The computer then constructs a histogram of the numbers of pixels in each category as the SOON scans each block/sunspot group. Since the system is calibrated to relate pixels to millionths of the solar hemisphere, the SOON can actually measure the plage or flare area in each region.

The SOON can also electronically alert the observer to a flare while calculating its size, location, and brilliance. Flare brilliance is

determined by the brightest pixel in the region. Hence a brilliant flare may be mostly of normal intensity with only a few bright points. This is one of the system's limitations. Consistent setting of the thresholds for each brightness category (plage, normal flare, etc.) and determination of region boundaries are additional limitations.

The SOON is not limited to flare patrol. It is also equipped with a spectrograph and magnetometer. The spectrograph can be used for detailed study of the sun at wavelengths other than hydrogen-alpha. When used to create a full disk image at a single wavelength, this capability is termed the spectrohelioscope. The magnetograph makes use of the Zeeman Principle to determine the magnetic field strength and polarity of individual sunspots. This permits magnetic classification of a spot group and may, in other ways, prove invaluable in flare forecasting.

The SOON system is the most advanced solar patrol system yet developed. It meets the multiple tasks of routine, objective patrol and research while providing capability for expansion to incorporate new techniques. It does all this while providing a high-resolution image of the sun for analysis -- clearly revealing the importance of the instrumentation available to an observation.

3.5 Radio Telescopes and Resolution

The technical term for the "blur" referred to earlier is resolution. Resolution, usually measured in arc seconds, provides an evaluation of size of the smallest features that can be clearly identified. If two points of light are very close together, the resolution of an instrument will be the minimum separation at which the telescope will still permit the viewer to determine that two points, and not one, are present. This means that resolution is, ideally, a very small number. SOON resolution is about 0.75 arc second, for example. Conversion of angular measure to physical distance requires knowledge of the distance separating the object and observer. Mathematically, the resolution of a given telescope is proportional to the wavelength of radiation observed divided by the aperture of the telescope. Hence, improved resolution is available by viewing at shorter wavelengths (higher frequencies), using a larger telescope, or both.

Radio waves have wavelengths of approximately 100 meters compared to 1 micron ($10^{-6}$ meter) for optical wavelengths. Radio waves have, therefore, about $10^{8}$ times longer wavelengths. To achieve radio resolution similar to that available with an optical telescope would require a radio telescope with an aperture one hundred million ($10^{8}$) times greater than that of the comparable optical instrument. The most obvious consequence of this is that, while a SOON telescope can identify the source region (sunspot group) of a solar flare, a RSTN telescope can discern only that a radio burst originated from the general vicinity of the sun.

An instrument called an interferometer was designed to partially compensate for this lack of resolution. Radio interferometers (not to be confused with a sweep-frequency interferometer) consist of several small radio telescopes located a great distance apart and connected electronically. When all telescopes are turned on a given object, the resolution obtained is approximately equivalent to that of a single telescope with an aperture equal to the separation of the two most widely separated telescopes. Of course, the light gathering (or EM gathering) power of the interferometer is not equivalent to that of the larger instrument.

The sweep-frequency interferometer is intended to increase frequency - not spatial--resolution. A RSTN instrument scans the frequency range 25 - 80 MHz once each second and measures the intensity of radio emission at each frequency. This type of sweeping permits an observer to determine the variation in frequency with time for a low frequency radio burst. Different frequencies originate at different altitudes in the solar atmosphere. Sweep-frequency measurements determine the rate at which the burst source (say an ejected plasma cloud) changes altitude in the solar atmosphere.

3.6 Summary

Astronomical instruments are essentially extensions of the observer's senses coupled to various objective measuring devices. All telescopes (including the eye) can "see" equally far. Larger aperture telescopes can "see" fainter objects. Larger aperture telescopes also permit greater magnification to be used effectively and provide improved resolution. Radio telescopes are typically of poorer resolution than optical instruments due to the much longer wavelength of radio emissions. Some improvement results from the use of radio interferometers. Sweep frequency interferometers permit us to measure the speed of flare ejecta as it moves through the sun's atmosphere.

# CHAPTER 4

## SPACE

Everything that exists is termed the universe. Although we know very little about the universe because of the vast distances and times involved, we do know that it is apparently sparsely populated. Astronomers estimate that in excess of a billion galaxies are scattered throughout the visible universe. It seems reasonable to suppose that there are many more galaxies, but the universe is too young to permit us to see them. We estimate that the universe is nearly 15 billion years old. This means that only the light of objects within 15 billion light years (a light year is the distance covered by light in one year - about 6 trillion miles) has had time to reach us since the universe formed. Not surprisingly, the most distant observed galaxies are approximately 15 billion light years from the earth. Some recent work suggests that the age of the universe (and its size) may be only half that previously thought. This work, based on quasars, highlights our lack of knowledge of the process of galaxy formation and evolution and its importance.

A galaxy is a vast island of stars. Some are amorphous blobs; others are ellipsoids; and still others are pinwheels. Our home galaxy, the Milky Way, is one of the latter and contains perhaps 200 billion stars similar to our sun and a lot of gas and dust. The Milky Way is about 100,000 light years in diameter and 1,000 to 10,000 light years thick.

### 4.1 The Solar System

Located about 30,000 light years from the galactic nucleus, our sun is fairly typical in many ways. It would be indistinguishable were we looking back from a nearby galaxy (the nearest is about 100,000 light years away). Yet, the region of space within a few light hours of the sun is the only area of space about which we can claim any significant knowledge. This is the region of the planets known as the solar system.

The solar system is truly our celestial backyard. No man-made object has yet escaped its grasp. Pioneer 10, launched in the early 1970s, will be the first to achieve this feat, in 1987. The solar system technically extends beyond the outer planets to the heliopause. This marks the boundary of the sun's control. Inside the heliopause, the sun's magnetic field and the solar wind dominate. Outside, the interstellar medium is in control. The exact radius of the heliopause has yet to be determined, but it probably varies significantly with time and location.

Inside the heliopause, the sun dominates in many ways. It contains more than 95% of the solar system's mass. Roughly centered in the solar system, it is the major - but not the only - energy source. The planets actually exist in the tenuous outer atmosphere of the sun known as the solar wind. The remaining 5% of the system mass is tied up in the planets (mostly Jupiter), asteroids, comets, dust, and gas. Conversely, the planets have more than 90% of the system's angular momentum.

The planets seem to come in two types, each named for a prototype. The inner planets are small and rocky. They have few moons, little or no atmosphere, and high mean densities. These are the terrestrial planets. The Jovian planets roam the outer portion of the solar system. Located beyond the asteroid belt, these gas giants have many moons, low mean densities, and are

predominately gaseous. In fact, their composition is similar to that of the sun. The planets move in nearly circular orbits about the sun, in or near the plane of the earth's orbit about the sun (the ecliptic).

## 4.2 The Terrestrial Planets

Innermost of the planets, tiny Mercury orbits the sun in just 88 earth days and completes an axial rotation every 59 earth days. With a radius of about 2500 km, it is slightly larger than the earth's moon. No moons fill the night sky of Mercury. The planet is pockmarked by craters from over four billion years of meteorite bombardment. The absence of any atmosphere results in negligible erosion. The high mean density (5.4 $gm/cm^3$) suggests a small metallic core which, in turn, may account for the tiny planetary magnetic field. The absence of tectonic activity raises important questions about the formation of planetary magnetic fields.

Over 60 million miles from the sun, Venus is very nearly the earth's twin. Its radius is about 95% that of the earth, and its mean density is about the same as that of Mercury and the earth. Nonetheless, it has no measurable magnetic field. A Venusian year lasts about 225 earth days. Enshrouded by clouds, Venus' surface is invisible to optical telescopes. The clouds are apparently composed of sulfuric acid and move at 200-500 mph. The dense cloud cover traps considerable heat. This, combined with Venus' proximity to the sun (Venus receives about twice the solar radiation received by the earth), results in a surface temperature of nearly 900°F--sufficient to melt lead. A surface pressure of nearly 90 atmospheres is sufficient to squash all but the sturdiest of satellites (unless, of course, they are vented to the atmosphere). Mapped by radar from space and injection probes above the surface, Venus is slowly revealing its secrets. Three continents and little or no liquid water are present. Intense lightening storms are common near the continents. There may be active volcanoes and some tectonic activity.

The most tectonically active planet in the solar system is the earth. Orbiting the sun at a distance of 8 1/3 light minutes (93 million miles), the earth is the largest of the terrestrial planets. Moreover, it has the largest moon by comparison to the planet size of any solar system planet. This has lead some astronomers to call the earth-moon system a double planet.

The moon is approximately 1 1/4 light seconds from the earth. The moon and sun account for terrestrial tides, which may be slowing the earth's rotation through friction. Meanwhile, the earth and moon are slowly drifting apart (at a rate of about 5 cm/year). This separation is a result of conservation of angular momentum and, possibly, a secular change in the gravitational constant. Two types of terrain are present on the lunar surface. The lighter colored lunar highlands contain an abundance of aluminum and volatile gases. The highlands are heavily cratered and probably little changed since the moon's formation over four billion years ago. Approximately 20 dark, circular basins are less heavily cratered. They are called maria and are probably ancient (3 billion years) lava flows resulting from meteorite strikes. They are rich in iron and titanium. Although 18 are visible on the near side of the moon, only 2 are known on the far side. The lunar crustal thickness also seems to vary from about 60 km on the near side to 120 km on the far side. There is evidence of a primordial lunar magnetic field, but none exists today.

Mars, the most distant of the terrestrial planets, is larger than Mercury but considerably smaller than Venus. Two known moons orbit the red planet, and a tenuous atmosphere accounts for a surface pressure of as much as 7 millibars. With a mean density of about 4 $gm/cm^3$, Mars is markedly less dense than the other terrestrial planets. Polar caps, recently active volcanoes, and an axial tilt and rotation rate similar to that of the earth drive a feeble weather pattern. Mars has apparently undergone ice ages like the earth, and it experiences vast seasonal dust storms. A few clouds have been observed near volcanoes, and some liquid water is apparently tied up in the polar caps and permafrost.

Although Mars is small, some of its surface features are truly vast. It seems to possess the largest volcano in the solar system - Olympus Mons. These large volcanoes (that largest are as large as the state of Arizona) are probably a consequence of the absence of plate tectonics (continental drift) which, on earth, accounts for island chains and many small volcanoes. On Mars, hot spots in the mantle can generate large volcanoes, since the overlying plate is immobile. There are also signs that, at some time in the past, a lot more liquid (water?) was present on the Martian surface than is now apparent. Where it went is unknown.

## 4.3 The Jovian Planets

Jupiter, prototype of the gas giants, is the largest of the known planets. Although it is only a thousandth the size of the sun, it is larger than all of the other planets combined. The planet is predominately hydrogen and helium, resulting in a mean density of about 1.6 $gm/cm^3$. Nonetheless, its atmospheric mass is thought to produce interior pressures in excess of a million atmospheres. At these pressures, hydrogen is thought to be a conductor, and internal temperatures probably exceed 30,000°K. This may account for the large, offset magnetic field of the planet. (The magnetic axis of the planet is not parallel to its rotational axis; nor do the two intersect.)

Jupiter rotates rapidly - once every ten to eleven hours. In addition to the planetary magnetic field, this rotation produces a marked equatorial bulge. The planet's polar diameter is almost 20,000 km less than its equatorial diameter. Rapid rotation probably also accounts for planetary meteorology. The great red spot and many, smaller brown and white spots are actually storms. Intense electrical storms were observed by the Voyager spacecraft in the nighttime hemisphere.

Spacecraft exploration of Jupiter has also revealed a delicate ring structure and numerous moons. Since Jupiter emits about twice as much energy as it receives from the sun, the Jovian system is often studied as a miniature solar system. One of the four large (Galilean) moons, Io, is known to have active volcanoes. It orbits Jupiter well inside the trapped radiation belts and is probably heated by tidal forces due to Jupiter's proximity.

Saturn shares many of Jupiter's features - rings, moons, energy emission, and gaseous composition. Its smaller size, slightly lower mean density, and greater distance from the sun probably explain why its atmosphere is less colorful than Jupiter's. Many gases freeze out at lower altitudes, and the

planet may be enveloped in haze. Like Jupiter, it emits nearly twice as much energy as it receives from the sun. The Saturnian ring system is much more imposing than that of Jupiter. Like Jupiter's, it consists of many tiny particles and is thin. The rings actually consist of numerous ringlets separated by relatively empty regions (divisions). Particles in these divisions would have orbital periods synchronized with one or more moons. The resulting tidal forces keep the divisions relatively empty. Saturn's mean density is less than that of water.

Lying beyond Saturn are Uranus and Neptune. These planets are both smaller than Saturn, yet they are each nearly 15 times the earth's mass. Primarily gaseous, they have roughly twice the mean density of Jupiter and Saturn. Uranus is known to have at least five moons and a ring system nearly as tenuous as that of Jupiter. The rings and moons orbit in the planet's equatorial plane which is roughly perpendicular to Uranus' orbit plane about the sun. Hence, Uranus' poles alternately point directly towards or away from the sun. Neptune has so far revealed only three moons, but more probably exist. It, too, is thought to have rings. Until 1999, it will be the most distant planet from the sun.

Smallest of the known planets, Pluto is usually the most distant. Its highly elliptical orbit carries it inside the orbit of Neptune for about 25 of every 250 years. In 1978, a moon was discovered in orbit about Pluto. Named Charon, the moon provided the first accurate determination of Pluto's size and mass. There is now some question as to Pluto's status as a planet. It is considerably smaller than the earth's moon and only about three thousandths the mass of the earth. Its mean density suggests that Pluto is primarily frozen gas. Sunlight requires nearly six hours to reach Pluto, where the sun is little more than a bright star in an eternally dark sky.

## 4.4 Interplanetary Matter

Planets and moons are not alone beyond the sun. Between the orbits of Mars and Jupiter, more than 75,000 tiny objects called asteroids circle the sun. Each is a tiny world, though even the largest (Ceres, of 400 km radius) is barren and airless. Some asteroids have tiny moons, but none is truly a planet. In fact, only Vesta shows signs of previous heating (which would be expected in a planet). The asteroid belt is itself segmented into inner and outer belts, with composition being the primary difference. Inner belt objects are rocky, while carbonaceous objects populate the outer asteriod belt (about 270 million miles radius orbit compared to 200 million miles for the inner belt).

Not all asteroids are confined to the asteriod belt. Some, like the Trojans (asteroids at Jupiter's distance from the sun but displaced 60° ahead of and behind Jupiter), are trapped by the gravitational fields of the planets and move with their parent planets. Still others, the Apollo asteroids, occasionally cross the earth's orbit. This group of asteroids is considered a possible source of <u>meteors</u>.

The earth accumulates between 5-50 tons of meteor dust each day. Most of it is deposited near 100km in the atmosphere in the form of metallic ions. They are thought to play an important role in maintaining the lower ionosphere. The heating and ionization which results from their passage

through the upper atmosphere also has a significant impact on HF, VHF, and UHF radio and radar systems. Meteors are actually tiny dust grains which burn due to friction in the atmosphere. Most are less than a gram in mass. Since the brightness of a meteor is proportional to the square of its velocity times its mass, higher speed meteors are generally brighter. The brightest meteors are generally visible in the midnight-to-dawn sector where the earth's speed in orbit is added to that of the meteor. Evening sector meteors must catch up with the earth in orbit. (The earth orbits the sun counter-clockwise when viewed from above or north of the ecliptic. It rotates counter-clockwise once each 24 hours when viewed from this same vantage point.)

If the meteor is not completely destroyed by its transit through the atmosphere, it may reach the surface as a meteorite. Most known meteorites are metallic, probably because people finding the more common stones cannot identify them as meteorites. Meteorites have provided the first direct evidence that the complex molecules necessary to the formation of life can exist in space. This discovery has lead to theories suggesting that the earth was seeded from space. In space, the grain is called a meteoroid.

Although more than 25 million meteors are thought to occur each day, the counting rates at any one location are typically small (1-5 per hour). The occurrence of fireballs (very bright meteors which leave a bright tail or train in their wake) is even more uncommon. The incidence of both increases during meteor showers. During a shower, counting rates of 30-40 per hour are not uncommon at some sites. Showers may result in intense, localized atmospheric heating and strong electrical currents. Such activity can have a marked impact on ionospheric operations -- sometimes over a wide area. Meteor showers are typically named for the constellation from which the meteors appear to originate (their "radiant"). Showers are thought to be a consequence of the earth's passage through a stream of meteoroids orbiting the sun (actually, like a solar ring which crosses the earth's orbit). Consequently, showers recur with fair regularity. These streams of meteoroids have, in many instances, been associated with comets and may be debris remaining from a cometary breakup.

Most comets are thought to be small (few kilometers in radius) rocky objects permeated with frozen gases. They are, for the most part, in highly elliptical orbits obout the sun. Some short period comets have periods as short as a few years, while the longer period comets may require hundreds of years to complete one orbit of the sun. The tiny heart of the comet is known as the nucleus. As it approaches the sun, it heats, and the frozen gases are boiled off (inside the orbit of Mars, typically) much like steam. This outgassing creates a thin cloud (the coma) thousands of kilometers in diameter about the nucleus. The pressure of the solar wind combined with the orbital motion of the comet results in the formation of a long ($10^6$ to $10^8$km) tail. There are, in fact, two distinct tails. One, the dust tail, curves gracefully away from the sun. It responds to photon pressure and orbital motion. The second tail points radially away from the sun. This is the ion tail and is a consequence of the interplanetary magnetic field transported outwards by the solar wind. The effect is like a magnetic "rake" pulling ions out of the coma. Comet tails provided the first evidence of the nature and magnetic structure of the solar wind. Today, they are still used for remote measurements of the solar wind.

## 4.5 Summary

The solar system contains at least one star and, perhaps, nine planets. Exact numbers depend on the definitions chosen for "planet" and "star". The planets are loosely grouped into the terrestrial, or inner, and the Jovian, or gas giant, planets. Among the Jovian planets, at least two -- Jupiter and Saturn - are known to emit more energy than they receive from the sun. The discovery of rings about Jupiter and Uranus raises the question of why all planets don't have some sort of ring structure. Considerable research is already underway in the areas of comparative geology (planetology) and meteorology. From the distant planets, we are beginning to learn how our home planet functions. Even the tenuous interplanetary medium has proven significant. Meteors and comets are both geophysically important. The further we look, the greater are the mysteries which we uncover. To paraphrase Albert Einstein, as the circle of light expands, so, too, does the circumference of darkness about it.

# CHAPTER 5

## STELLAR EVOLUTION

Only in its proximity to the earth is the sun unique. It is but one of roughly 200 billion stars which populate a fairly average spiral galaxy--the Milky Way. There are many stars which are as much larger than the sun as the sun is larger than the earth. Similarly, there are stars which are smaller than the earth. The sun's proximity is important not only for its life-giving energy but also because it permits us the luxury of detailed analysis. A stellar model for a star like the sun serves as a benchmark for our models of all other stars.

Inferring the life history of a star is still not a simple task. Since stars live for such long times (millions of years at a minimum), they change very little within a few hundred years. We must be content with taking a snap-shot of many stars in various stages of evolution and attempting to guess which is where in its evolution. Such a process has numerous pitfalls. One way to begin is by cataloging the properties of the stars.

### 5.1 Stellar Properties

Perhaps the first question one is tempted to ask is how we can be certain that stars ever change. Assuming that what we know of science is moderately accurate and generally applicable, a simple observation suffices. Stars shine. This implies that something inside the star (or possibly external? to it) is being converted into the energy which the stars so readily radiate. If evolution is taken to mean change, then stars must be evolving if they shine. This, however, is not the only information revealed by our snapshot.

Most stars are remarkably similar in composition. The bulk (75% to 80%) of a star is hydrogen, and helium follows closely (20% or more). Astronomers loosely group all other elements under the title of "metals." The metal content of the stars ranges from far less than 1% to as much as 5% and may be related to the age of the star. Content also seems to be related to a star's location in the galaxy, but this is probably due to parts of the galaxy forming before other parts. Since most of the heavy elements are thought to result from nuclear fusion inside massive stars, it is probable that several generations of stars presently coexist in the Milky Way (and other observed galaxies).

A star's temperature is generally taken as its effective (or <u>surface</u>) temperature and is based on the blackbody nature of its radiation. The surface temperature establishes the character of the star's emission, while the internal temperatures can only be inferred by use of mathematical models. Stellar temperatures seem to range from about 3000°K to 35,000°K. There are some stars which seem to have temperatures of nearly 100,000°K, but this may be a transient phenomenon. For reference, the sun's surface temperature is about 6000°K. Surface temperature establishes the "color" of a star, since most stars are moderately good blackbodies.

Many stellar parameters are most conveniently stated in terms of solar units. Stellar radii range from about .01 solar radii to about 900 solar radii. A 900 solar radius star located in the sun's position would extend to the orbit of Jupiter. The luminosity of a star is a function of its size and temperature. Stellar luminosities range from $10^{-4}$ to $10^{4}$ solar luminosities. Stellar mass, probably the most important of all stellar properties, is the most difficult to measure accurately. Except for the sun, accurate masses are known only for stars located in certain multiple star systems. This would seem not to be a significant constraint, since more than half of all stars occur in multiple systems. Nonetheless, these masses may not be directly applicable to similar stars found without companions. There is strong evidence that close companion stars can significantly influence a star's evolution. This is particularly significant for companions of widely different masses. Stellar masses seem to have a moderately limited range: 0.1-60 solar masses. This is small by comparison to the range of size and luminosity found among stars.

## 5.2 The Hertzsprung-Russell Diagram

In the early part of this century, astronomers were attempting to make sense out of the numerous observations available. Two--Hertzsprung and Russell--independently chose to graph luminosity against temperature (or some other similar measure, such as color). The result, know as the H-R or color-magnitude diagram (Figure 5.1), was successful beyond the wildest dreams of its conceivers.

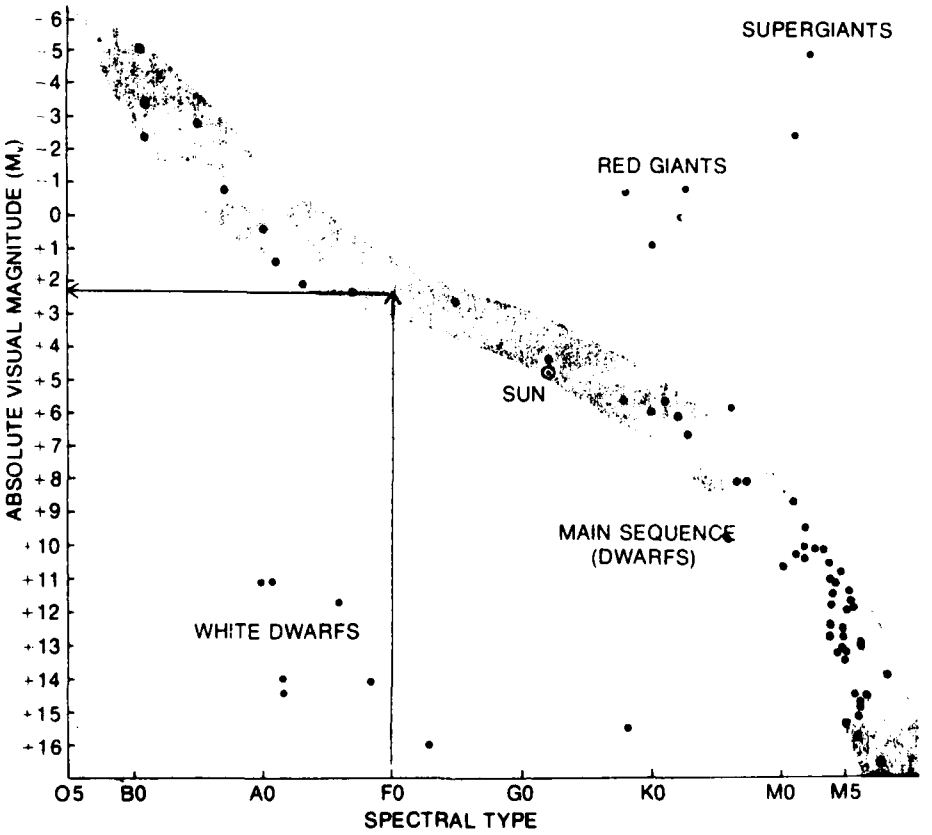



**Figure 5.1 Color Magnitude Diagram. Luminosity increases upwards. Temperature increases to the left (from Pasachoff, 1977).**

Several important observations can be made. More than 90% of all stars lie along the portion of the plot known as the main sequence. The upper right of the diagram, containing less than 1% of all stars, is termed the gaint region. About 10% of all stars, those known as white dwarfs, are located in the lower left of the diagram. The terms giant and dwarf relate to the relative size of the star.

Some giants have temperatures equal to those of main sequence stars, but they are much more luminous. Since luminosity is dependent on temperature and size, it must be that some stars are larger than others. Moreover, the effect on luminosity of even a considerable size difference is equivalent to only a slight variation in temperature. This can be seen from looking at the equation defining luminosity, L, in terms of the star's temperature, T, and radius, R:

$$L = (\text{constant})\ R^2T^4.$$

It should be noticed that the arrangement of stars in the H-R Diagram admits at least two possible explanations. It may be that stars spend 90% of their life on the main sequence and about 10% as white dwarfs. Conversely, it may be that 90% of all stars are born main sequence stars and never change; 10% are formed and die as white dwarfs; and less that 1% come in and go out as giants. To differentiate between these two explanations we must develop a theory to explain how a star produces the energy it radiates. Such an evolutionary theory should help us identify which of the above answers to the H-R Diagram riddle is the correct one.

## 5.3 Stellar Energy Sources

In order to conceive an evolutionary theory, we must first determine the origin of stellar energy. A viable energy source must have certain longevity traits in addition to providing a given level of output. We can establish suitable criteria in both these areas by considering the earth-sun system. We know that the earth and moon have been solid for approximately four billion years. Consequently, the sun must have been moderately stable for about four billion years. Given total mass to energy conversion, the sun is producing energy at a rate of $4 \times 10^9$ kg/sec. If this rate of energy production has been relatively level (as we suspect it has) since the sun's formation, then the sun will have released about $2 \times 10^{-4}$ of its mass as energy. We can now evaluate possible energy sources in light of their capability to fulfill these requirements.

Burning of fossil fuels was early considered as a possible stellar energy source. Calculations suggest that even if the sun were solid coal, the rate of energy release would be insufficient to produce the observed emission. In four billion years, only $5 \times 10^{-10}$ of the sun's mass could be converted into energy.

The release of gravitational potential energy through contraction or cooling seems somewhat more viable, though not a long-term energy source. To produce the currently observed rate of energy release, contraction could last only about 30 million years. Since there is no evidence for significant size changes in the past hundred million years, it seems unlikely that contraction has played a recent role in solar energy production.

Of the known energy sources, nuclear energy is the most promising stellar fuel. Two types of nuclear reactions can produce energy. The splitting, or fission of elements heavier than iron will release energy. Likewise, the combining (fusion) of elements lighter than iron releases energy. Neither fission nor fusion will release energy from iron. If permitted to occur throughout the sun for four billion years, fission would by now have released $5 \times 10^{-4}$ of the sun's mass as energy. Under the same constraints, fusion would release .01 solar masses as energy. Clearly, both are sufficient, although fission is just barely so.

To differentiate between the two we must look deeper. Since the primary constituents of the sun are hydrogen and helium, only a fraction of the sun can be undergoing fission at any given time (since only a tiny fraction of the sun will be made up of elements heavier than iron). Conversely, fusion would be possible throughout the star. Moreover, no energy production is observed at or near the solar surface. Given the already marginal performance of fission, it seems likely that fusion is the primary energy source for sun-like stars. Moreover, fossil fuels have probably never been a significant stellar energy source, because they are relatively rare inside stars. Contraction, on the other hand, may occasionally be significant. It seems particularly probable during the early and latter stages of stellar evolution.

## 5.4 Star Birth

The space between the stars is not empty. It is filled by a tenuous mixture of gas and dust. In some parts of the galaxy, this interstellar medium forms into clumps. These clumps are known as nebula and are thought to be the birthplaces of stars. The interstellar medium contains varying percentages of metals, depending on location. The metal concentration probably results from the demise of massive stars. As we shall see, very massive stars produce heavy elements through fusion and end their evolution in disruptive explosions. These supernova hurl 10% - 90% of their mass back into the interstellar medium, thereby enriching the metal content of the ambient gas. This gas will eventually be incorporated into a new generation of stars.

Some of the gas clouds are quite dense. In such clouds, the central temperature will be only a few degrees above absolute zero. Complex molecules can form in such an environment, for they are shielded from the destructive effects of the radiation from nearby stars. For some reason, the cloud will begin to contract, or collapse, in free fall. Several mechanisms have been suggested to trigger these collapses and include photon pressure from nearby stars, shockwaves from a nearby supernova, or the spiral density wave thought to maintain the arm structure of the galaxy. These mechanisms share a common fault: they require stars in order to create more stars.

Once a cloud begins to collapse, it will continue to contract in free fall. In the process, considerable energy is released. Since the gas is so cold, little of the energy is trapped. Most of the atoms are in the ground state and are unable to absorb the emitted photons. Rather, the energy expands freely into space. Eventually, the gas will start to heat. When this occurs, the cloud is becoming opaque to the radiation. The opacity is provided by atoms with electrons so arranged as to be able to absorb the photons being produced.

Working against this new-found opacity, the radiation pressure of the released energy acts to slow the free fall collapse of the protostar. Simultaneously, the interior temperature of the cloud rises. At this stage, the protostar is surrounded by a dense, cool envelope of gas and dust. It is visible as a strong source of radio and infrared emission. The speed with which it contracts to an equilibrium point and the temperatures attained at equilibrium are functions of the star's mass. Eventually, the central temperature of the star will be sufficient to trigger nuclear fusion. At this point, the star settles down into what will become the longest, most stable stage in its evolution. It is now a main sequence star.

## 5.5 Main Sequence Evolution

A main sequence star undergoes nuclear fusion in its core. The outer envelope of the star shields the core and acts to restrain the nuclear reaction. The star's main sequence lifetime is approximately equal to

$$\left[\frac{(\text{mass of the sun})^2}{(\text{mass of the star})^2}\right] \times 10^{10} \text{ years.}$$

Although a more massive star has more fuel, it consumes its fuel much more rapidly. All stars might live longer if they could make use of all their hydrogen. Typically, only about 10% of the star's mass, located in the stellar core, will undergo fusion. Despite the vast amount of potential fuel, it is inaccessible to the stellar core.

Changes in the stellar core cause a gradual change in the star during its main sequence lifetime. The build up of helium "ash" in the core changes the average mass of material found there. In order to maintain the constant pressure required to support the overlying layers, the core temperature must increase. An increase in temperature produces an increase in the rate of fusion and the rate of energy production. The star's luminosity grows.

The star's mass is thought to determine much of its internal structure. Stars greater than about 1 1/4 solar masses have convective cores, radiative envelopes, and are rapid rotators. The conditions are reversed for less massive stars, where radiative cores and convective envelopes combine with slow rotation. Moreover, the more massive stars seem to have smaller cores than do the less massive objects.

## 5.6 Post Main Sequence Evolution

As core hydrogen exhaustion approaches, structural changes are more rapid and discernable. The cutoff of core fusion marks the onset of internal collapse. This releases gravitational energy in an attempt to maintain sufficient pressure to support the star's overlying layers. This rapid release of energy into the star's hydrogen-rich envelope eventually results in the explosive ignition (fusion) of a shell of hydrogen surrounding the core. This sudden burst of energy in addition to that released by the core collapse results in a rapid expansion of the stellar envelope. As the envelope expands, it cools. The star's size increases, and it "moves" (by changing size and temperature) into the giant region of the H-R Diagram. During this phase of its evolution, the sun will envelop Mercury, Venus, and the Earth.

The core continues to collapse, and the star expands until the core temperature and pressure are sufficient to initiate helium fusion. The more massive the star, the more quickly it will achieve core helium fusion. The sun will require almost a billion years between hydrogen exhaustion and the onset of helium fusion. It will be a red giant during this period.

Many stars are markedly unstable during the red giant phase. Size fluctuations are common, as are strong, dense stellar winds. The stellar winds may carry away as much as 20% of the mass of a sun-like star during this billion years. Neutrinos escaping from the hydrogen fusion shell may carry away several percent of the energy released in and around the core without increasing the star's internal pressure. Such losses further delay the onset of helium fusion. Some stars, such as Mira in the constellation of Cetus, are thought to lose so much energy by this process that they are never able to initiate core fusion - or do so only periodically. This periodic core fusion may account for the pulsating variability of Mira type variables.

The core helium fusion episode is much shorter than was the core hydrogen fusion. More massive stars do it more quickly. Following core helium exhaustion, the cycle of core collapse and explosive shell source ignition begins again. Sufficiently massive stars will continue these cycles until an iron core exists. Less massive stars will be unable to create sufficiently high core temperatures or pressures to fuse carbon or other intermediate elements. When core fusion is no longer possible, the star moves into the final stages of its evolution. The exact result is again a function of the star's mass.

## 5.7 Star Death

Stars less than about 1 1/2 solar masses will probably be unable to develop iron cores. Following the final core fusion phase, these stars will again expand into red giants. The onset of shell fusion sources serves to softly "kick off" the outer shell of the star. The shell expands slowly outwards, aglow in the emission of the remaining stellar object. This entire object is known as a planetary nebula; the core star is known as a white dwarf. Powered solely by residual thermal energy, the star's temperature is typically near 20,000°K. It is nearly isothermal with a density of approximately 15 tons/cm$^3$. The star is about the size of the earth and slowly cooling. It will, after billions of years of slow cooling, become a black dwarf (not to be confused with a black hole).

A more massive star exits somewhat more dramatically. Following the creation of an iron core, the star attempts to fuse iron. This is an endothermic reaction. The effect is similar to turning on a refrigerator in the core of the star, and soon the core is in nearly free-fall collapse. Iron fusion continues, but it does so at the expense of stopping the core contraction. The core continues to fall inward, releasing considerable potential energy in the process. Neutrino emission probably also carries away considerable energy at this time. The end result is an explosive ignition of fusion throughout most of the star outside the core. The star essentially blows itself apart, with the explosion accelerating the core inwards. Between

10% - 90% of the star's mass is lost within a few minutes. During this period, the star may outshine an entire galaxy. This cataclysmic event is known as a supernova.

The object remaining after the supernova depends on how much mass remains. If less than about 1 1/2 solar masses are left, a white dwarf results. If the remaining core has a mass of 1 1/2 - 2 1/2 solar masses, the core collapse is sufficient to force the atomic electrons into the nuclear protons. The collapse stops with the resulting neutrons pressing, physically, against other neutrons. The object is called a neutron star and may be 10-20 miles in diameter. Its density will exceed $10^{14}gm/cm^3$, and the spin rate may approach 100/sec. Intense magnetic fields are also likely. Such objects are thought to be at the hearts of pulsars. Perhaps the best known pulsar is found inside the supernova remnant known as the Crab Nebula in the constellation of Taurus.

For stars with a residual mass greater than 2 1/2 solar masses, even neutrons pushing against neutrons are thought to be incapable of supporting the overlying mass. The star literally collapses into itself until only its gravitational influence remains. Such an object has, theoretically, no radius--it is a singularity. We can define a surface known as the event horizon on which the escape velocity equals the speed of light. Inside the event horizon, not even light can escape the singularity, because light is <u>too slow</u>! An object with the mass of the earth would have an event horizon of radius 1 cm if it were compressed into a singularity. Since not even light can escape from such objects, they are called black holes. Although none have yet been directly observed, several are thought to exist. These are "found" by their gravitational effects on companion stars. Cygnus X-1 (the first x-ray source discovered in the constellation Cygnus) was the first such object. There have since been others. The physics describing such objects is completely unknown. What we do know is that as the mass of the black hole increases, so too does the radius of its event horizon. In fact, if we determine the density of a black hole using the size of its event horizon we find that the greater the mass, the <u>lower</u> the required density. Current observations of the size and density of the known universe imply that it is a black hole. So far as we know, no light has ever escaped the universe, so it does seem to meet all the requirements for being a black hole!

## 5.8 Summary

Modeling the evolution of a star is difficult because of its slow rate of change by comparison to a human lifetime. The vast distances to even the nearby stars effectively preclude direct observation. As a result, analysis of the sun is crucial, for it establishes our only benchmark in creating model stars. We believe that stars begin as diffuse gas clouds, spend considerable time fusing hydrogen to helium in their cores, and eventually partially disrupt themselves. The rate of evolution and the sequence of changes are uniquely determined by a single parameter--the mass of the star. Composition has some effect, but it varies little among observed stars. Most such variations are a consequence of earlier generations of stars enhancing the heavy element abundance of the interstellar medium with their death. The end product of stellar evolution is also a function of a mass and may be a white dwarf, neutron star, or black hole.

# CHAPTER 6

## POSITIONING THE SOLAR SYSTEM

We have seen that the angles between fields and particles in motion strongly influence the outcome of the interaction. It should not be surprising, then, that the magnitude of certain geophysical effects of solar activity vary with the time of year. The earth-sun system varies its orientation, and the orientation of the respective magnetic fields, continuously through the year. A large flare occurring in September may have very different effects from a similar flare in June. In order to understand these differences, we must be able to describe the motions and geometry of the system.

### 6.1 Kepler's Laws

Kepler's Laws are not really laws, but rather a generalized series of statements describing the motions of two objects with respect to one another. Only gravity is considered; electromagnetic interactions are disregarded.

The first of Kepler's Law's specifies that all orbits of one body about another are ellipses. Today, we know that this statement is not strictly true. An elliptical orbit is always possible, but so are other orbits--circles, parabolas, or hyperbolas. Different orbits can be obtained by altering the initial direction and energy of one body with respect to the other. A more general statement of Kepler's First Law is that all orbits are conic sections. This means that escape, or open orbits are possible.

Kepler's Second, or Areal Law relates speed in orbit to position in orbit. Shown graphically in Figure 6.1, this law requires that two objects in orbit about each other move so that the imaginary line connecting them sweeps out equal areas in space in equal times. One direct consequence of this deals with the length of the seasons. The period of time from the September equinox

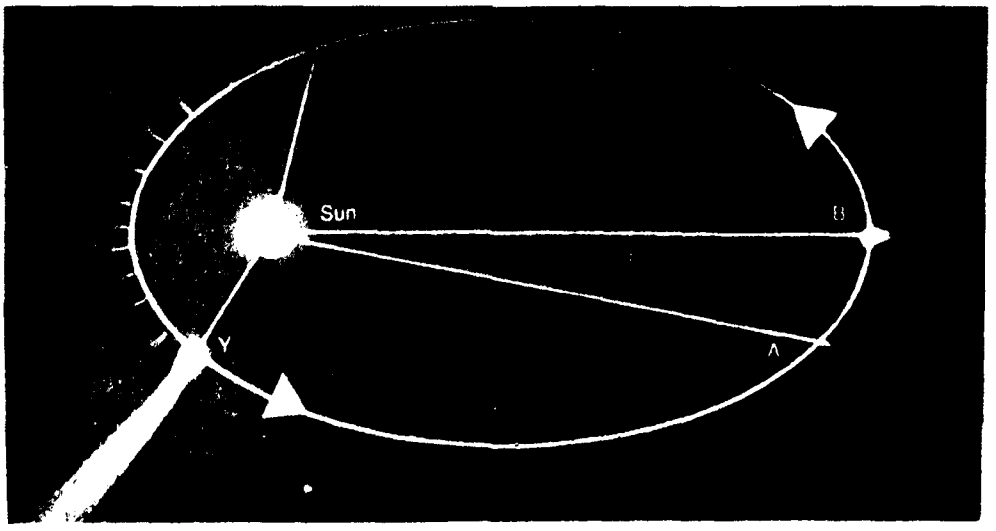



Figure 6.1 Equal areas swept out in equal times means a higher speed at closest approach.

to the March equinox is several days shorter than the time from the March equinox to the September equinox. By definition, the distance covered during each of these periods is equal. This means that the earth must move more rapidly in its orbit during northern hemisphere winter than during that hemisphere's summer. This results from the earth's passing nearest the sun in December - January of each year. If the sun were fainter (so that we could see other stars during the day), we would see that the sun <u>appears</u> to move most rapidly among the background stars in December and January. Remember the earth is doing most of the moving. The sun only seems to move.

Speed in orbit and size of orbit are brought together in the Third or Harmonic Law. This law, like the Second, assumes closed orbits. It can be best stated mathematically:

$$P^2 = (\text{constant})\ a^3,$$

where $P$ = sidereal period and
$a$ = semi-major axis of orbit.

(For an ellipse, the semi-major axis is 1/2 the length of the long axis; for a circle, it is the radius.) The constant accounts for, among other things, the difference in units. Newton later revised this equation by evaluating the constant. The revised form is given by

$$P^2 = \frac{4\pi^2 a^3}{G\,(m_1 + m_2)},$$

where $G = 6.67 \times 10^{-8}$ dyne $cm^2/gm^2$ and $(m_1 + m_2)$ is the mass of the two bodies. Of course, Einstein's work further revised this analysis, but Kepler's original work is sufficiently accurate for our purpose.

Geosynchronous and geostationary satellites are similar in that both have orbital periods (time to complete one orbit) of 24 hours. We can use Kepler's Third Law to determine the altitude above the center of the earth (why the <u>center</u> and not the surface, and is it really about the center of the earth?) at which these vehicles must orbit in order to sustain a 24 hour period. Note that the altitude is essentially the semi-major axis of the orbit (these orbits are nearly circular). If we solve the equation for a, we obtain

$$a = \frac{P^2 G\,(m_1 + m_2)}{4\pi^2}^{1/3}$$

$(m_1 + m_2)$ is the sum of the masses of the earth and the satellite. This is essentially the mass of the earth, $6 \times 10^{27}$ grams. Likewise, P = 24 hours = 86,400 seconds. Substituting and punching a few numbers into the calculator yields

$$a = 4.24 \times 10^9 \text{cm},$$
$$= 4.24 \times 10^4 \text{km}.$$

The earth's radius is $6.4 \times 10^8$cm, so
$a = 6.6$ earth radii.

It is important to note that the preceding discussion has tacitly assumed closed orbits. Not all orbits are closed (for example, Pioneer 10's orbit about the sun is an escape orbit). Kepler's Laws are also true for these orbits, but they yield significantly less information.

## 6.2 Earth Motions

The earth is constantly moving in response to many different forces. The net motion is extremely complicated, but two components of this motion can be easily identified--rotation and revolution.

The earth rotates once every 24 hours, turning from west to east about an axis passing through the north and south poles. Rotation causes the sun to appear to rise in the east and move westward during the course of the day. An extension of the earth's rotational axis into space defines two points--the celestial poles. The north star, Polaris, is close to the north celestial pole. The varying position of the sun with respect to a fixed spot on the earth results in marked changes, known as diurnal variations, in the ionosphere. Moreover, rotation transfers considerable angular momentum into the earth's atmosphere and magnetosphere which helps shape these structures.

Each year, the earth revolves around the sun. Revolution accounts for different stars being visible in the nighttime sky at different times of the year. Specifically, the earth's revolution causes the sun to apparently shift its position with respect to the background stars by about $1^{o}$ per day ($360^{o}/365$ 1/4 days). The shift is from west to east (opposite the apparent solar motion resulting from the earth's rotation) and causes the sun to rise about 4 minutes later each day. In the time required for the earth to complete a $360^{o}$ rotation (sidereal period), the earth will have revolved about $1^{o}$ in its orbit. Hence, sunrise to sunrise requires approximately $361^{o}$ degrees of rotation by the earth (the time for this is known as the synodic period).

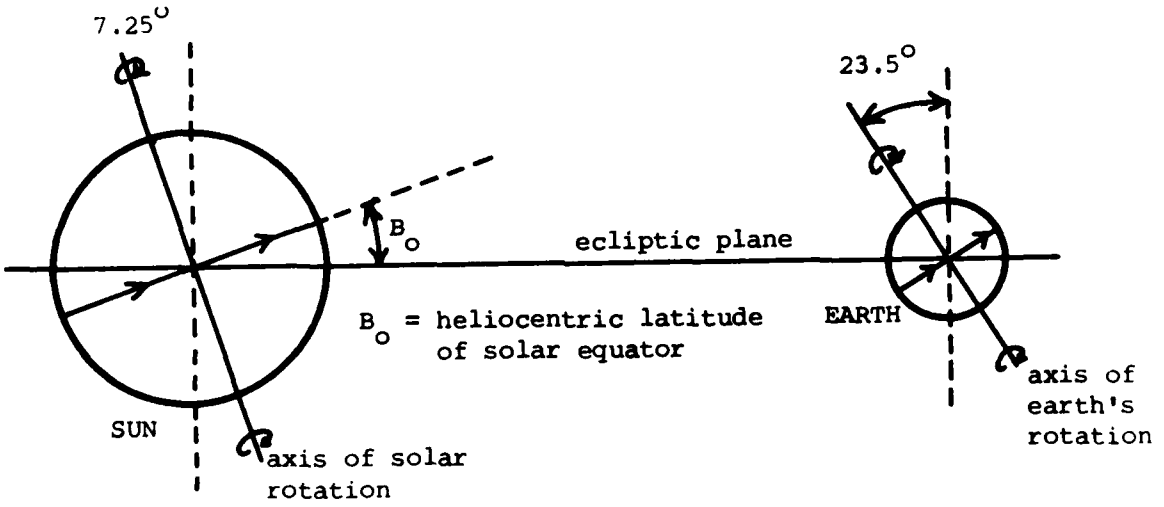



Figure 6.2 Relationship of earth and sun rotational axes to ecliptic.

From our discussion of Kepler's Laws, we saw that the earth passes perihelion (closest approach to the sun; perigee is closest aproach to the earth) during northern hemispheric winter and aphelion in that hemisphere's summer. The earth's orbit about the sun defines a plane called the ecliptic. The earth's rotational axis is not perpendicular to the ecliptic, but it is inclined about 23 1/2° away from perpendicular. This is termed the obliquity of the ecliptic. (See Figure 6.2)

Since the earth's equator does not lie in the ecliptic, the sun will appear to move both north and south of the equator during each year. This migration of the sun is responsible for seasonal variations. The equinox are defined as the two instants (approximately six months apart) when the sun is in the earth's equatorial plane (would appear directly overhead at midday for an observer at the equator). The solstices are the points in time when the sun reaches its extreme distance (23 1/2°) north and south of the earth's equator. Revolution of the earth causes the sun to appear to move along the ecliptic, from west to east. (As seen from north of the ecliptic, the earth revolves about the sun in a counterclockwise sense.) The earth's rotation causes the sun to appear to move east to west parallel to the earth's equator (seen from above the earth's north pole, the earth rotates in a counterclockwise sense). Be very careful not to confuse the ecliptic with the equator. We have seen that the earth's equator does not coincide with the ecliptic. Shortly, we shall find the same to be true for the sun.

## 6.3 Solar Motions

The net motions of both the earth and the sun result from complicated sets of motions. Of course, the sun and earth actually revolve around their common center of mass. Since this lies inside the surface of the sun, it is simpler to assume that only the earth moves, and that the sun is the center of the system.

While we may neglect the revolutionary period of the sun (it didn't have many patriots anyway), we must consider its rotation. Viewed from above the solar north pole the rotation is counterclockwise. Near its equator, which is inclined about 7° to the ecliptic, the sun rotates at about 2 km/sec. This is slow for stars, many of which spin at rates of 100-200 km/sec. A 360° rotation of the sun requires between 25 and 32 days, depending on latitude.

The wide range in rotational period is not due to observational uncertainties. Rather, it is a result of the gaseous nature of the sun. The sun is not a rigid body, so the equator rotates much faster than the regions near the poles. The time for a 360° rotation (sidereal period) varies from about 24.9 days at the equator to approximately 31.5 days near the poles. Hence, the latitude of a sunspot group determines how long it will be visible and how long it will be hidden on the far side of the sun.

For rough calculations, we usually assume that the sun rotates once in 27 days. This assumption is based on two items. First, most sunspot groups form at low latitudes (north or south of the solar equator), so we choose the area of most interest to select our rotation period. The average sidereal rotation period for low latitudes is 25.4 days. Why 27 days? The second item involves the earth's activities during this 25.4 days. During this period, the earth will have moved about 25° in its orbit about the sun. That means that,

although a given spot group has rotated 360° in 25.4 days, it will not yet appear to be back where it started as seen from earth. It will need to rotate an additional 25° (at about 13°/day) in order to return to its apparent starting point. The time required for this 385° (360° + 25°) rotation is called the synodic period (for 17° heliographic latitude in this case) and can be calculated. At a rate of 360°/25.4 days, this latitude will take 127/72 days = 1.76 day to rotate 25°. So we have a synodic period of (25.4 + 1.8) days; which is approximately 27 days.

## 6.4 Solar Coordinates

In order to coordinate observations of the sun, we must adopt a uniform means of specifying locations on the sun. Some systems are geocentric. They refer to apparent positions--as seen from the earth--and are not "fixed" to the sun. In these systems, a given feature will continuously alter its coordinates. Some systems (heliocentric) are fixed to the sun. Each has advantages and disadvantages.

When we observe the sun, it appears as a flat disk against the sky. Features may extend outwards from the visible edge, or limb of the sun. Both disk and limb are terms commonly used in discussing the sun. As the earth rotates, the orientation of the sun appears to change. At sunrise, the west limb (nearest the observer's western horizon) will rise first. At midday, an observer in the middle latitudes of the northern hemisphere will see the solar north pole highest in the sky and the west limb nearest the western horizon. The sun rotates so that the east limb is approaching the observer, and the west limb is receding. If one draws a line connecting the apparent solar north and south poles (central meridian) and a line connecting the midpoint of the east and west limbs (equator), a system of coordinates can be constructed. Although easy to construct, this system has little physical significance due to the inclination of the sun to the ecliptic. It can, however, be corrected to a useful system by accounting for the tilt of the sun's rotational axis.

As we pointed out earlier, the sun's rotational axis is not perpendicular to the ecliptic; nor is it parallel to the earth's axis. This means that the true solar north pole may be tilted towards or away from us each day, and it may be east or west of the apparent north pole. The amount of towards (+) or away (-) tilt of the pole is measured by the $B_0$ angle and ranges from about +7.25° in September to -7.25° in March. This means that the earth (which orbits in the ecliptic) is north of the solar equator in September and south of it in March. The east (+) or west (-) tilt of the axis is measured by the $P_0$ angle. $P_0$ ranges from -26.32° in April to +26.32° in October. Observations are made in the apparent (geocentric) coordinates mentioned earlier and corrected to heliocentric values using the values of $P_0$ and $B_0$ for the day. These data are tabulated in the Astronomical Almanac (see Figure 6.3) for each day of the year. These corrections are made easier by use of a Stoneyhurst Disk for the specific $B_0$ angle (see Figure 6.4).

After selecting a Stoneyhurst Disk for the correct $B_0$ value, the apparent solar equator is aligned with the appropriate $P_0$ angle using the side scales. The Stoneyhurst coordinate grid then provides corrected, or heliocentric coordinates for each plotted feature. The coordinates are

FOR 0$^h$ UNIVERSAL TIME

| Date | | Position Angle of Axis $P$ | Heliographic Latitude $B_0$ | Heliographic Longitude $L_0$ | H. P. | Semi-Diameter | Ephemeris Transit |
|---|---|---|---|---|---|---|---|
| | | ° | ° | ° | ″ | ′ ″ | h m s |
| Aug. | 16 | +16.24 | +6.69 | 259.61 | 8.69 | 15 49.15 | 12 04 16.88 |
| | 17 | 16.58 | 6.73 | 246.39 | 8.69 | 15 49.32 | 12 04 04.48 |
| | 18 | 16.91 | 6.78 | 233.17 | 8.69 | 15 49.50 | 12 03 51.58 |
| | 19 | 17.24 | 6.82 | 219.95 | 8.69 | 15 49.68 | 12 03 38.18 |
| | 20 | 17.56 | 6.86 | 206.74 | 8.70 | 15 49.86 | 12 03 24.30 |
| | 21 | +17.87 | +6.90 | 193.52 | 8.70 | 15 50.06 | 12 03 09.93 |
| | 22 | 18.18 | 6.93 | 180.31 | 8.70 | 15 50.25 | 12 02 55.08 |
| | 23 | 18.49 | 6.97 | 167.10 | 8.70 | 15 50.45 | 12 02 39.77 |
| | 24 | 18.79 | 7.00 | 153.88 | 8.70 | 15 50.66 | 12 02 24.00 |
| | 25 | 19.09 | 7.03 | 140.67 | 8.71 | 15 50.86 | 12 02 07.80 |
| | 26 | +19.38 | +7.06 | 127.46 | 8.71 | 15 51.07 | 12 01 51.17 |
| | 27 | 19.66 | 7.09 | 114.24 | 8.71 | 15 51.29 | 12 01 34.12 |
| | 28 | 19.94 | 7.11 | 101.03 | 8.71 | 15 51.51 | 12 01 16.69 |
| | 29 | 20.22 | 7.14 | 87.82 | 8.71 | 15 51.72 | 12 00 58.87 |
| | 30 | 20.49 | 7.16 | 74.61 | 8.72 | 15 51.95 | 12 00 40.69 |
| | 31 | +20.75 | +7.17 | 61.40 | 8.72 | 15 52.17 | 12 00 22.17 |
| Sept. | 1 | 21.01 | 7.19 | 48.19 | 8.72 | 15 52.40 | 12 00 03.32 |
| | 2 | 21.26 | 7.21 | 34.98 | 8.72 | 15 52.62 | 11 59 44.17 |
| | 3 | 21.51 | 7.22 | 21.77 | 8.72 | 15 52.85 | 11 59 24.73 |
| | 4 | 21.75 | 7.23 | 8.56 | 8.73 | 15 53.08 | 11 59 05.03 |
| | 5 | +21.99 | +7.24 | 355.35 | 8.73 | 15 53.32 | 11 58 45.09 |
| | 6 | 22.22 | 7.24 | 342.15 | 8.73 | 15 53.55 | 11 58 24.92 |
| | 7 | 22.44 | 7.25 | 328.94 | 8.73 | 15 53.78 | 11 58 04.56 |
| | 8 | 22.66 | 7.25 | 315.73 | 8.73 | 15 54.02 | 11 57 44.03 |
| | 9 | 22.88 | 7.25 | 302.53 | 8.74 | 15 54.25 | 11 57 23.33 |
| | 10 | +23.08 | +7.25 | 289.32 | 8.74 | 15 54.49 | 11 57 02.50 |
| | 11 | 23.29 | 7.24 | 276.12 | 8.74 | 15 54.73 | 11 56 41.56 |
| | 12 | 23.48 | 7.24 | 262.91 | 8.74 | 15 54.98 | 11 56 20.51 |
| | 13 | 23.67 | 7.23 | 249.71 | 8.75 | 15 55.22 | 11 55 59.39 |
| | 14 | 23.85 | 7.22 | 236.51 | 8.75 | 15 55.47 | 11 55 38.20 |
| | 15 | +24.03 | +7.21 | 223.31 | 8.75 | 15 55.72 | 11 55 16.97 |
| | 16 | 24.20 | 7.19 | 210.10 | 8.75 | 15 55.97 | 11 54 55.70 |
| | 17 | 24.37 | 7.17 | 196.90 | 8.75 | 15 56.23 | 11 54 34.42 |
| | 18 | 24.52 | 7.16 | 183.70 | 8.76 | 15 56.49 | 11 54 13.14 |
| | 19 | 24.68 | 7.13 | 170.50 | 8.76 | 15 56.75 | 11 53 51.88 |
| | 20 | +24.82 | +7.11 | 157.30 | 8.76 | 15 57.02 | 11 53 30.66 |
| | 21 | 24.96 | 7.09 | 144.10 | 8.76 | 15 57.28 | 11 53 09.48 |
| | 22 | 25.09 | 7.06 | 130.90 | 8.77 | 15 57.56 | 11 52 48.37 |
| | 23 | 25.22 | 7.03 | 117.70 | 8.77 | 15 57.83 | 11 52 27.35 |
| | 24 | 25.34 | 7.00 | 104.51 | 8.77 | 15 58.10 | 11 52 06.44 |
| | 25 | +25.45 | +6.96 | 91.31 | 8.77 | 15 58.38 | 11 51 45.65 |
| | 26 | 25.56 | 6.93 | 78.11 | 8.78 | 15 58.65 | 11 51 25.01 |
| | 27 | 25.66 | 6.89 | 64.91 | 8.78 | 15 58.93 | 11 51 04.54 |
| | 28 | 25.75 | 6.85 | 51.71 | 8.78 | 15 59.21 | 11 50 44.26 |
| | 29 | 25.84 | 6.81 | 38.52 | 8.78 | 15 59.49 | 11 50 24.20 |
| | 30 | +25.92 | +6.77 | 25.32 | 8.79 | 15 59.76 | 11 50 04.37 |
| Oct. | 1 | +25.99 | +6.72 | 12.12 | 8.79 | 16 00.04 | 11 49 44.79 |

**Figure 6.3 Sample Page from Astronomical Almanac (1982).**

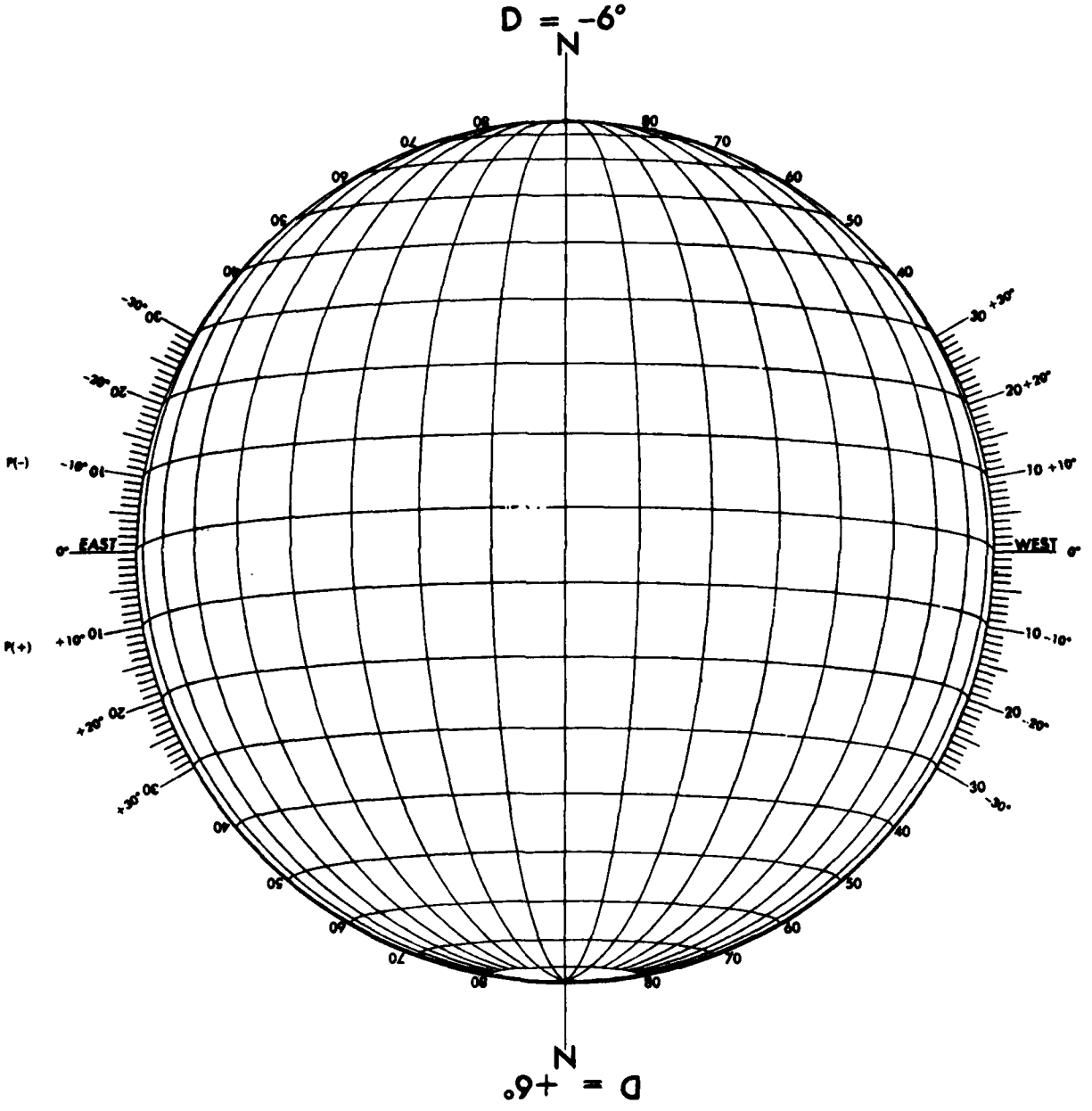



Figure 6.4 Stoneyhurst Disk.

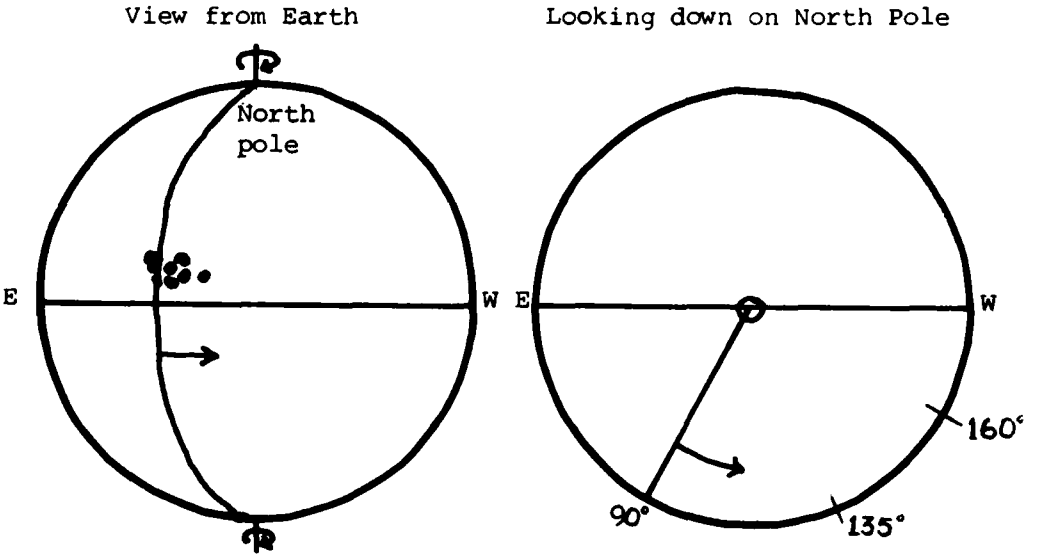
heliographic latitude ($0^o$-$90^o$N or S) and heliographic longitude ($0^o$-$90^o$E or W). Latitude is measured from the solar equator, and longitude is measured east (towards east limb) or west from the true central meridian (CM). The heliographic longitude of a given feature changes daily due to solar rotation and earth revolution. SESS observatories typically correct photographs for $P_o$. A similarly simple $B_o$ correction is, of course, not possible (Why?).

A longitude system which co-rotates with the sun was defined in the 1850s. Since latitude already remains constant, a fixed-to-the-sun system is possible, and "active longitudes" can easily be identified. This co-rotating longitude is identified as $L_o$ (see Figure 6.3) and is named Carrington Longitude after its originator. Carrington defined the meridian at the sun's center on a certain date and time (1200 U.T., 9 November 1853) as the prime meridian. Carrington longitudes are measured eastward (counterclockwise from above the solar north pole) from the prime meridian. Figure 6.5 gives a plan - view of this system.

A "Carrington Rotation" is said to begin each time $0^o$ Carrington crosses CM, and the Carrington Rotation number is a way of keeping track of "solar months" (compare to Julian Day system).

## 6.5 Celestial Coordinates

While solar coordinate systems are useful for positioning features on the sun, a different type of coordinates are required to identify the location of objects in space. Certain signposts must first be identified. The celestial




Carrington longitude lines rotate with the sun (period 27 days). As seen from the earth, they move from east to west across the face of the sun.

**Figure 6.5 Carrington Longitude.**

poles have already been mentioned. Two additional points are useful. The point directly overhead is the zenith. It is opposite the direction a plumb bob would point (known as the nadir). The great circle which passes through the celestial poles, the zenith, and the nadir is known as the local celestial meridian. Although we know that the earth is not the center of the universe, picturing celestial coordinates is easier if one imagines all the stars fixed on a sphere of infinite radius--the celestial sphere. The earth is fixed at the center of this sphere, and the sphere rotates on an axis passing through the celestial poles (and the center of the earth).

The first celestial system of interest is known as the altazimuth system. A plane tangent to the earth's surface at the observer's location will intercept the celestial sphere to define the celestial horizon. The observer can see only objects above this horizon. Since the radius of the earth is negligible by comparison to the radius of the celestial sphere, we can consider the plane as passing through the center of the earth (and oriented tangent to the observer's location). Arcs which lie on the celestial sphere, extend from the zenith to the celestial horizon, and intercept the celestial horizon in a right (90°) angle are called meridians. The meridian which also passes through the north celestial pole (called the prime meridian) intercepts the celestial horizon at two points known as the north and south cardinal points. The coordinate called azimuth is measured from the north cardinal point (0° or 360°) through east (90°), and on around the celestial horizon. Azimuth is much like a normal compass rose. The second coordinate of this system is altitude. Altitude is measured up from the horizon along a meridian passing through the body. For example, the altitude of the zenith is 90°. The altitude of the north cardinal point is zero. Zenith distance is just (90° - altitude)--the distance down from the zenith to the body. Altitude ranges from 0°-90°, and azimuth has a range of 0°-359°.

Altitude and azimuth are easy to visualize. Unfortunately, they are tied to the observer. This means that the altitude and azimuth of an object such as the sun will vary continuously due to the earth's rotation. At midday, the sun will be on the prime meridian and have its maximum altitude. It will be at the zenith for only one latitude at any one instant and never at the zenith for latitudes above 23 1/2°. At sunrise and sunset, the altitude of the sun will be zero (neglecting refraction).

Since the sun's apparent position varies with time, it follows that we can use it to measure time. We do this by defining a coordinate system known as the local hour angle (LHA) system. This system is an extension of the earth's latitude/longitude system onto the celestial sphere. An extension of the earth's equator intercepts the celestial sphere in a great circle called the celestial equator. (The angle between the celestial equator and the zenith measured along the prime meridian is equal to the observer's latitude.) Since the earth rotates about an axis perpendicular to the celestial equator, all stars appear to move parallel to the celestial equator. An object's angular distance above or below the celestial equator (similar to latitude) is termed its declination. Most objects (not the sun, moon, and planets) are essentially fixed in declination. The declination of the north celestial pole is 90°. (Polaris is near the north celestial pole.)

The second coordinate, (LHA), is predicated on that portion of the observer's prime meridian which extends from the north celestial pole to the south cardinal point (for a northern hemisphere observer). LHA is measured westward from the local celestial (prime) meridian to the meridian passing through the object in question. Meridians intercept the celestial equator in a 90° angle and describe a great circle to the celestial poles (much like terrestrial meridians). LHA is a measure of how long ago the object was on the observer's prime meridian. It can be measured in units of arc (0° - 359°) or time (0 - 24 hours). The conversion is easy, since the earth rotates 360° in 24 hours. A bit of division yields 1 hour. = 15°; 4 minutes = 1°; etc.

A sundial measures Local Apparent Solar Time (LAST). This is accomplished by noting the varying position of a shadow cast by an appropriately aligned wire. Mathematically, LAST = LHA of the sun + 12 hours. This means LAST = 1200 when the sun is on the observer's meridian. Since the sun moves with a varying rate over the course of the year, LAST flows discontinuously. As a consequence, we define LMST: Local Mean Solar Time = LAST-Equation of Time. The Equation of Time can be determined from a table for any day. Even LMST does not eliminate all time problems. Observers at different longitudes will measure different times at the same instant. To minimize the confusion which would result from this, we introduce the concept of time zones. All longitudes within 7 1/2° either side of a standard meridian use the time observed on the standard meridian. Standard meridians are spaced at 15° intervals east and west of the Greenwich (England) Meridian. Local Mean Solar Time on a standard meridian is called Standard Time for that particular time zone. Greenwich Mean time (also known as Z time, GMT, and Universal Time) is standard time on the Greenwich meridian.

The more eastern the observer's location, the greater will be the observed LHA for any given object. This means that an eastern observer will have a later time at any instant than a more westerly observer. Were it not for the International Date Line, "time travel" would be possible simply by circumnavigating the globe.

## 6.6 Summary

There are a number of coordinate systems. Each is designed for a particular purpose or to simplify a particular analysis. A good understanding of the motions of the earth and sun and their respective coordinate systems is essential to relating these systems to each other. Since many solar-terrestrial interactions are a function of solar altitude and relative earth-sun positions, the systems discussed above are of primary interest to SESS.

# CHAPTER 7

## THE QUIET SUN

The sun, our nearest star, is circled by the earth in an elliptical orbit. The average earth-sun separation, 93 million miles or about 150 million kilometers, is defined as one astronomical unit (1 AU). Since the orbit is an ellipse rather than a circle, this distance varies from a minimum of 91.4 million miles (147 million kilometers) to a maximum of 94.6 million miles (152 million kilometers). It takes the earth 365.256 earth days to complete one orbit around the sun. The sun has a radius of 696 thousand kilometers (432 thousand miles). By comparison, the radius of the earth is 6378 kilometers (3964 miles); so the solar radius is 109 times the earth radius. It would take nearly 12 thousand earths to cover the face of the sun and 1.3 million earths to fill the volume of the sun. The mass of the sun is $1.989 \times 10^{30}$ kilograms, about one-third million times the mass of the earth. The average density of the sun is about 1.4 grams per cubic centimeter, slightly more dense than water. The sun is not composed of water, but mainly of hydrogen and helium. Traces of at least 62 other elements exist in the sun, with the most common ones being carbon, nitrogen, oxygen, neon, magnesium, silicon, sulfur, argon, calcium, iron, and nickel. The quiet sun has an overall magnetic field of about one or two gauss. It is thought to be about 5 billion years old.

The sun emits huge amounts of energy and mass. The radiated energy of the entire sun is estimated at $4 \times 10^{33}$ erg/sec or $4 \times 10^{23}$ kilowatts. At the earth, the received energy flux is still large -- 1.36 kilowatts per square meter per second or about 1.5 horsepower per square yard per second. This energy is spread across the electromagnetic spectrum with approximately 52% in infrared (heat), 41% in visible, 7% in near ultra-violet, 0.1% in extreme-ultra-violet and x-ray, and $10^{-8}$% in radio waves. The sun emits a steady stream of particles, known as the solar wind. The solar wind carries about one billion kilograms of mass away from the sun every second. At this rate, it would take $3 \times 10^{12}$ years (30 billion centuries) to decrease the solar mass by 10%.

Despite the nearly incomprehensible numbers attributed to the sun, it is really nothing more than a fairly average, garden variety star. Lest we become complacent, there is an aspect of stars which is most important. How do they effectively "manage" their vast resources? For the most part, the sun seems very stable. Yet, there are indications of considerable variability in the sun, as in other stars (the so-called "flare stars"). In this and the following chapter, we address the operation of a star.

### 7.1 Energy Transport

The sun is thought to convert about 4 million tons of hydrogen into energy each second. In order not to explode, the sun must manage this vast amount of energy. At least three transport mechanisms are thought to function routinely within the sun: conduction, radiation, and convection. Although all three typically function simultaneously, one will usually dominate. Which one dominates depends on the temperature gradient.

Conduction is present, to a small degree, throughout the sun. It is probably not dominant anywhere. Collisions between energetic particles, typically electrons, are the actual means of conduction. The mean free path (average distance between collisions) is probably short for electrons at all levels in the solar interior. This means that conduction is a slow, somewhat inefficient means of energy transport. As energy "backs up" awaiting transport up from the solar interior, the temperature rises. Higher levels remain relatively cool because of the slow rate of energy transport via conduction. This means an increasing temperature gradient. Eventually, it will become high enough to permit radiation to function efficiently.

Radiation is energy transport by photons. Except in certain types of degenerate stars, the mean free path for photons generally exceeds that for electrons. This means radiation will dominate, since it is more efficient. Photons travel with the speed of light (because, they are light). The "light time" from the center of the sun to the earth is slightly more than eight minutes. Yet, calculations reveal that the average photon requires thousands of years (perhaps millions) to complete this journey. This means something must be acting to retard the escape of (hence, energy transport by) photons from the center of the sun. This something is called opacity.

Opacity is a resistance to the passage of light. In stars, it is created by placing atoms capable of absorbing photons (i.e. atoms with some electrons attached) in the path of the light. Actually, the requirements for an opacity source are somewhat more involved than this. In addition to having retained some of its electrons, the atom must be in the proper configuration (electrons in the right orbits) to absorb the particular photon energies available. This restriction is critical. If an atom is too cold, its electrons will be in their lowest energy or ground state orbits. The only photons which these atoms can absorb are very high energy photons capable of exciting the electrons out of their ground state. These atoms, while retaining their electrons, would be essentially transparent to middle and low energy photons. They would not create an effective opacity source unless high energy photons were present. Hence, opacity sources within stars can be said to turn on and off at various levels as temperatures alter atomic states and photons present. Where there is no effective opacity source, the mean free path for photons will be long. Hence, radiation will be an effective energy transport mechanism. When an effective opacity source "turns on", the mean free path for photons will be short, and radiation will be inefficient. Again, the temperature gradient will rise.

At very high temperature gradients, convection will easily dominate the other two transport mechanisms. Moving parcels or gas currents become the vehicle for energy movement. The parcel, being warmer than its surroundings, is bouyant and rises. As the parcel gives off energy, recombination occurs within it to release additional heat and maintain bouyancy. Eventually, all sources of internal heat are depleted, and the parcel merges with its surroundings.

All three transport mechanisms are thought to operate in the sun. Which one dominates at each level of the sun determines much of the structure of that level. In fact, we can describe the interior of the sun by discussing the regions in which each mechanism functions.

## 7.2 Internal Structure

In the quiescent model, the sun is viewed as a static, spherically symmetric ball of hot gases. Solar properties are assumed to change only with radius and to be uniform over any spherical layer. Thus, we can define solar properties in terms of distance from the solar center. Using this system, we divide the sun into six layers: core, radiation zone, convection zone, photosphere, chromosphere, and corona. The chromosphere and corona constitute the solar atmosphere. Below the solar atmosphere, we find the solar interior.

### 7.2.1 The Solar Core

The sun is massive. Under its own gravitational attraction, the solar material is compressed to such a high central density and temperature that nuclear fusion takes place. Fusion is the process of combining light elements such as hydrogen to make heavier elements. Fusion also drives a hydrogen bomb. (This process requires such high temperatures and pressures that Man has not yet been able to produce a controlled fusion reaction.) Fusion occurs when two nuclei move close enough together to interchange components or to combine. The total mass after the fusion reaction is less than before it, because the "missing" mass was converted to binding energy.

The core of the sun is defined as the region within .25 $R_{sun}$. This small sphere (1/64 the solar volume) contains 1/2 the solar mass. The central pressure is 250 billion atmospheres, and the gas density is 158 grams per cubic centimeter (158 times that of water or 11 1/2 times that of lead). Since particles move freely, the matter is a gas. The temperature is about 15 million degrees Kelvin, and the gas particles are 100% ionized (i.e., the gas is a plasma). In this core, 99% of the sun's emitted energy is generated.

Fusion in the solar core consumes 600 million tons of protons per second to produce alpha particles ($^4He^{++}$) and 4 X $10^{23}$ kilowatts of energy. Chargeless, massless particles called neutrinos are also thought to be produced. These particles interact only weakly with matter and so escape readily from the core. Earth-based detectors have been designed to measure solar neutrino fluxes. So far, they have detected only about 20% the number predicted by theory--a discrepancy which raises major questions about solar structure. The hydrogen "burning" in the core also produces helium "ash." As much as 10% of the core may now be helium. Eventually, this helium ash will quench the hydrogen fusion. Strangely, much hydrogen will still remain in the sun, but it will be unable to reach the core. Current theory suggests there is no mixing between the core and overlying, hydrogen rich layers. Nonetheless, sufficient fuel remains in the core to power the sun for at least another 5 billion years. Since the gas in the core is completely ionized, it is transparent to photons. Consequently, energy is transported out of the core by radiation. It is said to be in radiative equilibrium with the region surrounding the core.

### 7.2.2 Radiation Zone

The Radiation Zone is, in many ways, the simplest of all regions of the sun. Fusion has virtually ceased by the time we reach 0.25 $R_{sun}$ from the center of the sun. This occurs because the temperature has dropped to less

than ten million degrees, and fusion is very dependent on high temperature. The radiation zone is named for the dominant process occurring in the region.

Energy is transferred outward through this region by radiation. Although much of the region inside .86 $R_{sun}$ is highly ionized, it remains opaque to gamma rays and ultraviolet (UV) photons. Each photon is absorbed and reradiated many times as it works its way outward through the sun. It has been estimated that this radiative transfer is so slow that if fusion in the solar core ceased today, it would be many thousands of years before we could measure a drop in solar EM output. The frequency of the emitted radiation depends on the temperature and nature of the emitting material (blackbody radiation). As the photons move outward and the gas temperature decreases, the frequency emitted by progressively higher layers drops from gamma rays to x-rays to ultraviolet radiation. However, radiation is not the dominant energy transfer process all the way to the surface of the sun.

At approximately 0.86 $R_{sun}$, the gas properties have changed significantly. The steady drop in temperature has cooled the gas to under a million degrees, and hydrogen - the most abundant element in the sun - recombines. The electrically neutral hydrogen provides a marked increase in opacity and a corresponding decrease in the mean free path of photons. This results in a sharp drop in the efficiency of radiation. A greatly steepened temperature gradient initiates convection.

### 7.2.3 The Convection Zone

In the region we call the convection zone, gas motion is the dominant energy transfer process. The temperature decreases rapidly with height, and this steep vertical temperature gradient results in convective instability. Near the top of the layer, the hot gas particles radiate energy and cool. The cool gas particles subside to the bottom of the layer, absorb electromagnetic energy from the radiation zone, and begin their ascent through the layer.

The turbulent convection creates noise (low frequency sound waves) which propagates upward into higher layers of the solar atmosphere. The top of the convective zone appears as an energetically boiling and bubbling layer. These convective bubbles give the surface of the sun its mottled or granular appearance. Individual convective cells seem to reach depths of about 400 km. In the central portion of the cells, upward motions of about 2 km/sec are observed. Material then flows horizontally at about 1/2 km/sec beyond the edge of the cell and sinks. This motion also produces gravity waves, which, with the sound waves, carry large amounts of non-radiative energy upward. Some of this energy, in the form of shock waves, is thought to heat the solar atmosphere.

Our examination of the solar interior has, thus far, been based on theory. Since we cannot observe the solar interior, we calculate its characteristics based on theory and surface observations. Above the top of the convective zone, we enter the photosphere, the lowest layer we can directly observe.

### 7.2.4 The Photosphere

The distance an electromagnetic wave will travel before it is absorbed is dependent on the density of potential absorbers present. Just above the convective zone is the layer where emitted radiation first has a chance of traveling to the earth without being absorbed. This layer is called the photosphere. It is the visible surface of the sun.

Photos is Greek for light, so photosphere translates as "light sphere." Almost all the light emitted by the sun comes from the photosphere. The photosphere is a very thin layer (500 to 1000 kilometers thick) which emits primarily at visible and infrared wavelengths. The edge of the visible disk appears sharp in "white" (frequency integrated) light, because the photosphere is so thin. Moreover, photospheric radiation establishes the blackbody, or effective temperature of the sun as 5280°K. This means that the sun's emission is most closely approximated by an ideal blackbody of 5280°K. Most energy radiated by such a body is in the so-called "optical" portion of the EM spectrum.

The photospheric gas density ($10^{-8}$ grams/cm$^3$ or $10^{14}$ particles/cm$^3$) is approximately one hundred thousandth that of the earth's atmosphere at sea level, and the gas pressure is approximately 0.01 atmosphere. The temperature at the bottom of the photosphere is approximately 6000°K, and at the top, it is near 4300°K (Figure 7.1). The effective temperature of the entire layer is 5280°K. Solar gravity at the photosphere is 27 times that at sea level on the earth.

Examination of a photospheric photograph reveals a center to limb drop in disk brightness. This is a direct result of the photospheric temperature gradient. When we observe the sun, we see down to a given level, known as optical depth, past a certain number of photospheric particles. By looking at the solar limb we are looking into the sun at an angle and reach our optical depth physically higher in the photosphere than when we observe straight down at disk center. Higher in the photosphere, the temperature is lower and thus, less energy is emitted. This effect is known as limb darkening.

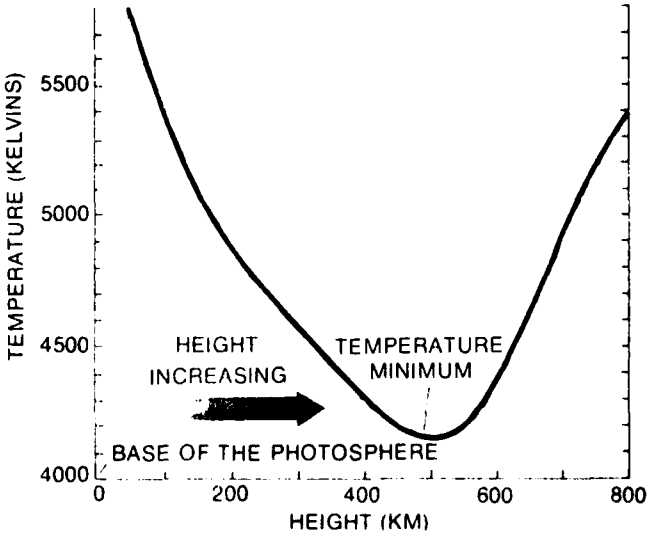



**Figure 7.1 Photospheric Temperature Profile (from Pasachoff, 1977).**

Further examination of a photospheric photograph reveals a cellular pattern known as granulation. Granules are seen as relatively bright, irregularly shaped polygons separated by narrow dark lanes. The average cell diameter is approximately 1800 kilometers. We are observing the tops of convective cells which have penetrated the photosphere. Relatively hot convective zone gases rise from the bright (hot) granule center, and cool by radiating away excess energy. Then, they spread out horizontally across the top of the granule. The cooled, denser gases subside along the darker lanes which outline the cells. The flow pattern is not long-lived, for a granule will last an average of only ten minutes. By the end of that time, the cell will have faded, broken up, or coalesced with another cell.

A second convective structure, known as supergranulation, is known to exist through the photosphere and into the chromosphere. Supergranules have an average diameter of 30 thousand kilometers and an average lifetime of 20 hours. This pattern, which is much harder to observe, originates deeper in the convection zone than the granular one and probably results from the change in opacity caused by the recombination of helium (which occurs at a higher temperature than hydrogen recombination). Supergranules seem to be associated with solar magnetic fields. Spicules, which surround supergranules like a picket fence, are a visible manifestation of these magnetic fields. The magnetic fields are thought to connect the photosphere through the chromosphere to the overlying corona. Moreover, material in the short-lived spicules is thought to be "thrown" into the corona.

## 7.3 The Solar Atmosphere

To the unaided observer, the photosphere marks the edge of the sun. Filters, occulting disks of a coronograph, and the fortuitous solar eclipse permit viewing the tenuous outer layers of the sun. The outer levels of the sun are so low in density that, despite their high temperature they pale into invisibility in the sun's brilliance. In fact, this thin solar envelope apparently extends to the edge of the solar system (defining the heliosphere).

### 7.3.1 Chromosphere

Early solar eclipse accounts tell of a brilliant red flash just after the moon obscured the photosphere. At one time, astronomers (led by Johannes Kepler) thought this was due to scattering of the photospheric light by the atmosphere of the moon. They reached this conclusion, because they thought the solar atmospheric density went to zero just above the observed disk. Later work revealed that the flash was due to emission from a solar layer above the photosphere. Chromos is Greek for color, so this layer was named the chromosphere.

The chromospheric gas is less dense but hotter than the lower photosphere. The base of the chromosphere is defined as the height of the temperature minimum in the solar atmosphere, 4300°K. Through the thin chromosphere (approximately 3000 kilometers thick) temperature increases rapidly with height. In the lower portion of the chromosphere, the increase is relatively gradual, reaching 8000°K in the first thousand kilometers. However, near the top of the chromosphere the increase is very rapid, going to nearly a million degrees at the top of the chromosphere. The chromosphere is not in thermodynamic equilibrium, so these temperatures are kinetic, not

sensible. Density decreases through the layer to a value at the top near $10^{10}$ particles per cubic centimeter, about one ten-thousandth that of the photosphere and one billionth that of the earth's atmosphere at sea level.

Most chromospheric observations are made at either hydrogen-alpha or calcium-K wavelengths. Hydrogen-alpha is a particular wavelength (6563 angstroms) of EM radiation due to hydrogen. It is the strongest visible absorption line of the most common solar gas. In emission, hydrogen alpha is a brilliant red color. Calcium K is the strongest absorption line in the visible spectrum of the sun. It occurs in emission at approximately the same altitude as hydrogen-alpha and is about 3934 angstroms (yellow). How can we see both an absorption and an emission line at the same wavelength?

Absorption occurs at these wavelengths in the photosphere, so dark lines are created in the otherwise continuous emission welling up from lower levels. In the hotter, overlying chromosphere, emission lines occur at these wavelengths. The result is an emission line superimposed on a broad, dark absorption line. (The absorption line is broadened (more than a single wavelength wide) by collisions which alter the energy levels of some atoms, Doppler shifts due to turbulence and rotation, and local magnetic fields.) For this reason, the chromosphere is often called a reversing layer. Hydrogen-alpha emission is formed in a very thin range of altitudes in the chromosphere. The calcium K line forms over a slightly broader range of altitudes and so is somewhat brighter (more atoms contributing emission photons). Unfortunately, features viewed in the light of calcium are somewhat less distinct (like an out of focus photograph) than the hydrogen-alpha view. Much solar activity occurs in the relatively thin chromosphere; plage and flares, for example.

We frequently see bright, stable structures on the solar limb in hydrogen-alpha. These features may be millions of kilometers long. They are called quiescent (quiet) prominences. They are condensations of gas with chromospheric temperature and density, suspended in the corona. We also see dark, stable, threadlike features against the disk. These are called filaments. Filaments and prominences are names for the same material. When seen on the disk, the suspended gas absorbs more photospheric energy than it gives off along the line of sight, so it appears dark. When seen on the limb, it emits more energy than is given off by the less dense corona that surrounds it, so it appears bright. Both prominences and filaments extend through the chromosphere into the outermost layer of the solar atmosphere -- the corona.

### 7.3.2 Corona

Corona, Latin for crown, is the name given the outermost layer of the sun. The corona is the very hot, tenuous outer layer of the solar atmosphere. Near the sun, the coronal temperature is nearly two million degrees. The particle density is approximately one hundred thousand ($10^5$) particles per cubic centimeter ($5 \times 10^{-17}$ g/cm$^{-3}$). By comparison, earth's sea level density is $10^{19}$ particles/cm$^3$. The corona's low density accounts for its low intensity -- being visible only when the bright disk of the sun is occulted by an eclipse or artificially by a coronagraph.

The hot corona poses a dilemma. "How do we get a hot corona above a cooler photosphere if the energy is all produced by fusion below the

photosphere?" The answer comes by considering non-electromagnetic waves. The gas motion in the convective zone produces waves which carry energy upward to heat the corona. The rising fluid may be likened to a boiling pot of water. In the water, the rising steam bubbles burst through the top of the water in the pan and produce a "bubbling" sound. Although the rising solar gas in the convective zone doesn't escape the sun, it does overshoot into the stable photospheric layer above and produces sound (acoustic) and gravity waves. These waves propagate upward through the chromosphere, and give up their energy in the lower corona. Thus, the coronal temperature is due to energy from the convection of solar gas below the photosphere.

The corona is strongly affected by the solar magnetic field. The magnetic field controls the motion of coronal particles, and indirectly determines local coronal density. Magnetic field lines are the imaginary field lines which connect a north to a south magnetic pole. The field lines may be divided into open and closed lines. An open solar magnetic field line has only one end which connects to the sun (penetrates photosphere), while a closed field line has both ends rooted in the photosphere. Charged coronal particles follow field lines. Closed field lines will trap protons and electrons in the corona and keep the particle density relatively high. Open field lines will allow the protons and electrons to escape the corona and let the particle density decrease to relatively low values. This low density region in the corona is termed a coronal hole and was one of the major discoveries of the Skylab flights.

### 7.3.3 Coronal Holes

Coronal holes can be observed directly as areas of low EUV (extreme ultraviolet) and x-ray emission. From such observations, we can construct a brief climatology of coronal holes. Some, perhaps most, coronal holes extend non-radially outward from the sun. In other words, they tilt. This tilt may account for observations of particles escaping high latitude ($40^{o}$) holes reaching the solar equator. Low latitude (between $10^{o}$-$30^{o}$) holes seem equally important. Most high speed particle streams (greater than 500 km/sec) observed at 1 AU seem to originate in holes at or below $40^{o}$ heliographic latitude.

Four categories of holes are known. Except near the maximum of the solar cycle, large coronal holes dominate the solar poles. When polar magnetic fields are weak (near solar maximum), the polar holes are replaced by large, high latitude ($30^{o}$-$70^{o}$) holes. This may be a manifestation of the solar magnetic field reversal to be discussed later. The third and longest-lived flavor of coronal hole is the low latitude hole. These large, unipolar regions typically appear during the declining phase of the solar cycle between $10^{o}$-$30^{o}$. Near sunspot maximum, the high latitude holes are accompanied by short-lived equatorial (below $10^{o}$) holes. These holes are usually located between the sunspot belts and often last for little more than one solar rotation.

The corona expands outward from the sun, filling the solar system. This outward flux of material accounts for the small mass loss of the sun and is termed the solar wind. Coronal holes are thought to be the primary sources of high speed solar wind streams. Some holes have no observable wind streams,

while others have wide streams. Small (less than 13° in longitudinal extent), low latitude holes usually have very narrow wind streams and little observable effect at 1 AU.

## 7.4 Surface Velocity Features

We observed earlier that the sun, being a non-rigid body, rotates differentially. Each latitude has a different velocity than any other latitude. This shearing action causes observable shifts in sunspot groups. Recently, observers (Howard and LaBonte, 1980) noted organized longitudinal speed variations of 3 m/sec at the same latitude. It is not yet clear whether this cyclic behavior is only surface deep or tied to a deep-rooted phenomenon. Four zones, two fast and two slow, are observed in each solar hemisphere. The zones are of alternating speed and seem to move, like a barber pole stripe, from the solar poles to the solar equator. As Figure 7.2 shows, every 11 years a new, high speed stream originates at each pole. During the ensuing 22 years, it moves steadily towards the equator. When the stream reaches the middle latitudes (near the beginning of a new sunspot cycle), sunspots form along the poleward boundary of the fast stream and accompany it to the solar equator, where both disappear. This suggests that the activity centers, sunspots, are just a side effect of a much larger scale phenomenon. These large scale current streams are the most deeply rooted features yet observed which seem to vary systematically. They may well be the link between the quiet, stable star we have discussed thus far and the strangely variable sun to which we next turn our attention.

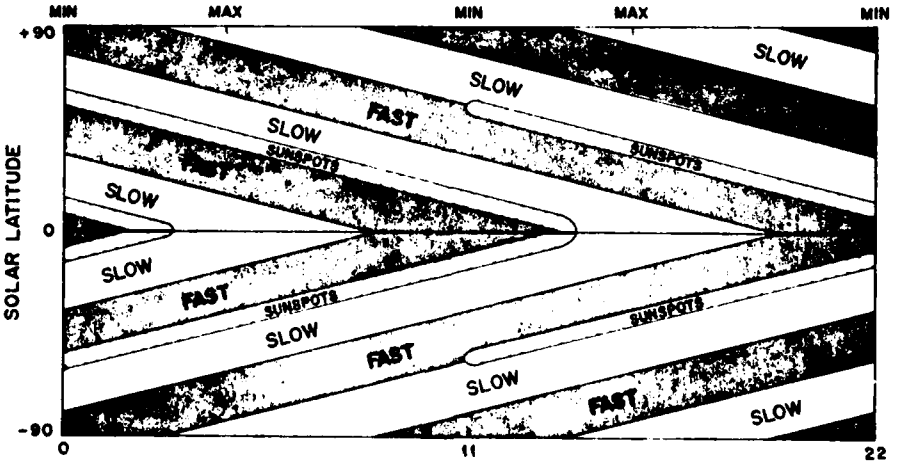



Figure 7.2 Latitudinal Variability in Solar Rotational Rates (from Howard and LaBonte, 1980).

## 7.5 Summary

Although the sun is immense by comparison to the earth, it is really a fairly average star. Nuclear fusion in the sun's core is thought to power our nearest star and strongly influence its structure. We infer the existence and structure of the core, radiation zone, and convection zone from observations of the visible solar surface, the photosphere. The chromosphere, a bubbling froth atop the photosphere, is the site of most solar activity and the origin of most hydrogen-alpha emission. The corona is the near-sun term for what, near the earth, is called the solar wind. The solar wind is structured by the solar magnetic field.

# CHAPTER 8

# THE ACTIVE SUN

The sun's ultimate energy source is nuclear fusion. Since this energy is released at a steady rate, one might expect a spatially uniform and temporally steady release of radiation at all wavelengths over the entire solar surface. However, over short periods of time and from certain locations, the intensity fluctuates rapidly. Large emission increases occur at the shorter and longer wavelengths. The major factors causing these fluctuations are the sun's large-scale magnetic field and its differential rotation. The sun's background magnetic field of 1-3 gauss, frozen into the surface and initially running from pole to pole, is distorted by differential rotation. By twisting and shearing of the magnetic field, energy is stored. This stored energy can be released by solar active regions. Several features are associated with these active regions, though not all are always present. The most common indicators of locally enhanced magnetic fields are facula, plage, and sunspots.

## 8.1 Plage and Facula

Solar observers have frequently observed vast regions of enhanced emission in certain chromospheric lines. These regions are due to a higher temperature, higher density region of gas in the chromosphere. Early French observers compared the appearance of these regions to the white sand of a beach and named them plage, the French term for beach. The gas of a plage is heated and compressed by a strong magnetic field extending from the solar surface into the chromosphere. A plage is usually associated with a magnetic field of 200-500 gauss, compared to 1 to 3 gauss for the undisturbed sun. The plage is the first chromospheric indicator, and longest lived feature, of this strong magnetic field. Plage is routinely observed in both hydrogen-alpha and calcium light. The former is now more common, because it provides greater detail. Plage area is usually measured in millionths of the solar hemisphere. For comparison, the earth would cover about 100 millionths of the solar hemisphere.

When a typical plage forms, it is relatively small and intense. As it grows in size, it initially maintains its intensity. On reaching maturity, it often decreases in intensity while continuing to grow in size. This "dulling" is generally thought to be correlated to the magnetic field passing its most complex, and least stable, phase. The decay in size of the plage is a fairly slow process. The lifetime of a plage may be several days, several weeks, or even a few months.

Sometimes, long-lived, enhanced emission regions appear in the photosphere. These regions usually occur in conjunction with chromospheric plages and are called faculae. A facula, like the plage above it, is evidence of a concentrated magnetic field. It is often visible in white light and most easily seen near the limbs as a result of limb darkening providing greater contrast. It is more easily visible on the disk at discrete wavelengths. The structure (magnetic) associated with facula (and plage) is visible over a considerable range of altitude in the solar atmosphere and tends to spread out

slightly with increasing altitude. Facular structure shows structural variations over time periods of a few hours or less. For 1-2 years before sunspot cycle minimum, faculae are also observed at high latitudes (near $67^{o}$) and may be associated with the newly evolving polar magnetic fields (referred to elsewhere as background solar fields).

## 8.2 Sunspots

The most commonly reported solar active feature is the sunspot. Sunspots were first observed by such astronomers as Galileo, who left sunspot drawings. The spots which he drew were dark features, normally with a very dark, nearly circular center and gray boundary region. Sunspots are located in the photosphere under plage regions. The term "UMBRA" is applied to the dark core of the spot and "PENUMBRA" to the gray boundary region of the spot. Sunspots are regions of very strong magnetic fields (typically several thousand gauss) which constrain relatively cool (approximately $4000^{o}K$) gas. This gas is emitting light, but since its temperature is lower than the ambient photosphere, the spot appears very dark. The spot is relatively cool, because the magnetic field both confines the gas and reduces the interaction between the spot and the surrounding gas. The gas in the spot cools by radiating more energy than it receives. The difference between the magnetic field of a spot and that of a plage is intensity; the weaker (few hundred gauss) plage magnetic field concentrates energy coming up from below, while the intense sunspot field effectively blocks the energy flow. Many sunspots are found to lean, vertically, to the west. They are found to be displaced by $1/2^{o}$ to $7^{o}$ from vertical, possibly due to differential rotation rates at different atmospheric levels. High resolution observations reveal that gas flows out of the umbra, parallel to the solar surface, and is deposited at about twice the penumbral radius from the spot center. The flow rate is about 2 km/sec and is known as Evershed flow.

Less than half of all sunspots develop penumbra. The existence of penumbra is probably more dependent on sunspot magnetic field geometry than on spot size or intensity (McIntosh, 1981). The penumbra results from transverse magnetic fields and may briefly (less than a day) exist without interior umbra. Mature penumbra is usually darker than rudimentary penumbra and is often associated with an older spot group.

Young spots which lack penumbra are often mistakenly referred to as pores. A pore is in reality less intense and shorter-lived than a sunspot. Its diameter is usually less than 2500km; its magnetic field is about 1500 gauss, and its lifetime is typically 15 minutes to 1 hour. Pores produce a photospheric darkening similar to that of intergranular spaces. The transition from pore to sunspot is generally a change in intensity as opposed to the coalescence of several pores. Many pores form near sunspot groups.

### 8.2.1 Sunspot Groups

A large percentage of solar plage regions develop more than one sunspot. These spots are evidence of magnetic field eruptions through the solar surface, and they normally behave as a unit. They are collectively termed a sunspot group.

A sunspot group is oriented primarily east-west. Thus, the longitudinal extent of the group is its most commonly used size designator, although sunspot area (in square degrees or millionths of the hemisphere) is also used. The western-most sunspot in a group is known as the leader (or preceding) sunspot, and the more easterly spots are called follower (or trailer) spots. These names come from the fact that, as the sun rotates, the leader appears to lead the way across the disk.

The leader spot is of special interest in studying the group. Generally, it is the first spot to form and last spot to disappear. Normally, it is also the first spot in the group to form a penumbra and last to loose it, and is the largest spot throughout the life of the group. Most sunspot groups in a given solar hemisphere have leaders of the same magnetic polarity. The leaders in the northern hemisphere have the same magnetic polarity as the trailer spots in the southern hemisphere. Moreover, the leader polarity generally matches the background magnetic field polarity of the hemisphere.

### 8.2.2 Sunspot Classification

Several schemes have been devised to classify sunspot groups. The objective, in most cases, is to identify increasingly complicated spot structures. The more complicated the magnetic structure of a group, the more potential it has for the explosive release of its trapped energy in a flare. Thus, sunspot classification schemes provide insight into flare forecasting.

The Zurich classification system classes sunspot groups by size. See Figure 8.1.

| *Classification* | *Meaning* |
|---|---|
| A | A single (long lived) pore or group of unipolar pores. |
| B | A long lived group of bipolar pores. |
| C | A bipolar group in which only the largest spot (generally the leader) has a penumbra. |
| D | A bipolar group whose main spots all have penumbrae and which is less than 10 degrees (heliographic) in longitudinal extent. |
| E | A large bipolar group whose main spots all have penumbrae and which is 10 to 15 degrees in longitudinal extent. |
| F | A very large bipolar group whose main spots all have penumbrae and which is over 15 degrees in longitudinal extent. |
| H | A spot or unipolar group of spots with at least one penumbra (generally on the leader spot). |

The magnetic complexity of a sunspot group is the criterion for the second classification system, the Mount Wilson magnetic system. The categories are shown in the indicated figures, where the dotted lines represent magnetic polarity reversals (inversion lines).

| Classification | Meaning |
|---|---|
| Alpha | Unipolar group; that is, all plus or all minus magnetic field (Figure 8.2). |
| Beta | A bipolar group; that is a mix of plus and minus magnetic polarities exist, with the plus well divided from the minus with one polarity in each end (E-W) of the group. (Figure 8.2). |
| Beta-Gamma | A group which is generally bipolar but which is lacking a well marked dividing line between the opposite polarity regions (Figure 8.3). |
| Gamma | A group in which the polarities are completely mixed (Figure 8.3). |
| Delta | A subclassification for non-unipolar regions. It means at least two opposing polarity umbrae are within two heliographic degrees of each other and share the same penumbra (Figure 8.4). |

The Mount Wilson Observatory analysis is accomplished by the use of a solar magnetograph. This instrument makes use of the Zeeman Effect to measure the strength and polarity of individual sunspot magnetic fields. Field strengths are normally specified in hundreds of gauss (average 1000-4000 gauss), and polarities are noted by color. Red identifies a plus polarity (directed away from the solar surface), and blue is used for negative (towards the solar surface) polarity fields. These two measurements permit the calculation of the magnetic gradients within a sunspot group. These are tabulated in gamma/km and range from near zero to one or more. Stronger gradients appear conducive to flare activity (an attempt to relieve the pressure generated by the magnetic energy density -- force line tension -- in regions of large gradients). Gradients in excess of a few tenths of a gamma/km are considered significant. The classification is completed by use of the letters p and f to identify which component of a group is dominant -- the preceding or following, respectively. For unipolar groups, these letters identify the portion of the plage containing the bulk of the spots.

Other complimentary classification systems exist for classifying spot distribution and penumbral shape. These parameters are important in more precisely assessing the flare potential of a region. One of these, attributable to Pat McIntosh of SESC (McIntosh, 1981), is now in routine use by SESS. It has a direct relationship to flare activity. This system, known as the modified Zurich class, is intended to identify dynamic solar regions. It is a three dimensional system. The first parameter is the Zurich class of the group (discussed above). The second stage of the classification deals with the penumbra of the largest sunspot.

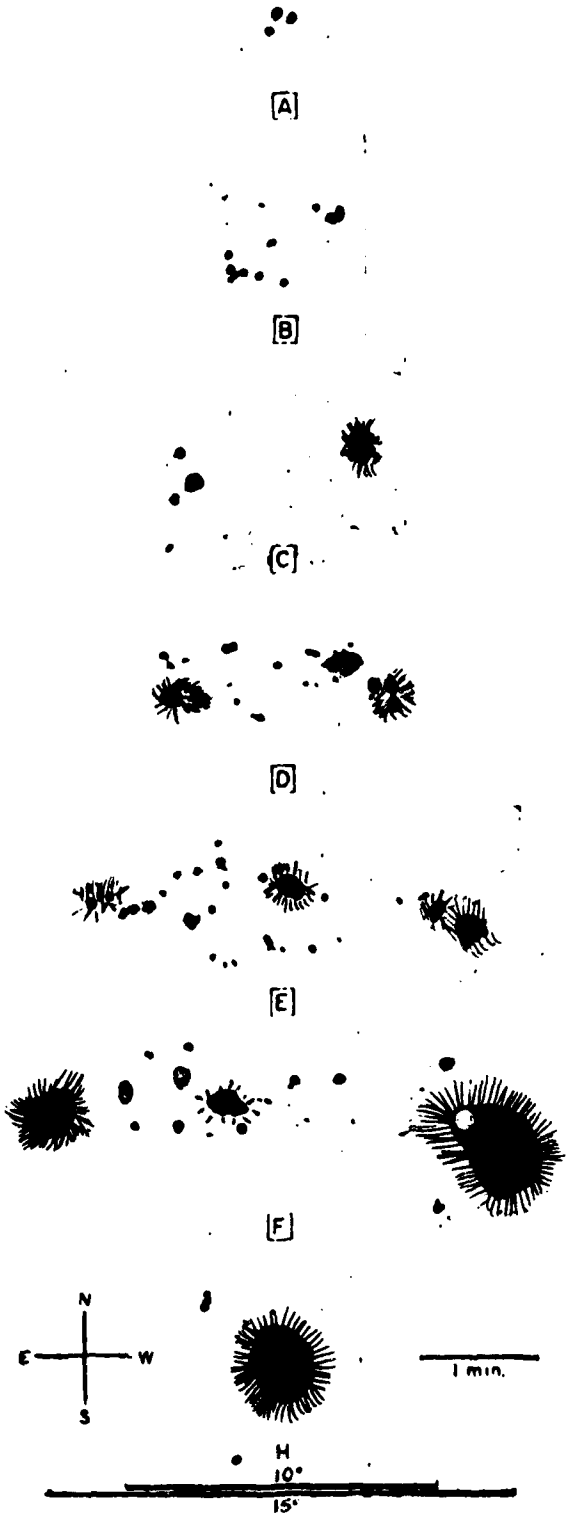



Figure 8.1 Zurich Classification System.

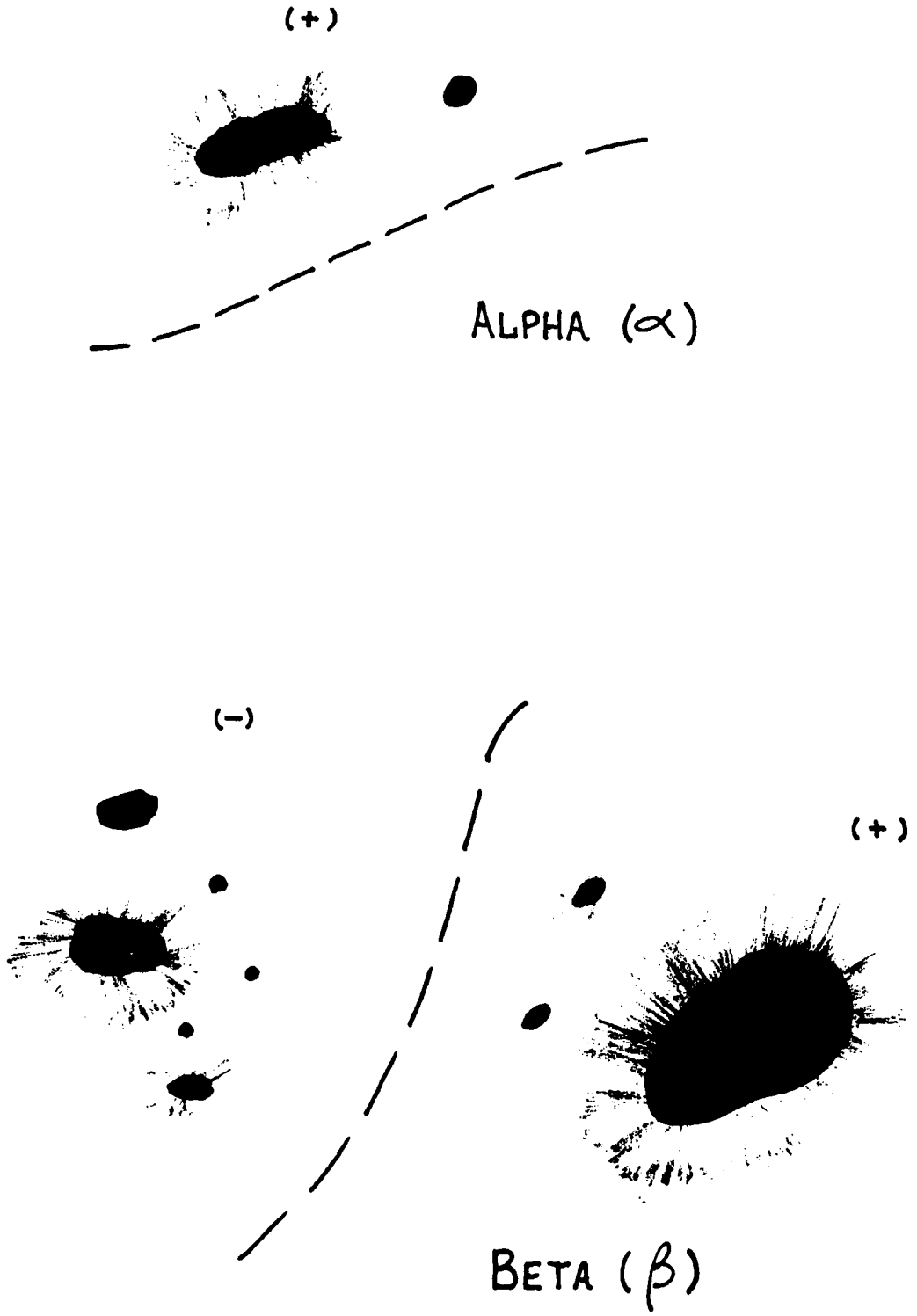



**Figure 8.2** Mount Wilson magnetic classification, alpha and beta classifications.

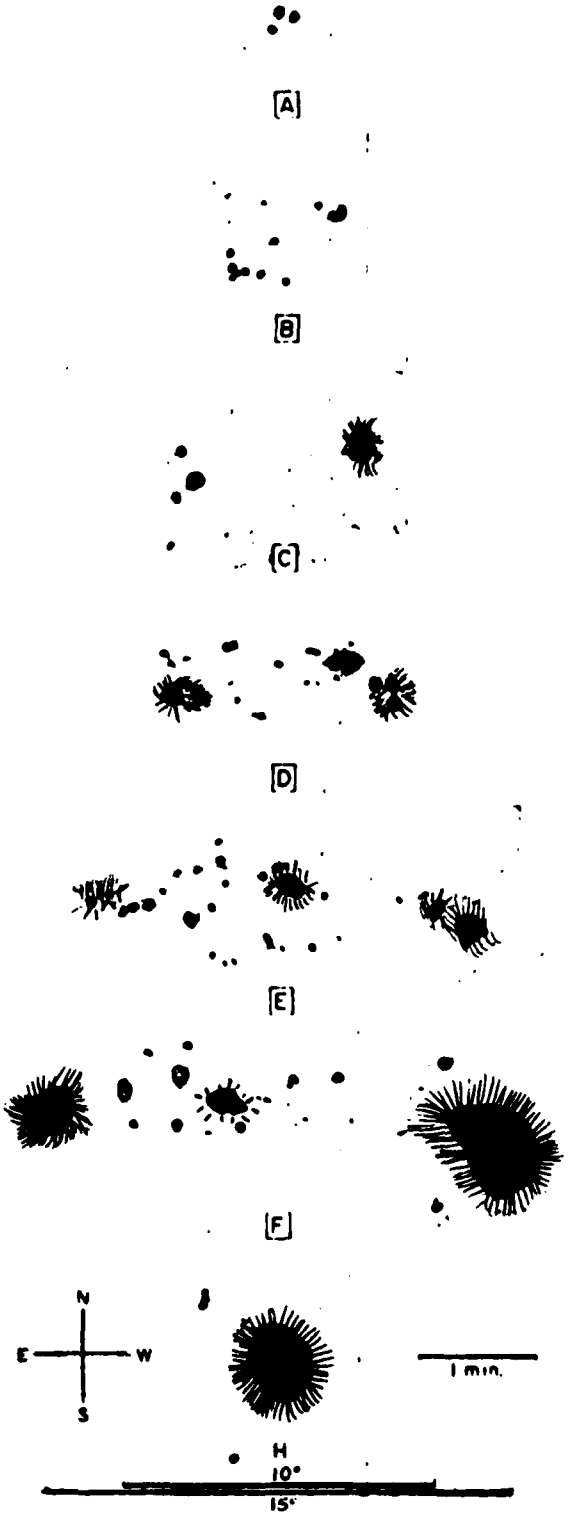
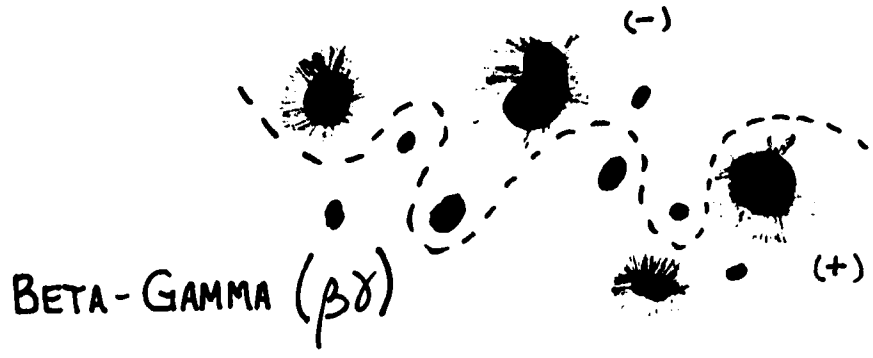



Figure 8.1 Zurich Classification System.



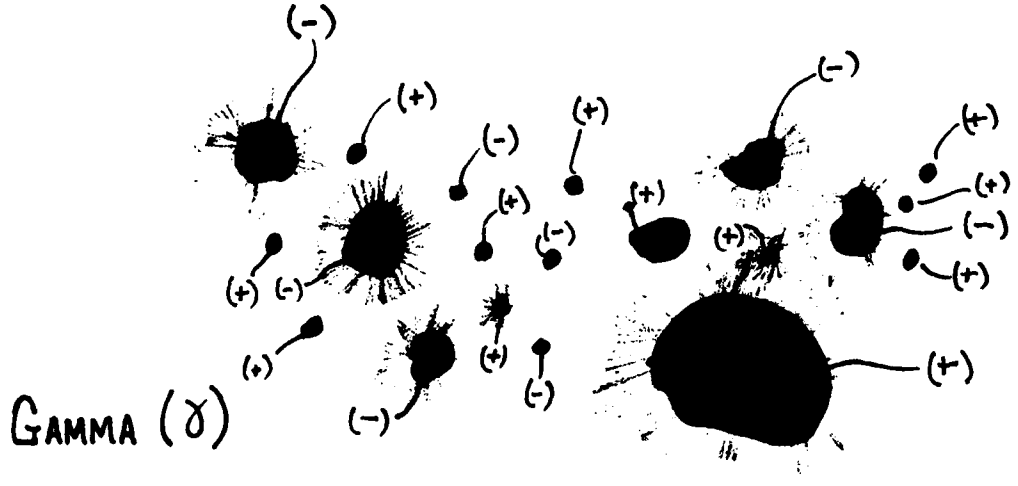




**Figure 8.3 Mount Wilson magnetic classification, beta-gamma and gamma groups.**

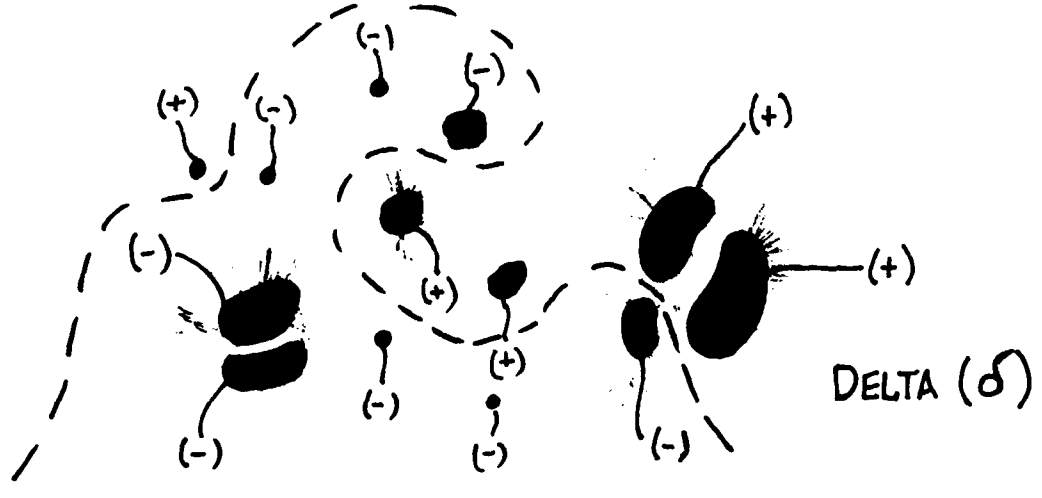



**Figure 8.4 Mount Wilson magnetic classification, delta configuration.**

X No penumbra. The width of the gray area bordering spots must exceed 3 arc seconds in order to be classified as penumbra.

R Rudimentary penumbra. Penumbra is usually incomplete, irregular in outline, and as narrow as 3 arc seconds. It has brighter intensity than normal penumbra and a mottled, or granular, fine structure. Rudimentary penumbra represents the transition between photospheric granulation and filamentary penumbra. Recognition of rudimentary penumbra will ordinarily require photographs or direct observation at the telescope.

S Symmetric. Nearly circular penumbra with filamentary fine structure and a spot diameter not exceeding 2 1/2 heliographic degrees. The umbrae form a compact cluster near the center of the penumbra. Also, elliptical penumbra symmetric about a single umbra. Spots with symmetric penumbra change very slowly.

H A large symmetric penumbra with diameter greater than 2-1/2 heliographic degrees. Other than size, it has the same characteristics as an S penumbra.

A Asymmetric, or complex, penumbra with filamentary fine structure and spot diameter along a solar meridian not exceeding 2 1/2 heliographic degrees. Asymmetric penumbra is irregular in outline or clearly elongated (not circular), with two or more umbrae scattered within it. Asymmetric spots typically change from day to day.

K A large, asymmetric penumbra with diameter along a solar meridian greater than 2 1/2 heliographic degrees. Other than size, its characteristics are the same as "A" penumbra. When the longitudinal extent of the penumbra exceeds 5 heliographic degrees, it is almost certain that both magnetic polarities are present within the penumbra, and the classification of the group becomes DKC, EKC, or FKC.

The third element of the modified Zurich system deals with the distribution of sunspots within the group.

X Single spot.

O An open spot distribution. The area between the leading and following ends of the group is free of spots, so that the group appears to divide clearly into two areas of opposite magnetic polarity. An open distribution implies a relatively low magnetic field gradient across the line of polarity reversal.

I An intermediate spot distribution. Some spots lie between the leading and following ends of a group, but none of them possesses a penumbra.

C A compact spot distribution. The area between the leading and following ends of the spot group is populated with many strong spots, with at least one interior spot possessing penumbra. The extreme case of compact distribution has the entire spot group enveloped in one continuous penumbral area. A compact spot distribution implies a relatively steep magnetic field gradient across the magnetic inversion line.

### 8.2.3 Sunspot Variations

Sunspots often show considerable motion in longitude. Some of the rotation is due to differential rotation of the solar surface. Since the leader (western-most) spot is closest to the equator, it moves slightly faster than the more poleward spots, and the group lengthens. If this effect is removed, a "proper motion", or motion relative to the differentially rotating solar surface, remains. Proper motion is apparently due primarily to growth and expansion of the magnetic field of the region. Thus, proper motion is evidence of a changing magnetic field. In a group, the leader tends to migrate slowly westward (in Carrington longitude) until the group reaches its maximum size, and then it reverses to move slowly eastward. The follower spot(s) will migrate slowly eastward for a few days and become stationary on a given Carrington longitude. This motion leads to lengthening the sunspot group by approximately eight degrees when and if the group becomes an F group. Slight latitudinal drifts are often also seen (see Figure 8.5).

Rotation of a sunspot or group of sunspots is occasionally observed. The cause of this motion is not understood, but researchers agree that it shows dynamic changes of the magnetic field in the region. Similarly, a sudden darkening of the larger umbrae is thought to portend flare activity. Some spot groups show a leader spot with the "wrong" polarity. Such groups are known as reversed polarity spot groups, and are often unstable and capable of explosive growth.

The life cycle of most spot groups is fairly simple. A group begins as a unipolar umbra or small group of umbrae with leader spot polarity (A-group). Then, follower polarity spots will develop (B-group), and later, the leader will develop a penumbra (C-group). As the magnetic field of the spot group grows, the region may expand (D, E, or F-groups), but at some point the magnetic field strength reaches maximum intensity and begins to weaken. Shortly after that time, the spot group reaches its maximum area and begins its decay. As it decays, the follower spots shrink and disappear, while the leader actually grows in size. The group becomes an H-type, with one large, symmetric sunspot with penumbra. This spot will slowly decay, marking the end of the group.

A few decaying spot groups will show sudden regrowth. This is due to the emergence of new magnetic flux and results in a sudden increase in the region complexity (and decrease in its stability). The interaction of the "old" and "new" magnetic fields can be very explosive. The formation of an arch fila-

ment system (AFS) or emerging flux region (EFR) in an old spot group can signal the increase. Depending on the location of the newly emerging spots, magnetic gradients and complexity may increase dramatically.

The so-called "neutral line" is one of the most commonly used optical tools for inferring the complexity/flare potential of a particular sunspot group. Since the magnetic fields in a sunspot group are generally closed, the field lines must be parallel to the solar surface at some point. The locus of all points at which sunspot field lines lie parallel to the photosphere is defined as the sunspot group's neutral line. Actually, the field is not neutral, its direction is intermediate between away and towards polarities. It is more correctly (but less commonly) referred to as the magnetic inversion line. Although the use of a magnetograph greatly simplifies inversion line analysis, this analysis can be done by connecting the "dots". The "dots" include filaments, plage corridors, and the corridors implied by fibril alignment. (Visible in hydrogen-alpha, fibrils are long, dark streaks. Associated with active regions, they seem to parallel magnetic field lines and emanate from large spots. They are sometimes known as superpenumbra and seem to conduct material into the individual sopts.) The sharpness of the curves and kinks in the resulting line is an indication of the magnetic complexity (and energy storage) of a particular sunspot group. For example, a generally east-west neutral line is more significant than a north-south line, because it is more unusual--indicating that opposing polarities are not well-separated longitudinally (as quiescent groups normally are).

### 8.2.4 Sunspot Model

The Babcock-Leighton model of sunspots (Babcock, 1961 and Leighton, 1964 and 1969) assumes that each magnetic cycle begins with a very simple overall solar magnetic field configuration, running from the solar rotational north pole to south pole with field lines just below the photosphere. Since the solar surface undergoes differential rotation, the equator moves at a higher (angular) rotation speed than those areas nearer the poles. The magnetic field lines are frozen into the plasma. Thus, the magnetic field is slowly wound about the sun, and the field lines become nearly parallel to the equator. The magnetic field intensity is strengthened by energy taken from the solar rotation in winding it tighter. As its strength grows, it tends to expand, and the gas volume containing the magnetic field grows in size. But, because the plasma particles are "frozen" to the magnetic field, the plasma density inside the magnetic field decreases. The magnetic field becomes bouyant, and it "floats" to the surface of the sun. The projection of magnetic field lines through the surface produces a sunspot group, one end of which has the leader polarity, and the other of which has the follower polarity. The leader polarity will be the same as the pole in that hemisphere, since the leading edge of the magnetic field line will be the more westerly (see Figure 8.6). Studies have shown that the initial location of highest probability of the magnetic field becoming bouyant is near 40 degrees latitude, with successively lower latitudes occurring as time progresses.

## 8.3 Prominences

A prominence is a relatively cool, dense condensation of gas immersed in the hotter, less dense corona. A filament is also a condensation of gas, but it is seen against the disk, where it appears as a dark object. In general, a prominence is just a filament seen on the limb. The terms filament and prominence are used interchangeably in this section, since the location of the phenomena in relation to the earth is not critical to what follows.

Prominences may be divided into two categories. Quiescent (or quiet) prominences are long-lived and slowly changing and are seen away from active regions. Active prominences are short-lived and rapidly changing and are seen in and around active regions. Prominences are phenomena of the lower corona and usually result from locally strong magnetic fields which cause and support the condensation of coronal material. These fields are either associated with a current sunspot group or remain from an old active region. Such regions project considerable material into the corona during their lifetime. Much of this material remains trapped for long periods of time. Prominences are evidence of the projection of this material into the lower corona.

Quiescent prominences are thin ropes of material (less than 300 kilometers in diameter and tens of thousands of kilometers long). Gas condenses through a quiescent prominence with a velocity of five to ten kilometers per second. Quiescent prominences form in pre-existing plage areas considerably after the plage. The region may contain sunspots when the prominence forms, but the region is typically in the mature or decaying stage. As the prominence ages, it moves out of the plage and slowly migrates towards the pole. Quiescent prominences may persist for several solar rotations.

The magnetic field lines in the corona support the condensed gas of the prominence. Closed magnetic field lines in the corona typically assume stable arch shapes when not distorted by other forces. Coronal gas is thought to condense at the top of the arch. The weight of this condensed gas pushes the top of the arch downward, to form a "magnetic saddle" which supports the material. Quiescent prominences lie along magnetic inversion lines and are supported by field lines crossing the inversion line. Quiescent prominences may be activated, or undergo changes in shape and location. The activations range from slight perturbations to complete destruction of the filament. At least four different activation mechanisms are known. They include:

Increased internal motions: Knots of material in the prominence undergo chaotic motions while that section of the prominence slowly rises and/or falls. This motion may cease and the prominence return to its non-activated state, or the motions may continue for long periods of time. In the extreme case, the gas may rise or fall up to 300 kilometers a second without destroying the prominence. These motions are thought to be due to the slow release of stored magnetic energy.

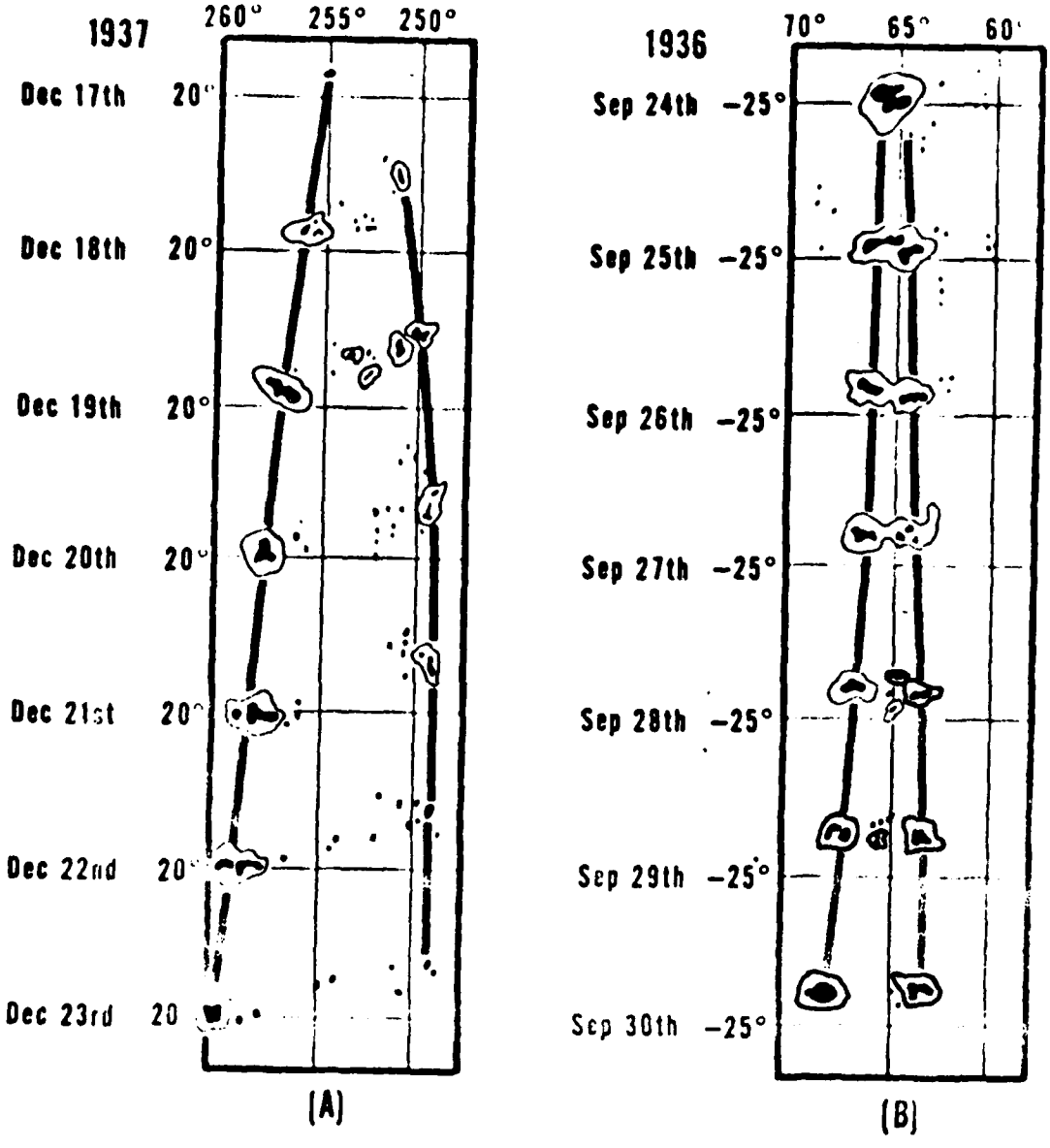
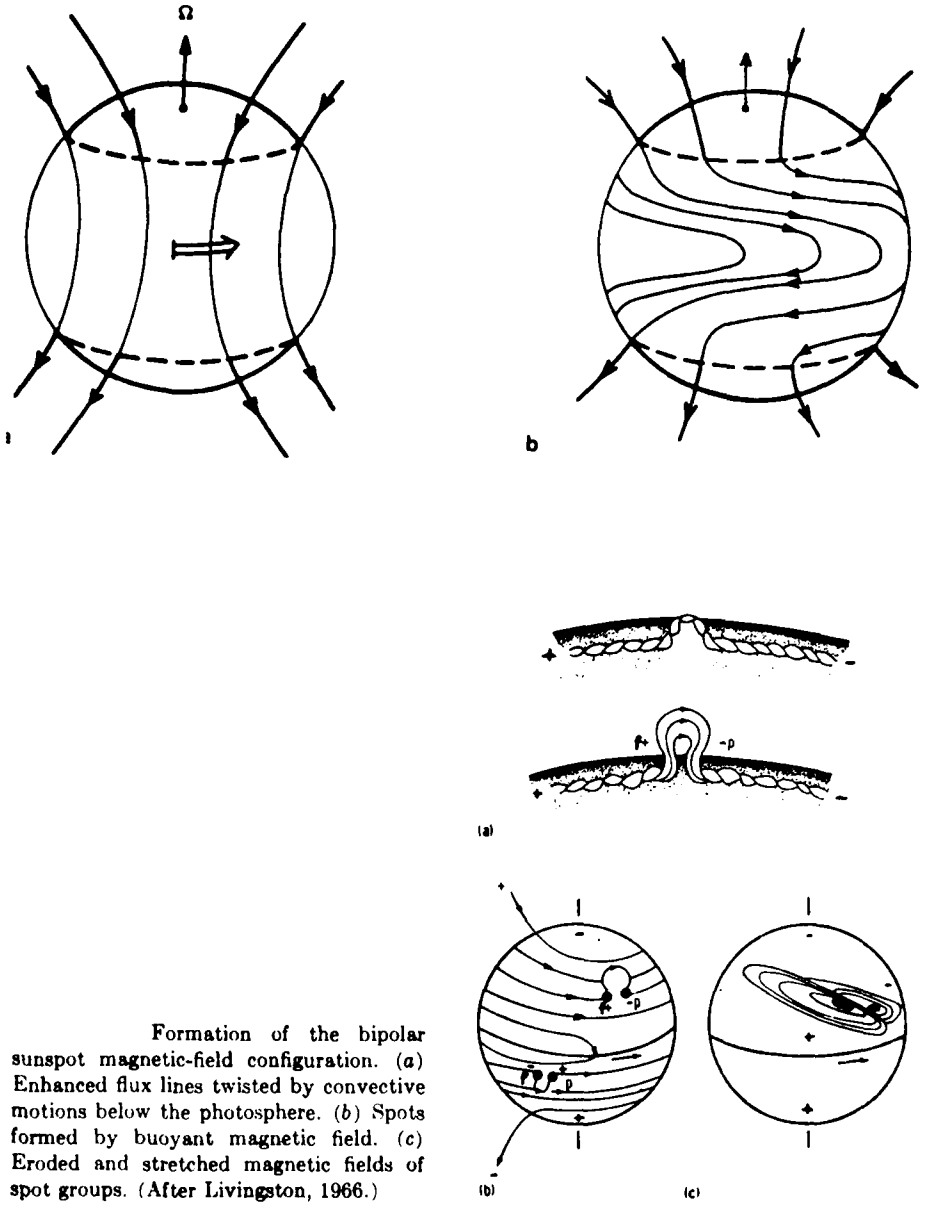



Figure 8.5 Sunspot Group Motions. (a) Proper motion in a newly formed group (b) Production of a spot group by splitting a larger spot (from Bray and Longhhead, 1964).

Winking filaments: Since we observe filaments with narrow-band-pass filters, and since a Doppler shift can change the precise frequency at which the filament is visible, a disturbance of the filament along our line of sight may cause the filament to flicker in and out of our filter's frequency width. This winking of the filament is normally due to a flare-produced shock wave which activates but doesn't destroy the filament. This activation is not due to a local release of energy, but rather to a flare on another (frequently distant) portion of the sun.




Formation of the bipolar sunspot magnetic-field configuration. (*a*) Enhanced flux lines twisted by convective motions below the photosphere. (*b*) Spots formed by buoyant magnetic field. (*c*) Eroded and stretched magnetic fields of spot groups. (After Livingston, 1966.)

Figure 8.6 Differential rotation causes the leader spot to possess the hemispheric magnetic polarity (after Gibson, 1973, and Jordan, 1981).

Eruptive prominences: Sometimes an apparently ordinary quiescent prominence will become unstable, rapidly ascend, and disappear. Generally the prominence will later reform in the same location. Studies have shown that this is common in the development of quiescent prominences. A special case of eruptive prominence is labeled a "disparitions brusque", which generally is a flare-induced activation in which the gas may reach escape velocity (618 km/sec at photosphere to 400 km/sec out in the corona).

Sinking and shrinking filaments: As the name implies, the material appears to sink lower in solar atmosphere and evaporate. While this is not as dramatic as an eruptive prominence, its consequences are more drastic because the prominence generally doesn't reform. It probably signals the weakening of the coronal magnetic field to the point where material cannot condense out on the field lines.

Active prominences are generally thinner, more intense, and shorter than quiescent prominences. Gas appears to flow along an active prominence, and the active prominence lies along the magnetic field lines and perpendicular to the magnetic inversion line. (Quiescent prominences lie along the neutral line and have gas condensing downward through them.) Several types of active prominences are particularly important.

Arch filament system (AFS): Material streams down the legs of an arch-shaped prominence with a velocity of approximately 50 kilometers a second. The top of the arch ascends with a speed of up to 10 kilometers a second. The flow of material outlines magnetic field lines which are rising in the solar atmosphere. They are indications of emerging magnetic flux in the photosphere. Arch filament systems are found early in the life of most active regions and are located between the spots of a developing bipolar group. An AFS may also form in an old active region, signalling possible instability. An arch filament system is typically 20,000 to 30,000 kilometers long. Although individual filaments endure less than an hour, the AFS may persist for several days.

Surges: Originating in point brightenings near sunspots, surges are radial projections of photospheric plasma into the corona. They are sometimes flare associated. Surge material is shot outward with a velocity of 100 to 200 kilometers a second and reaches 100,000 to 200,000 kilometers solar altitude. The material then returns along the same path, with the total surge lasting 10 to 20 minutes. The material usually does <u>not</u> reach escape velocity. Surges show a strong tendency for recurrence.

Sprays: An exploding flare mound ejects fragmented pieces of material at a velocity of 200 to 2000 kilometers a second. Sprays (if they occur) invariably occur during the expanding phase of a solar flare. Most spray material is moving faster than the escape velocity, though a small fraction usually falls back to the surface. Sprays indicate energetic solar flares and are more energetic events than surges. They sometimes occur in conjunction with the disruption of an active filament.

Loop prominences: A bright flare mound may expand into a number of loops with material streaming down both legs. Loops may also originate without the formation of a mound. Single loops may originate as surges.

During the next several hours, the loop system grows and may reach a height of 50,000 kilometers. However, individual loops do not expand, but fade and are replaced by higher ones. A loop prominence system is evidence that a very energetic solar flare has occurred.

Active region filaments: Very dark features seen on the disk inside active regions. Superficially, they resemble small quiescent filaments except for their orientation to the magnetic field. Typically, one or both ends are rooted near sunspots with material flowing along the filament into the spots. They typically last a few days. Sometimes, one end of an active region filament will connect to a quiescent filament.

Fibrils: Long, dark threads seen in hydrogen-alpha on the disk at the edges of plages and near sunspot penumbrae. They trace the magnetic field lines, and are useful in magnetic field line analysis.

## 8.4 The Solar Cycle

Early solar observers noted a cyclic variation in the occurrence of solar activity. The most common index of solar activity is the Zurich sunspot number (R). It is defined as:

$$R = K\,(10g + f)$$

where g is the number of sunspot groups; f is the number of individual spots; and K is a correction factor applied to compensate for differences in observations caused by variations in telescope size and habitual over-enthusiasm of observers. Daily Zurich sunspot numbers have been tabulated since 1749.

Heinrich Schwabe, in 1852, found that the long term (yearly) average sunspot number varied regularly with roughly a ten year period. The time of largest average sunspot number was labeled solar (or sunspot) maximum and that of lowest average number of sunspots is solar (or sunspot) minimum. Later analyses of longer samples of the sunspot cycle have revealed the average period to be 11.1 years, with a spread of 7 to 13 or possibly even 16 years. Each cycle begins at one solar minimum and continues through the following minimum. The typical cycle takes four years to rise from minimum to maximum and seven years to fall back to minimum. Solar cycle number 21, which began in June 1976, is assumed to have maxed during December 1979 with a sunspot number (yearly average) near 165. The largest sunspot cycle on record is number 19, which peaked in 1957-1958 with a value near 200 (see Figure 8.7).

The daily sunspot number varies much more widely than does the long-term average. During solar maximum, daily sunspot numbers over 100 are normal. During solar minimum, several consecutive spotless days are possible. A sunspot number forecast is published as part of the Prompt Report of Solar-Geophysical Data. It is the official long-term forecast of solar activity and is updated monthly. A typical graph is Figure 8.8.

In 1858, Carrington found that the mean latitude of sunspot groups varied in a cyclic fashion. At the beginning of a new solar cycle (just after solar minimum), the average spot group will occur near 40 degrees latitude. As the

cycle progresses, groups appear at successively lower latitudes until solar minimum, when most groups occur near the equator. If the location of sunspot groups is plotted versus solar cycle time, a diagram similar to that of Figure 8.9 (the Maunder Butterfly Diagram) results. Sunspots seldom occur poleward of 40° and never on the equator. The highest latitude spot group ever recorded was observed near 60°N.

The magnetic polarity of the solar poles undergoes a reversal every solar cycle. Early in the cycle, the polarity at one pole is positive (away from the sun), the other negative (towards the sun). Near solar maximum, there seems to be no dominant polarity at either pole. A reversal of solar polar polarities begins to be apparent within 2 years following sunspot maximum. Typically, one pole will show/complete reversal as much as a year or more before the other hemisphere. The solar northern hemisphere had completed its reversal (to towards the sun polarity) by mid 1980, while the south polar reversal was not complete until late 1982. By the end of the sunspot cycle, the polarity of the poles is reversed from that at the beginning. During any solar cycle, the leader sunspots normally have the polarity that their hemisphere had at the beginning of the cycle. Magnetic field observations have shown the weakening magnetic field of a decayed preceding sunspot drifts toward the solar equator, while those of decayed trailer spots drift toward the pole. The drift of the trailer spots is thought to over-neutralize the polar field and result in the reversal of the polar magnetic field direction. The drift of preceding spots toward the equator results in no net polarity in that region, because the preceding spots in the opposing hemispheres cancel each other. The effect of this slow reversing of overall solar hemispheric polarity is to unwind the stretched magnetic field lines and decrease the stored energy available to produce solar active regions.

The polarity of most of the preceding sunspots in either solar hemisphere reverses near solar minimum. Thus, each new cycle begins with leader spot polarity reversal. There is some overlap between the end of one cycle and the beginning of the next, so at solar minimum the sun may simultaneously have "old cycle" groups near the equator (with old cycle leader polarity) and "new cycle" groups near 40° (with new cycle leader polarity). Neither of these is considered a reversed polarity group, while a "new cycle" group (near 40°) with "old cycle" polarity (that of a group near the equator), or vice versa, would be reversed. The reversal of polarity, coupled with the sunspot cycle, is referred to as the 22 year cycle. The combined Maunder Butterfly Diagram and reversed polarity information is shown in Figure 8.10.

## 8.5 Radio Astronomy

The earliest known attempt to detect solar radio emission was by Sir Oliver Lodge about 1900. Because of the low sensitivity of his equipment and severe man-made electrical interference, he was not successful. Various other astronomers attempted in subsequent years to detect solar radio emission but also failed until World War II. In 1940, the Germans identified interference on their 1.7 meter (175 MHz) radar equipment in Denmark as produced by the sun. On 26 Feburary 1942, several 5 meter (55 to 80 MHz) British radars suffered such severe interference that they were useless for detecting and

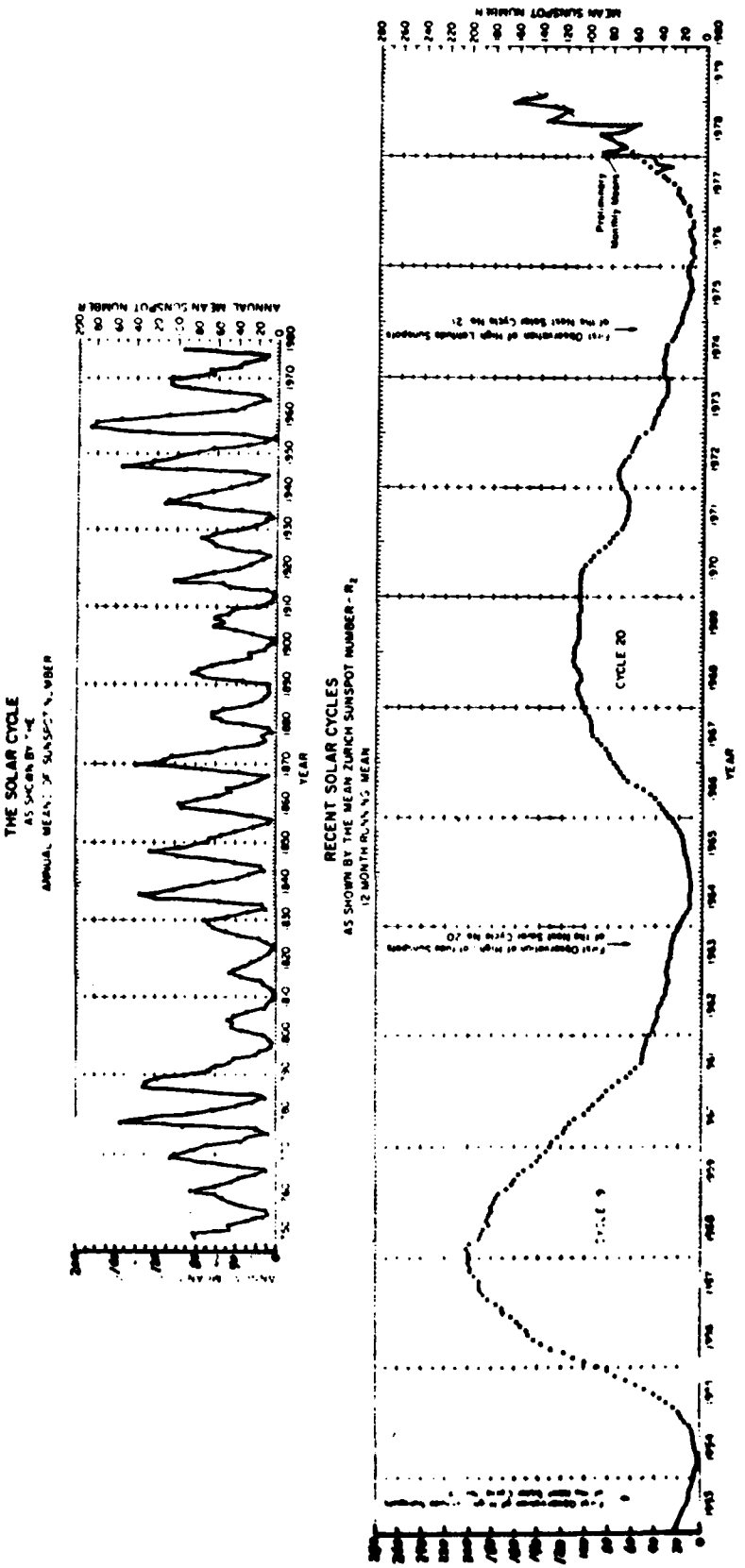



Figure 8.7 Plot of the Annual Zurich Sunspot Numbers.

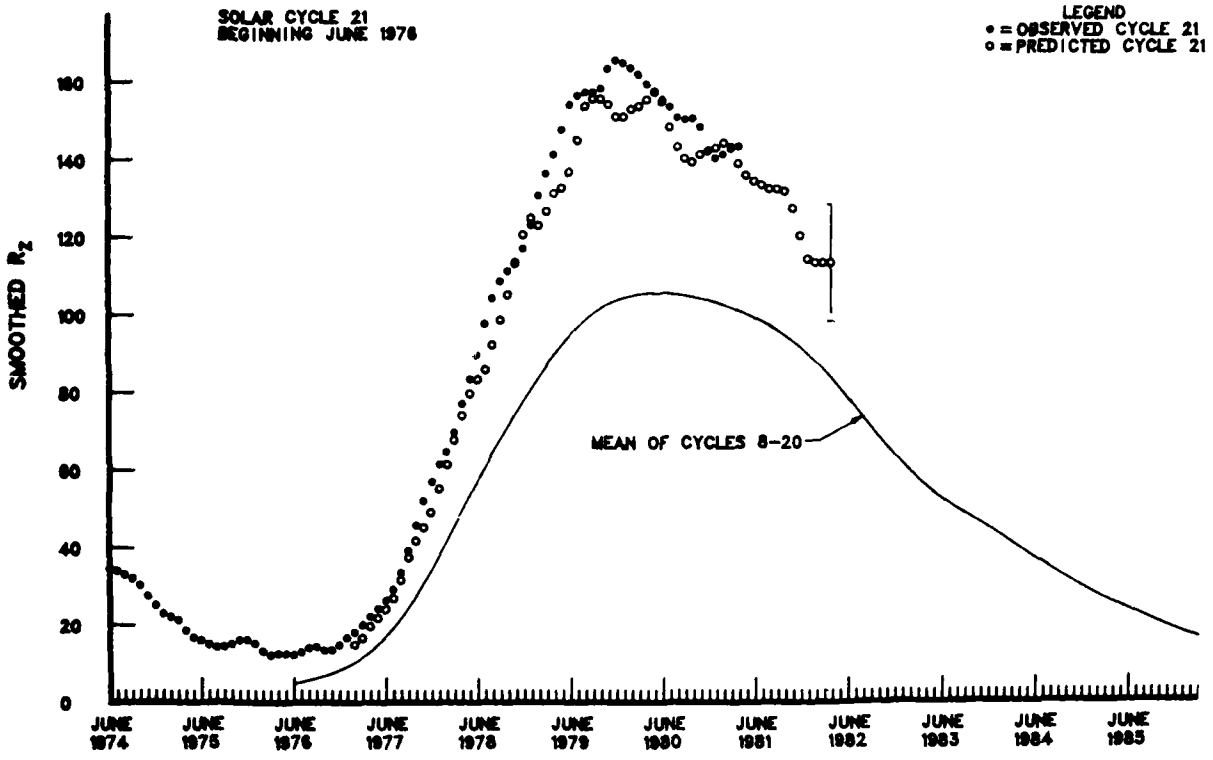



Figure 8.8 Long-term Forecast of Sunspot Number (from Caffey, 1981).

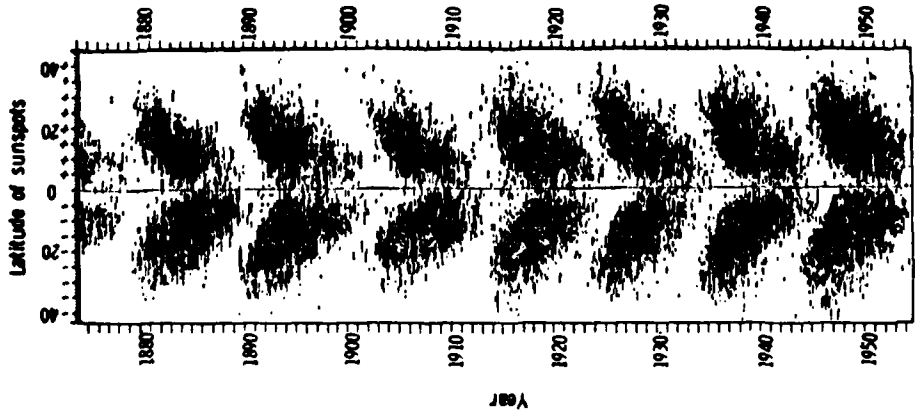



Figure 8.9 Maunder Butterfly Diagram (from Gibson, 1973).

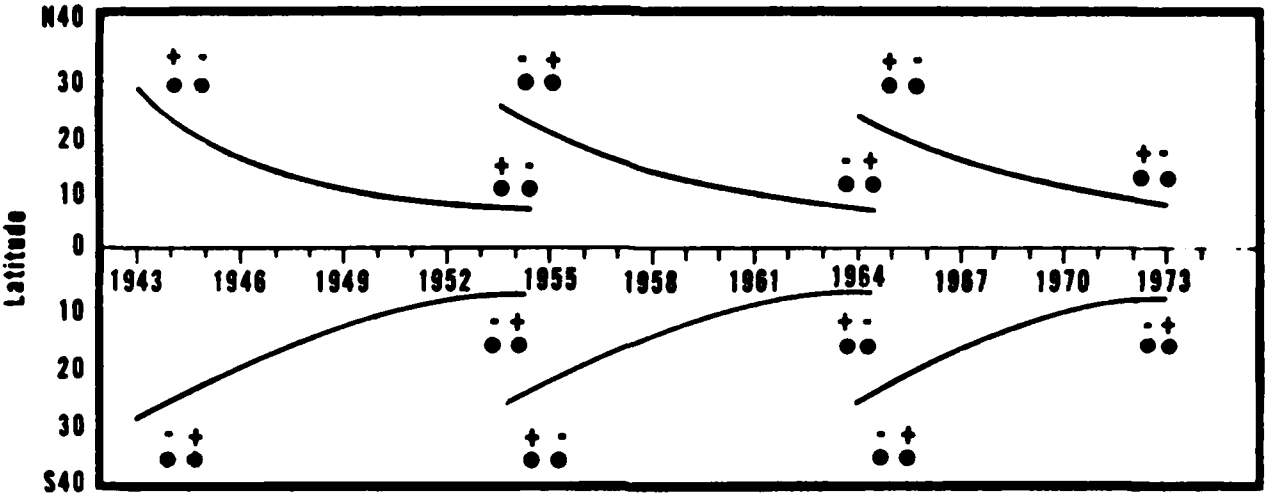



Figure 8.10 Sunspot Latitude Migration and Leader Polarity Reversal.

tracking enemy aircraft. At first, it was thought the Germans had developed a new *radar jamming system*, but Stanley Hey concluded the radio emission was associated with a large active region on the sun. G.C. Southworth, of Bell Telephone Laboratories detected steady solar radio emission near 10 cm. The first published report of solar radio emission was by an amateur, Grote Reber in 1944, who detected strong solar radio noise on a wavelength of 187 centimeters during September 1943. The earlier observations were not published until after the end of the war due to their military significance.

Since World War II, solar radio emission has been studied in great detail by scientists around the world. The largest early contributions to solar radio astronomy were made by the U.S., Australian, and British radio astronomers, who had a wealth of war surplus radar receivers available for their studies. Continued emphasis on radio astronomy has helped the Australians (and U.S. to a lesser extent) continue to lead the world in solar radio research.

Two basic techniques are currently used in observing solar radio emissions. Both use an instrument called a radiometer to measure the strength of radio emission. One method is called fixed frequency monitoring, and the other is sweep frequency monitoring.

Fixed frequency monitoring involves setting the receiver frequency at a particular value and recording the variation in intensity with time. This is much like tuning an AM radio to a station and listening to how the strength of

the station varies with time (without adjusting the volume). A significant solar event is like a burst of static during a thunderstorm. We measure the intensity of the outburst on several different frequencies chosen to span the range of solar atmospheric emission. This is a radio burst on a fixed frequency. We also measure the total output of the sun on each frequency when such an event is not in progress. This measurement is called the integrated radioflux on that frequency. It provides a "quiet" background for long-term comparison.

Sweep frequency monitoring is done by recording intensity versus frequency while sweeping back and forth across a particular range of frequencies. The sweeping of frequencies is similar to moving the tuning knob of your AM radio back and forth continuously. We divide these plots of frequency versus time into burst types according to their appearance.

Radio emission from the sun may be divided into three sections: the quiet sun, the slowly varying component, and the rapidly varying component. No matter which component we observe (they are impossible to separate), it is the frequency (or wavelength) which determines the altitude we are monitoring. Electron density decreases vertically in the solar atmosphere. Consequently, the frequency emitted by a given layer decreases with altitude. (The frequency emitted at a given level is closely related to that level's plasma frequency.) Lower frequencies originate from higher in the solar corona. Monitoring a number of properly chosen frequencies gives us a vertical "sounding" of the solar atmosphere. Since solar radio emission is blackbody radiation from a gas whose temperature and density change with height, the measurements of emission at various frequencies are effectively measurements of emission at various heights. Wavelengths less than 1 cm (frequency over 30,000 MHz) are assumed to be photospheric (temperatures of 6000°K). Those up to 30 cm (1000 MHz) are assumed to be chromospheric (temperatures of 40,000°K); those up to 3 meters (100 MHz) are from the inner corona; and those from longer wavelengths (lower frequencies) are from the outer corona (coronal temperature 1,000,000°K). This change in frequency with height above the photosphere is shown in Figure 8.11. Our optical data is, of course, limited to the lower chromosphere and photosphere.

### 8.5.1 Basic Component

Radio emission from the quiet sun is believed to be thermal emission. The only time we can measure this basic (or, quiet sun) component is after several months with no sunspots, or plage. Since this occurs only rarely, the basic component is determined by plotting the measured daily integrated flux at a frequency versus the Zurich sunspot number and extrapolating back to a sunspot number of zero. The basic component of the quiet sun slowly varies over the solar cycle for longer wavelengths. Below one centimeter, it is unchanged. The basic component of 10.7 cm solar emission changes from near 70 solar flux units (SFU) during solar minimum to near 130 SFU during solar maximum (see Figure 8.12). An SFU (solar flux unit) = 1 X $10^{-22}$ watts/$m^2$/sec.

The sun oberved at radio wavelengths is larger than the sun observed in optical wavelengths. Since radio emission at lower frequencies originates at

a higher level in the solar atmosphere, the radio "disk" at that frequency is larger. How high above the photosphere a particular frequency is emitted is dependent on the level of solar activity, so how large the radio sun appears depends on solar activity. For a quiet sun, emissions of 150 MHz originate near one solar radius above the photosphere.

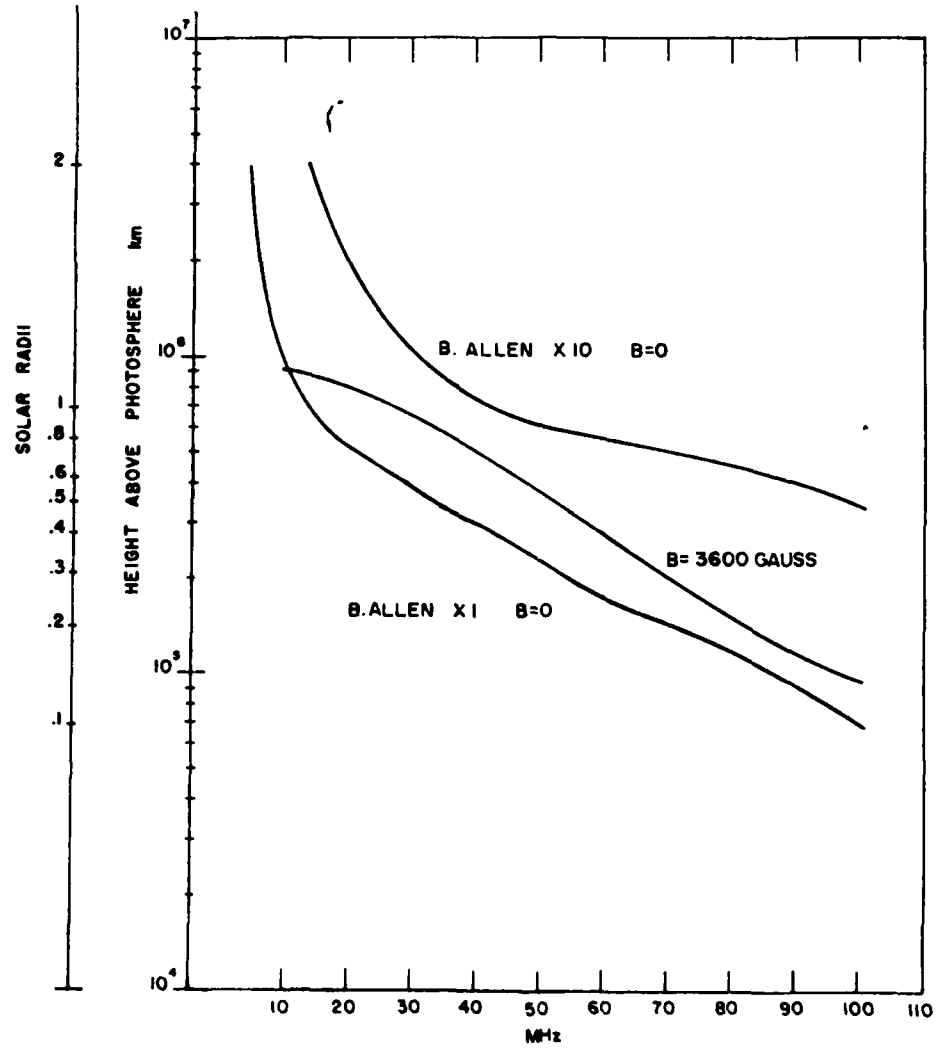



Figure 8.11 Change in Plasma Frequency with Altitude Above the Photosphere (from Eis and Rickard, 1978).

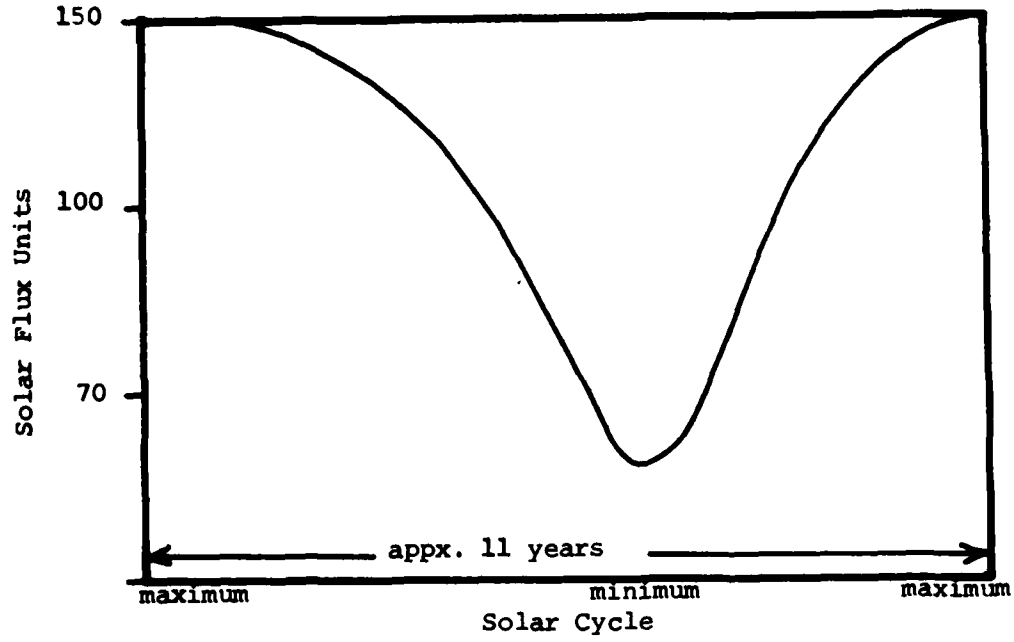



Figure 8.12 Quiet Sun Radio Variations.

Limb brightening occurs at radio frequencies. Optical limb darkening occurs, because optical limb emissions originate higher in the photosphere where the sun is cooler. Radio frequency limb brightening occurs because radio emissions originate higher in the corona than disk emissions. Thus, the temperature (kinetic temperature) increase with altitude in the solar atmosphere combined with the steadily falling electron density causes radio limb brightening.

### 8.5.2 The Slowly Varying Component

Solar radio emission in the wavelength range of one centimeter to one meter exhibits slow variability. The variations in emission strength show a "period" of roughly one month (Figure 8.13). This slowly varying component (S-component) is strongest on wavelengths between five and eight centimeters. The S-component causes the daily variation in integrated flux measurements, such as the Ottawa 2800 MHz (10.7 cm) daily flux value. (Here, integrated flux means all disk contributions at this frequency are added together.)

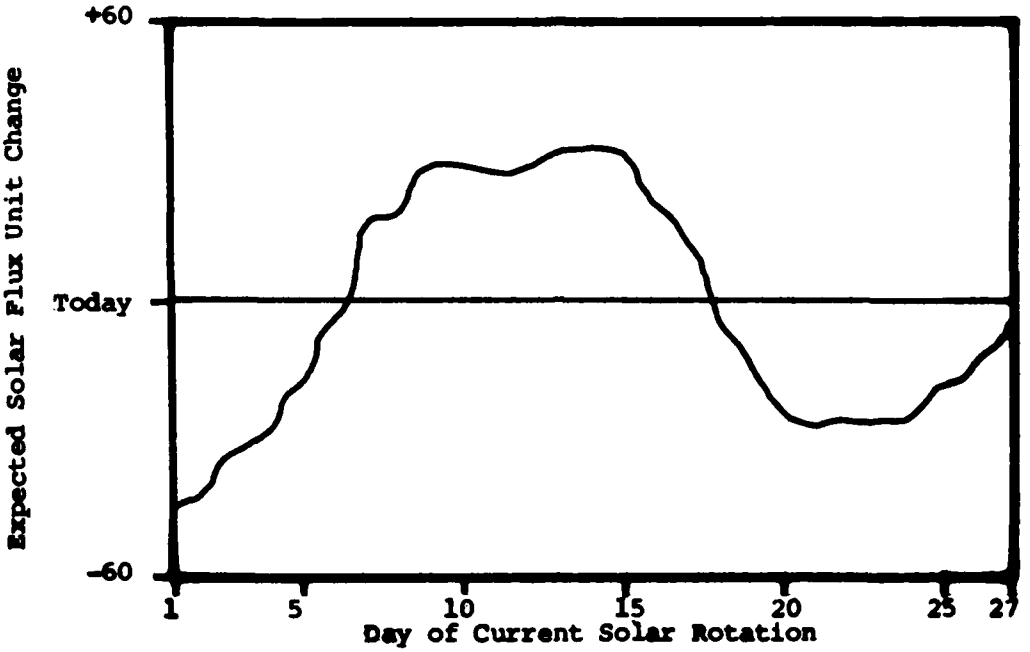



Figure 8.13 Slowly Varying Component.

Radio interferometers have been used to compare the locations of intense background emission to optical active regions. Optically complex, active regions underlie regions which are strong emitters in centimeter wavelengths. Although early correlations showed rough agreement between sunspot area and intensity of the S-component emission from the region, S-component emission remains high for weeks after the sunspots die. This implies a possible correlation between S-component and plage area. S-component emission is also believed to come from localized condensations of high electron density in the corona. These coronal condensations occur over active regions in the relatively strong magnetic field. The strong coronal magnetic field is typically maintained for several weeks after the sunspots disappear and until the photospheric magnetic field concentration that produced the plage dissipates.

Both the basic and slowly varying components of the solar radiowave emission vary with the solar activity level. The slowly varying component varies with the daily sunspot/plage area and is determined from daily integrated flux measurements. The quiet sun value varies with average solar "temperature" over a long time span, and is determined using a long term average of daily integrated flux measurements. The origin of this component is probably tied to the underlying causes of the sunspot cycle itself.

### 8.5.3 The Active Radio Sun

A solar radio event occurs when the corona is disturbed by energetic charged particles and/or a shock wave. Any charged particle which is accelerated (speeded up, slowed down, or changed in direction of travel) gives off energy in the form of an electromagnetic wave. Any solar flare produces some energetic electrons and protons, and these particles travel away from the location of the flare in all directions. These particles do not propagate outward through the corona with constant speed and direction of travel, but undergo many accelerations. Any solar flare also produces shock waves which similarly move outward from the location of the flare. These shock waves give up energy to protons and electrons to locally produce energetic particles. These particles, in turn, give up some energy and produce radiowaves. The electromagnetic waves produced by accelerations of energetic particles and by the shock produced energetic particles are what we track in measuring active solar radio emissions (Figure 8.14). Monitoring the emission at several frequencies permits a height analysis of flare effects in the solar atmosphere.

The active component is a result of a least three processes: bremsstrahlung, plasma, and synchrotron emission. For the latter two processes, frequency emitted decreases with altitude (for bremsstrahlung, it depends on the energy of the collision). We have already seen that plasma frequency decreases with altitude. For synchrotron radiation, the frequency produced is proportional to the magnetic field strength divided by the mass of the particle. Since the magnetic field strength decreases outward from the sun, the emitted frequency decreases as the particles move outward. Since plasma density and magnetic field strength do not decrease uniformly with height, this relationship is good in general, but may be violated for a particular case.

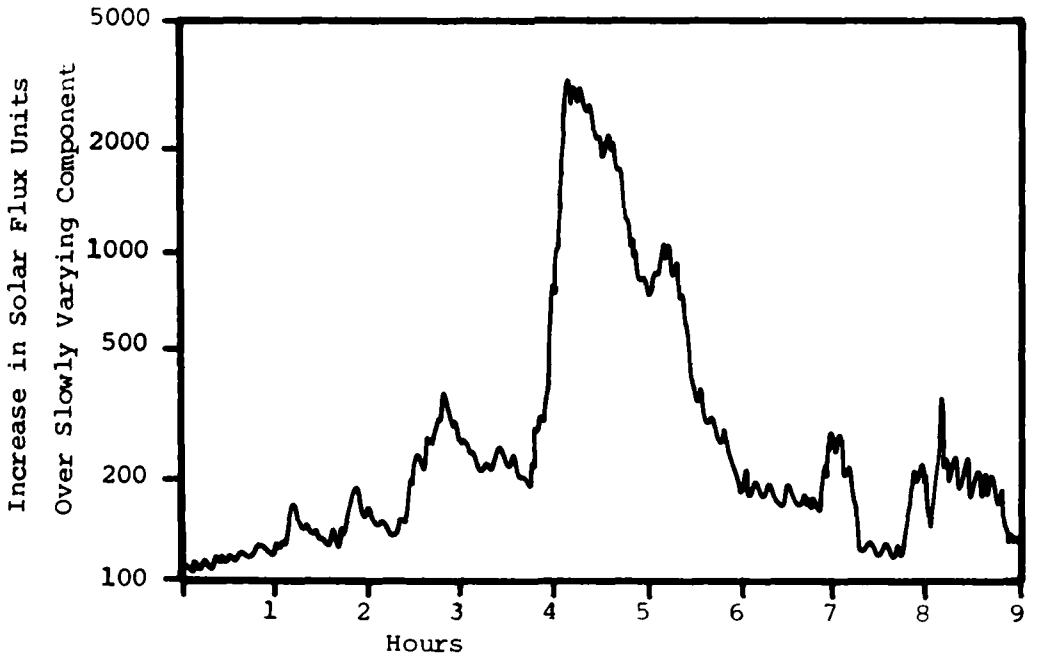



**Figure 8.14 Active Sun Component of Solar Radio Emission.**

### 8.5.3.1 Noise Storms

The great outburst of radio noise that jammed the British radars in 1942 and led to the discovery of solar radio emissions was what we now call a noise storm. Radiation of this type is usually found only at wavelengths longer than one meter (frequencies below 300 MHz). In that region of the electromagnetic spectrum, it constitutes by far the greatest part of the observed non-thermal radio flux. Near the time of sunspot maximum, noise storms are in progress about 10% of the time. During a period of intense activity, hundreds of storm bursts may occur each hour. Individual storms generally last from a few hours to several days, with the intensity of the received flux rising as high as 1000 times that of the quiet sun.

In most cases, the storm radiation consists of a mixture of two distinct components. There is a background of broad-band, relatively steady emission (the noise storm continuum) on which are superimposed large numbers of short bursts. The continuum normally covers a frequency range of 100 MHz or more and lasts hours or days. A typical burst has a bandwidth of a few MHz and lasts only a tenth of a second to a few seconds. High resolution radio telescopes indicate that noise storms come from small regions high in the corona over active regions. As many as two-thirds of all active regions with at least one well developed spot develop noise storm emission regions sometime during their lifetimes.

Noise storms are not traced back to particular flares, and so are not normally included in the time sequence of solar flare actions. Also, the occurrence of a noise storm does not appear to be an indicator of the amount of energy released by a flare. They do appear to be associated in some way, however, as 80 to 90% of radio noise storms follow flare activity by 2 hours or less. Theories of burst production involve the trapping of flare-energized particles by the magnetic field in the outer corona. The existence of a noise storm should be considered evidence of complex coronal magnetic fields rather than of a large solar flare.

Noise storm observations are used to assist customers who are experiencing radio frequency interference (RFI). Systems in the very high frequency (VHF) range (30 to 300 MHz) are particularly susceptible to interference by solar radio noise storms. This problem especially affects NORAD radar sites (BMEWS, PAVE PAWS, etc.). Radio observatories monitor 245 MHz for noise storms.

### 8.5.3.2 Sweep Radio Bursts

Sweep frequency radio bursts (other than type I noise storms) are classified by spectral type; that is, by their intensity versus frequency variation with time. Spectral typing of radio bursts is done using measurements from a swept frequency interferometer. Swept frequency means that the instrument runs through a range of frequencies and measures signal strength on each frequency. The output is displayed as a graph of signal intensity (darkness of plotted data) on a frequency versus time plot. By convention, the display has frequency increasing downward. Since plasma and synchrotron radiation frequencies increase with decreasing altitude in the solar atmosphere, the plot effectively shows altitude increasing to the top.

The AWS swept frequency interferometers cover approximately the frequency range 25 to 75 MHz (the outer corona).

Type I. These are noise storm spikes of a few seconds duration. Since much of this activity occurs in the 100-200 MHz range, AWS gear cannot normally detect these brusts.

Type II. In 1947, fairly intense radio bursts on meter wavelengths were reported. These bursts had durations of a few minutes, and their onset was progressively delayed as the frequency was decreased. These bursts were designated as type II, sometimes called slow drift bursts. A typical type II burst has a high frequency cutoff near 100 MHz with no burst detected at higher frequencies (for the fundamental frequency). In addition to the fundamental frequency, about one half the cases have a harmonic frequency, (double the frequency of the fundamental). Several mechanisms have been suggested to produce the harmonic frequency, but most studies are only concerned with the fundamental. Since the frequencies drift from high to low frequency, a low range swept frequency instrument like those run by AWS units will see the fundamental first. The remainder of this discussion is confined to the fundamental frequency.

More than 90% of type II bursts are clearly flare associated. The other 10% of the bursts may have been produced by flares behind the limb. So probably all type II bursts are flare-associated. Most flares do not produce type II bursts. They are primarily associated with large flares. Thus, a type II burst is a fairly good indicator of an energetic solar flare. Type II bursts are thought to be due to a solar flare-produced shock wave moving through the corona. The shock wave results in plasma frequency emission as it moves out. Typical bursts show drift rates of 0.05 to 1.0 MHz/sec. Theoretically, these drift rates can be used to determine the speed of the shock front in the solar atmosphere. They seem to average about 1000 km/sec. Perhaps 60% of type II bursts are preceded by type III bursts.

Type III. The most frequent sweep frequency radio events are classified as type III, sometimes called fast drift bursts. They occur alone or in groups, and there have been days when more than 50 such bursts were recorded. However, there are also days when no type III bursts are observed even when active regions and flares are observed. Type III bursts are predominantly region associated rather than flare associated; that is, they signal that activity is occurring in certain regions.

A typical type III burst has a short duration and fast frequency drift (1-70 MHz/sec) from high to low frequency. It lasts a few to a few tens of seconds and can be observed from a few hundred MHz (the base of the corona) to tens of KHz. The burst is produced by a stream of high energy electrons which cause synchrotron radiation and scattering as they move through the solar corona. These electrons have energies greater than approximately 20 KeV and travel at approximately one third the speed of light. They are commonly called "relativistic" electrons.

A subclass of type III radio bursts is the so-called "inverted U burst". In it, the frequency sweeps from high to low, and then reverses to sweep back toward higher frequency. The electrons which produce these bursts are thought to be trapped in magnetic arches which extend high (possibly even tens of solar radii) into the corona. These inverted U bursts, while often noted in technical publications, are of little interest to us except in that they imply the existence of magnetic arches in the corona.

Type IV. The most commonly studied solar radio bursts are type IV bursts. Type IV bursts in the frequency range of AWS swept frequency equipment are long-lived enhancements in emission intensity over practically the entire frequency spectrum. These events differ from type I noise storms in four important ways:

(1) Only a continuum is present, free of the individual storm bursts of type I emission. The emission is smooth and continuous across the band of frequencies;

(2) The intensity of the emission is higher than for type I noise storms;

(3) The duration of the type IV bursts is from 10 minutes to several hours instead of hours to days for type I; and

(4) Type IV bursts are associated with flares, usually large ones, while type I noise storms show no flare association.

Type IV bursts have also been studied in the microwave bands. We do not monitor these emissions with swept frequency radiometers, so they are of only passing interest. Their production methods differ from other type IV bursts. Type IV bursts are confined to frequencies below 200 MHz and start a few minutes after the onset of the associated flare with increased delay of onset for lower frequencies. The type IV burst may evolve into a noise storm lasting days. The emissions are thought to be synchrotron and bremsstrahlung emission from particles which are following the shock wave outward or which receive energy from the shock wave as it passes and which are tied to the coronal magnetic field. These particles give up their energy as radio waves at a frequency dependent on the magnetic field strength where they are trapped. The speed of the shock wave is comparable to that which produces a type II burst, 100 to 1000 kilometers per second. About 50% of all type IV bursts have a type II precursor.

Type V. Type III bursts are sometimes followed at metric wavelengths by a short-lived (1-3 minutes) broad band continuum which is called a type V burst. Type V bursts are thought to result from energetic electrons which have been trapped by the magnetic field in the corona, and which produce radio emission through synchrotron or plasma radiation. Type V bursts are always associated with type III bursts. They are often the most intense bursts observed because of the efficiency of synchrotron emission. Figure 8.15 is a comparison of the various types of sweep frequency bursts.

Not all flares are accompanied by radio bursts which can be classified by the scheme just described. Indeed, many flares occur which have no significant radio frequency counterparts. Significant radio emissions often occur without an optical counterpart. Large flares, those which release vast quantities of energy at optical wavelengths, are statistically more probable to have associated large radio emission. Not all of the types of radio bursts occur with all large flares.

The disturbances which produce radio bursts may sometimes be associated with an optical and x-ray flare occurrence. The streams of energetic electrons which produce type III and type V bursts may have the same origin time as the associated flare. The traveling shock wave associated with type II and type IV emissions has likewise been shown to originate in the time and vicinity of the flare. Therefore, the delays for onset of various bursts at a given frequency are due to the time it takes the flare disturbance to propagate to the appropriate level of the corona. The onset of the burst at lower frequencies is delayed due to the longer time required for it to reach higher coronal layers.

The typical flare time sequence is:

(1) Beginning of optical and x-ray flare (and production of stream of energetic electrons and shock wave);

(2) Type III burst produced as electrons move outward (elapsed time of 1 minute);

(3) Type V bursts produced by trapped electrons (elapsed time of 3 minutes);

(4) Type II burst produced as shock moves outward and excites electrons (elapsed time of 10-15 minutes);

(5) Type IV burst begins due to trapped electrons which received energy from shock wave and following flare plasma cloud (elapsed time of hours); and

(6) Type IV may become noise storm with type I bursts (elapsed time of days).

It is important to note that the radio bursts are produced by energetic electrons and/or shock waves. They are often associated with the production of energetic protons, but this is not a confirmed physical relationship. The overall sequence of emissions related to a large flare is summarized in Figure 8.16.

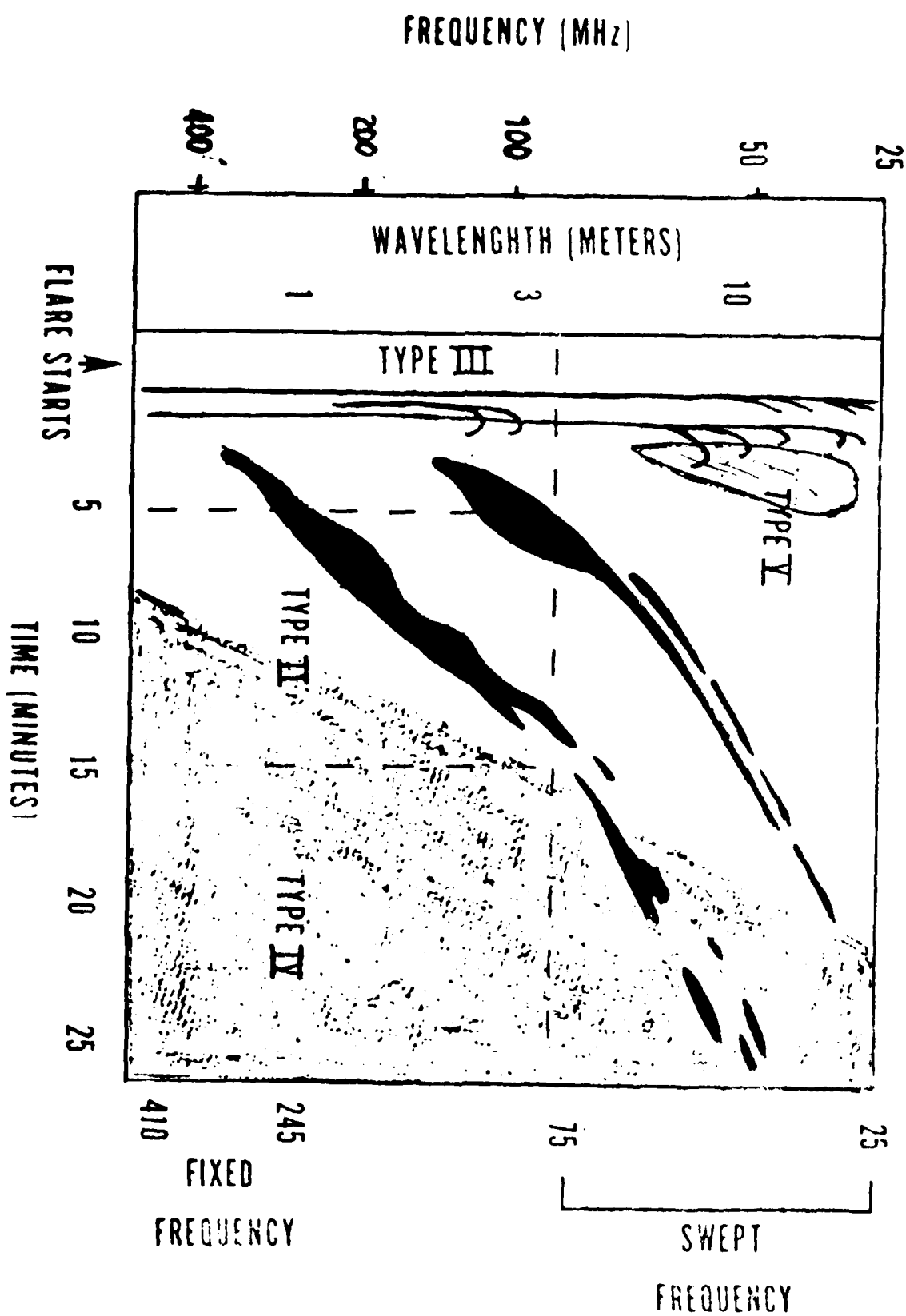



Figure 8.15 Comparison of Sweep Frequency Bursts.

8.5.3.3 Fixed Frequency Bursts

If we use a fixed frequency radiotelescope to monitor solar activity, we find that the energy flux on any frequency varies on a very short time scale. The short-term, often massive increases in output on a frequency are also called radio bursts -- discrete (or single) frequency bursts. AWS solar radio observatories record the energy output on a strip chart of energy versus time. The difference between the background flux before the burst (integrated flux) and the flux during the burst is the burst magnitude.

Solar radio bursts are indicators of short-lived disturbances in the solar atmosphere. The strength of the disturbance is equated to the energy output of the burst. Bursts can be classified by their rise and fall times and by the burst magnitude at peak emission.

AFGWC divides radio bursts into three basic categories based on the peak energy flux: below threshold, minor, and major. The lower threshold for a minor radio burst is 500 SFU. This has been found to be the lower threshold for significant effects on most radio sensitive systems. A burst maximum of less than 500 SFU is below threshold. If the burst maximum reaches 10,000 SFU, it is considered a major burst. This value equates to severe radio frequency interference (RFI) on many RF systems. These thresholds are based on average impact on all systems. Since the impact on various systems varies widely, division into minor and major bursts is, at best, a rough guideline. One system may be tolerant to a few thousand SFU without significant effects, while another may be rendered useless by a few hundred SFU. Each system is more sensitive to one range of frequencies than another, and directional systems complicate matters even more by adding antenna pattern effects.

The total energy of a burst is represented by the area under the energy flux curve. If we measure the flux under the curve during the burst, we have the integrated flux density (in watts per square meter) of the entire event. Active radio emission can be represented by a parabolic climb, and an exponential decay. The integrated flux density can be approximated by 1/3 FT where F is the peak flux value, and T the rise time. This is the semi-integrated flux density and is used in various prediction studies.

Classification of bursts by rise and fall yields four distinct categories (Figure 8.17):

(1) Simple bursts are characterized by a rapid rise and slow decay. The rise is smooth and rapid (up to several minutes long) and returns to the background level in a few minutes. These are generally flare associated and are thought to be due to synchrotron emission. Impulsive simple bursts, those which rise in less than a minute, are normally associated with "bright" optical flares. A complex burst may be modeled as a series of superimposed simple bursts.

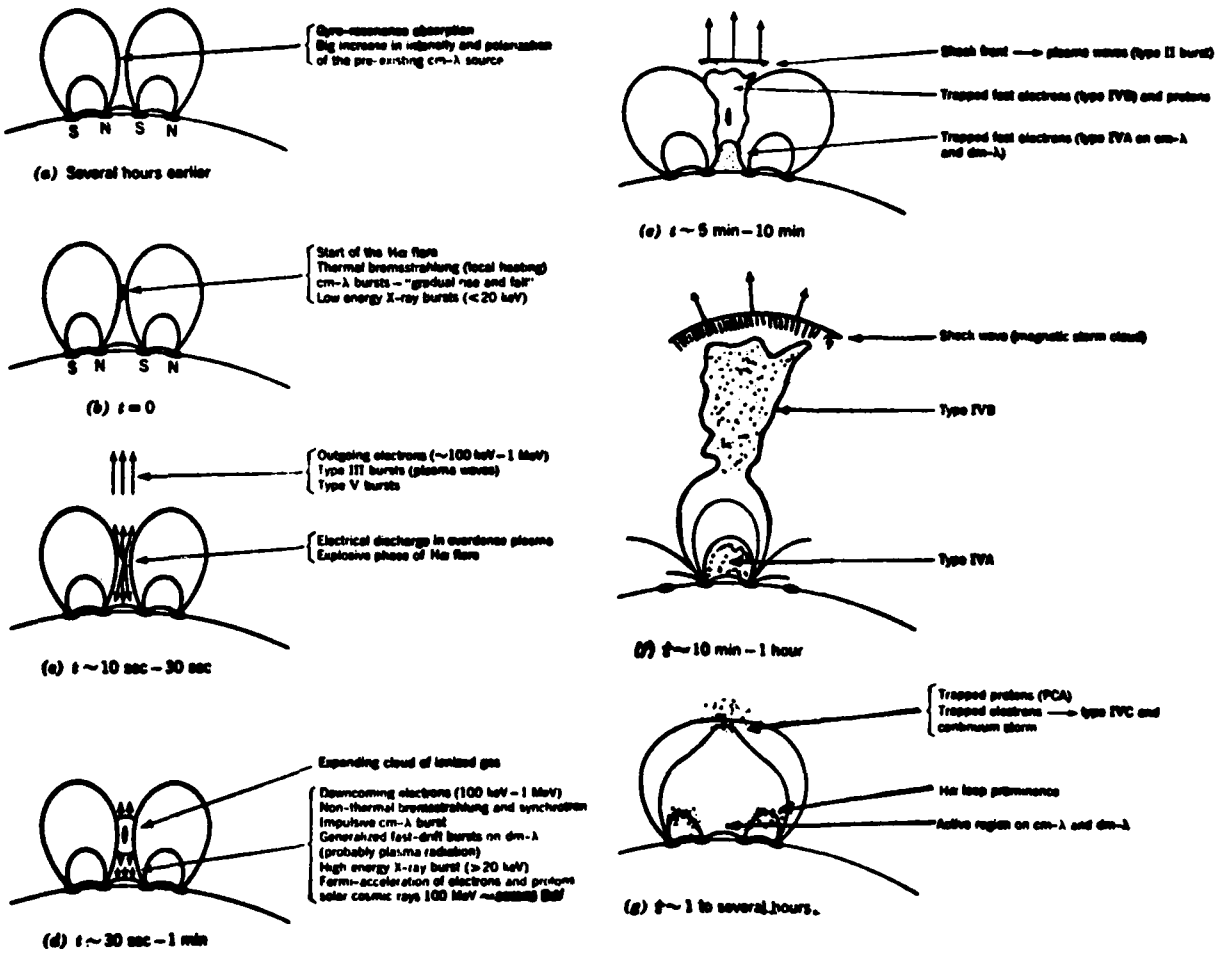



**Figure 8.16 Evolution of a Large Flare and Associated Emissions (from Kundu, 1965).**

(2) Gradual rise and fall bursts have slower rise and fall times and last ten minutes to several hours. Other bursts may be superimposed. They are thought to be due to preheating or compression in the magnetic field associated with active regions.

(3) Great bursts typically have impulsive beginnings, reach a maximum of 500 SFU or more, last tens of minutes to several hours, and are associated with major solar activity. Normally, they are very complex except at low frequencies.

(4) Post burst increases (PBI) follow other bursts (simple or complex) and last ten minutes to several hours. They are thought to be caused by bremsstrahlung radiation.

The so-called "Castelli U" (Castelli, et. al., 1967) signature on a frequency versus burst magnitude graph has been shown to identify very energetic events. This criteria, named for Dr. John Castelli, its originator, has certain minimum conditions. The frequency flux spectrum exhibits a distinct U shape with the minimum burst intensity occurring between 600 and

2000 MHz. The burst produces more than 1000 SFU near 9000 MHz, and rises to its maximum value on low frequencies (see Figure 8.18). The peak fluxes used to establish the U signature may (usually do) occur at different times during the event.

The two legs of the U are produced by different processes. The high frequency side of the U is formed by synchrotron radiation (and possibly some bremsstrahlung). The cause of this emission is dumping of high energy electrons which had been stored in the lower corona. The low frequency leg of the U is an extension of the type IV burst due to synchrotron radiation from relativistic electrons accelerated by the shock wave. The low point in the U is a crossover between dominant emission processes.

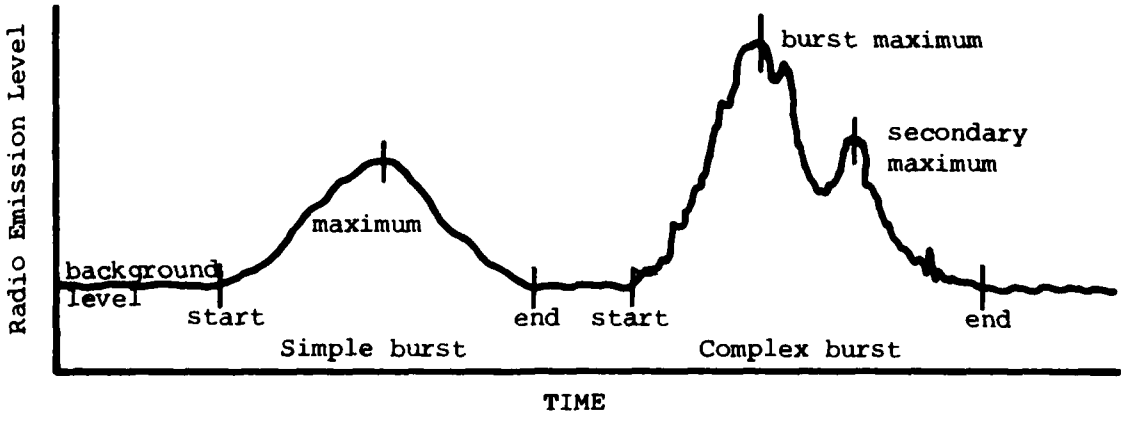



Figure 8.17 Classification of Discrete Frequency Bursts.

## 8.6 Solar Flares

A solar flare has been defined (Gibson, 1973) as "a highly concentrated, explosive release of energy within the solar atmosphere, followed by a gradual leveling off and decay of material motion and temperature". A more descriptive definition of a solar flare is "a sudden, short-lived brightening of a localized area in the solar chromosphere". We will look first at the systems for classifying solar flares, their life cycle, and their correlation to active region characteristics and production of energetic cosmic rays. Then, we will examine the cause of flare emission, and finally, we will correlate an energetic flare to the life cycle of a solar active region.

### 8.6.1 Optical Flare Classification

Solar flares are classified to approximate their total energy release. The total amount of energy released by a flare is the deciding factor in the severity of its effects on the near earth environment. Solar flares result in enhanced emission across the electromagnetic spectrum. Optical, x-ray, and radio frequency observations provide information on the flare at various portions of the spectrum. It is important to note that not all large flares produce high levels of emission at all wavelengths. Some flares concentrate most of their energy into just one portion of the EM spectrum.

The flare patrol maintained by solar observatories uses H-alpha, a chromospheric line. Flares are very obvious at this wavelength and are classified according to their size and intensity. The area which a flare covers at its time of maximum brightness is measured and corrected for solar curvature. This corrected area categorizes the flare in size or "importance" as follows:

| Importance | Corrected Area (in millionths of the solar disk) |
|---|---|
| 0 (subflare) | less than 200 |
| 1 | 200-499 |
| 2 | 500-1199 |
| 3 | 1200-2400 |
| 4 | greater than 2400 |

One optical flare intensity or "brilliance" classification is based on the Doppler shift of the hydrogen-alpha line. This Doppler shift is a measure of emitting gas particle velocity and is used by the observer in making his subjective estimate of flare intensity. Using this system we classify flares as follows:

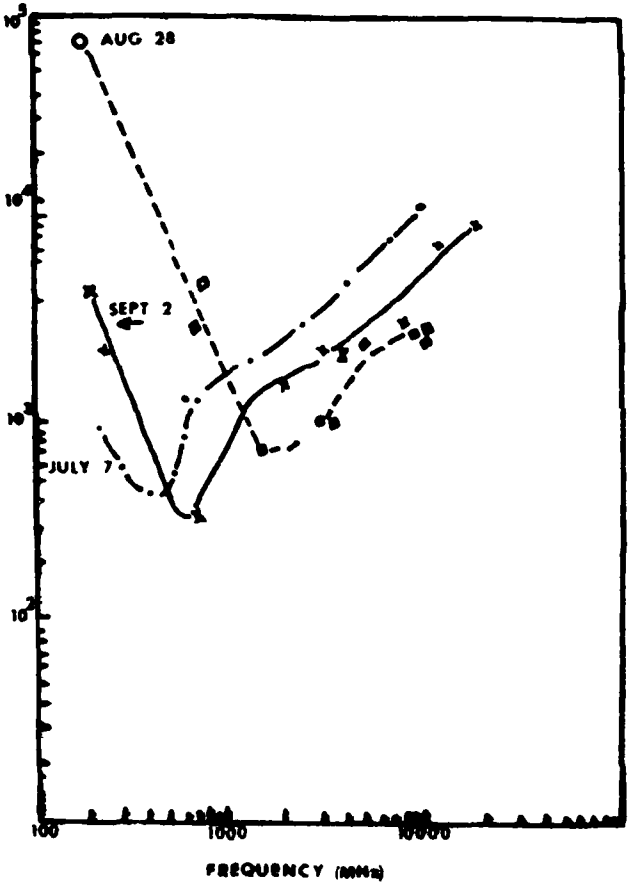



Figure 8.18 Castelli U Criteria (Castelli, et. al., 1967).

- U SHAPED FREQUENCY–FLUX SPECTRUM
- MAXIMUM FLUX INCREASES ON ALL FREQUENCIES NEED NOT BE SIMULTAEOUS BUT GENERALLY OCCUR WITHIN A FEW MINUTES OF EACH OTHER (AND ASSOCIATED OPTICAL FLARE)
- MINIMUM FLUX INCREASE OCCURS BETWEEN 600 AND 2000 MHz
- FLUX GREATER THAN 1000 s.f.u. NEAR 9000 MHz
- PEAK FLUX RISE IN LOW FREQUENCY DIRECTION

| Intensity | Doppler Shift of Flare Emission |
|---|---|
| Faint (F) | Seen over a line width of 0.8Å or greater. |
| Normal (N) | Seen over a line width of 1.2Å or greater. |
| Brilliant (B) | Seen at + and/or - 1.0Å off line center. |

The SOON (Solar Observing Optical Network) telescopes are capable of directly measuring the intensity of optical flare emissions. The SOON observatories report as their flare brightness the measured flare intensity. However, the observed intensity is strongly dependent on the seeing conditions, and only a slight amount of atmospheric pollution can drastically alter the measured intensity. The operator sets thresholds for each flare intensity criteria, but variations in seeing at various SOON sites can result in the same flare being assigned a different brilliance by different observatories. This is important in determining if the emission is a "faint" flare or a plage brightening. Moreover, flare intensity is based on the brighest element of a flare. Hence, a flare is encoded as "brilliant" if any segment reaches this level, even though most of the flare may be "faint."

Each optical flare is assigned an importance and an intensity classification. These values allow a forecaster to roughly estimate the total energy output of each flare. This determination is rough because a "large" 2B flare may give off more energy than a "small" 3N, and because the total energy output is also a function of output at other frequencies and the duration of the flare. Optical flare observation provides a rough classification of the energy (in addition to determining flare location).

One type of large flare, the Hyder flare, is generally non-energetic. Usually large (importance 2 - 3) and faint, these flares often occur following the eruption of a large prominence or distant flare. They probably result from coronal material falling into an old plage area and producing a large brightening. Consequently, they may occur in spotless regions. An energetic Hyder flare may result from the disruption or activation of a quiescent prominence.

### 8.6.2 X-ray Flare Classification

While visible flare emission increases by, at most, a few percent, the x-ray emission may be enhanced by a much as four orders of magnitude. Since x-ray sensors are located on satellites above the earth's atmosphere, atmospheric attenuation is not a problem. The sensors record energy output in a certain range of x-ray wavelengths. The sensors currently monitoring the sun are part of the GOES satellite system. Measurements are made in the one to eight Angstrom (soft) band and in the one-half to four Angstrom (hard)band. The satellites are located in geostationary orbits (22,000 miles altitude in the earth's rotational plane). They measure total solar output in their wavelength bands as long as the sun is visible from the satellite. Only during the equinox period does the earth eclipse a geostationary satellite. Two satellites are monitored and, provided they are sufficiently spread in longitude, only one can be eclipsed at a time. SESS receives continuous, real-time x-ray patrol from the GOES satellites.

X-ray flares are classified according to the peak energy flux of the flare. For "soft" x-rays, the classification system is:

| Classification | Minimum X-Ray flux (ergs/cm$^2$sec) |
|---|---|
| C | $10^{-3}$ |
| M | $10^{-2}$ |
| X | $10^{-1}$ |

The above categories are subdivided into nine sections according to the first digit of the actual peak flux. That is, a peak flux of 5.7 X $10^{-2}$ erg/cm$^2$ sec is an M5, soft x-ray flare. In the absence of soft x-ray data, "hard" x-rays are sometimes used:

| Classification | Minimum X-ray flux (ergs/cm$^2$ sec) |
|---|---|
| C | $10^{-4}$ |
| M | $10^{-3}$ |
| X | $10^{-2}$ |

As with soft x-rays, the first digit of the actual flux is used to divide each category into sections. Those in X-class are called major x-ray flares, and those in M-class are called minor x-ray flares. These names relate to the severity of the near-earth effects produced.

### 8.6.3 Flare Production

The life of a typical flare begins before the optical enhancement. A few minutes or tens of minutes prior to the onset of a solar flare a slight increase in soft x-ray emission occurs. The actual flare begins with an increase in the optical and x-ray emission of the flaring material by at least 50% above background in two minutes or less. This is the beginning of the "flash" phase which lasts until the peak intensity of the flare is reached. The area of the flaring region expands throughout the flash phase, and reaches maximum shortly after the flare intensity peaks. The optically brightened regions occur close to the region's magnetic inversion line. Energetic flares commonly have intense flaring regions on either side of the inversion line, and show as "parallel ribbons". These ribbons mark the photospheric foot-prints of the flare loops which lie perpendicular to the inversion line. Radio bursts begin shortly after the optical flare, and attain maximum intensity after it does. Bursts are seen first on higher frequencies and later on successively lower ones. On each frequency, the onset is very rapid. On each frequency, the flare begins a gradual decline as it passes maximum and enters its decay phase. The mean duration of an optical flare is loosely related to the flare's magnitude:

| Importance | Average Duration | Percent of All Flares |
|---|---|---|
| 0 | 17 minutes | 75 |
| 1 | 32 minutes | 19 |
| 2 | 69 minutes | 5 |
| 3 or 4 | more than 2 hours | less than 1 |

Similar durations occur in soft x-ray flares, while hard x-ray flares tend to be more impulsive and have shorter durations. Certain active region characteristics relate to higher probability of major flare occurrence. They are:

(1) A large, complex sunspot group; typically an E or F group; a $\beta\gamma\delta$ or (or even better yet a $\gamma\delta$ ) configuration; a kinked, complex inversion line with strong magnetic gradients;

(2) Modified Zurich classification of Fri, Fsi, Eai, Ehc, Dkc, Eki, Ekc, Fki, or Fkc;

(3) Reversed polarity;

(4) A second rotation (mature) sunspot group; and

(5) A history of other (including minor) flare activity.

When a very energetic flare occurs, certain characteristics are typically present. They are:

(1) Large, brilliant optical flare (2B or greater);

(2) Parallel ribbon flare;

(3) Loop prominence system, surges, and sprays;

(4) At least 20% of the umbra of a major sunspot covered by the flare emission region (possibly indicating flaring material is of sufficient energy or is producing sufficient breakup of region magnetic fields to encroach on the strong sunspot fields);

(5) A white light flare (implying strong emission over a wide range of frequencies);

(6) Large, long-lived soft x-ray flare (typically greater than X1 with a rise time in excess of 5 minutes);

(7) Intense hard x-ray flare (typically greater than X1);

(8) Large, long-lived microwave radio bursts (at least 1000 SFU with a rise time of over 5 minutes, and radio frequency maximum occurs after optical maximum);

(9) Castelli U spectrum of microwave bursts; and

(10) Type II and IV swept frequency bursts.

Typically, a very energetic flare is a slow-rising, long-lived flare which has large emissions across the electromagnetic spectrum.

8.6.4 A Typical Flare-Producing Region

The solar flare which produced an energetic proton event on 29 May 79 fits the "classic" proton flare picture:

On 10 January, a B group was first reported in McMath region 15097 (spots located at N19 E34). The spot group grew to a D-beta configuration before rotating off the disk, and produced numerous sub-faint and sub-normal flares. The region returned on 2 February as McMath region 15135, which grew to be an E-beta group. It produced one 1-brilliant and two 1-normal flares while on the disk. A region poleward of it at N27 (McMath region 15134) rotated on as a B-beta group and grew to be an E-delta group. This region produced numerous sub-normal flares and a 1-brilliant flare before it rotated off the west limb. In February, McMath region 15172 rotated onto the disk. At central meridian, this massive plage region covered an area of 8000 millionths (nearly 1%) of the solar disk. It contained as many as three separate spot groups during its transit of the disk (including those which had been in regions 15134 and 15135). The first spot group, at N19, was the third rotation of region 15097. This region began as an E group, but, by 10 March, when at W35 had declined to a C group. This group produced three 1-normal flares, one on 2 March and two on 4 March, and it continued to produce subflares as it rotated across and off the disk. The second spot group, at N27, was the second rotation of region 15134. It had between one and six spots through its transit, and was classified as a C or D group during most of the time. This spot group produced two normal flares and a 1-brilliant flare, on the 4th, 6th, and 13th of March. The third spot group, at N13, was initially a D group, but declined to a B group by 10 March (at W48). This group produced a 1-faint flare, its only significant activity, on 8 March. Region 15172 returned to the disk on 28 March as McMath region 15214. It contained a single B-beta spot group at N21, apparently the fourth rotation of region 15097 (15135, 15172). The plage areas varied between 300 and 3500 millionths of the disk. On 6 April, the region began producing subflares (at W38), while earlier no activity had been observed. This group produced a 1-brilliant flare which lasted nearly 50 minutes (on 8th). This flare produced a major radio burst on 606 MHz (14,000 SFU), plus significant bursts at 1415 MHz (6800 SFU) and 2695 MHz (750 SFU). The region also produced at least two other importance 1 flares. These flares indicate a possible reintensification of the previously dying region. On 23 April, an east limb flare with a loop prominence system occurred near NE25. The x-rays reached M7, and radio bursts stronger than 4000 SFU were reported on frequencies between 2500 MHz and 5000

MHz. This heralded the return of the spot system in McMath region 15266, even though the spots were still a day behind the limb. Type III and possible type II bursts accompanied the flare. Region 15266 covered more than 12,000 millionths of the disk with its plage (over 1%). As the region rotated onto the disk, this fifth rotation of region 15097 contained an E-delta spot group. On 25 April, it produced a 1-normal and two 1-brilliant flares during a four hour time period. The last of these flares was accompanied by a radio burst on 2695 MHz of nearly 8700 SFU. On the 26th it produced a 1-brilliant flare and a 1-normal flare six hours later. On the 27th, the region produced a 1-normal flare. On the 28th, the region was reported as an E-beta gamma delta spot group, containing 23 spots which covered 1000 millionths of the disk (0.1%). At 1304Z, a flare began which reached a maximum of 3-brilliant at 1335Z, covering an area of 1600 millionths. It was a parallel ribbon flare, with a loop prominence system, which lasted over nine hours. It was accompanied by an X4 x-ray burst (start 1308Z, peak 1337Z) and the following radio bursts:

| Frequncy | Start | Maximum | Intensity | Duration |
|---|---|---|---|---|
| 35000 MHz | 1316.8 | 1329.1 | 1856 | 84.6 Min |
| 15400 | 1315.0 | 1329.5 | 4580 | 85.8 |
| 10500 | 1309.7 | | 6561 | |
| 9500 | 1312.5 | 1328.4 | 5560 | |
| 9400 | 1310.5 | 1328.6 | 6461 | 73.8 |
| 8800 | 1310.6 | 1328.7 | 7530 | 92.9 |
| 4995 | 1306.7 | 1332.4 | 4080 | 103.5 |
| 2800 | | | | |
| 2695 | 1308.5 | 1324.3 | 3130 | 99.2 |
| 1415 | 1313.3 | 1332.0 | 23,800 | 95.7 |
| 930 | 1314 | 1330.6 | 26,978 | 66 |
| 606 | 1314.9 | 1332 | 23,800 | 89.1 |
| **Frequency** | **Start** | **Maximum** | **Intensity** | **Duration** |
| 410 | 1317.1 | 1334 | 20,200 | 93.4 |
| 245 | 1317.1 | 1322.4 | 142,000 | 92.9 |
| 237 | 1321.1 | 1323.2 | 143,600 | |

Secondary Bursts

| Frequency | Time | Intensity |
|---|---|---|
| 4995 MHz | 1353.8 | 1800 |
| 2800 | 1354.5 | 24,000 |
| 2695 | 1354.5 | 22,100 |
| 1415 | 1342.2 | 5500 |
| 930 | 1404.3 | 5330 |
| 606 | 1320.1 | 20,500 |

Certain features of this region's evolution are particularly significant:

This region was very large and complex and went through at least one reintensification. The spots which began in region 15097 grew, peaked, and were dying on their third transit, as is normal for a large sunspot group. This region had not been an energetic flare producer. The B-beta group seen in region 15214 apparently was the beginning of new flux emerging in the old region, a potentially explosive situation. The flares and radio bursts which began on 6 April showed the new spot group to be unstable. We have no record of an arch filament system in the region, but it is highly probable one existed at that time. If one did exist, it would have been further proof of the emerging flux.

The spot group was an E-delta when the major flare occurred. The group continued to grow and became an F-delta, comprised of 90 sunspots and covering 1150 millionths of the disk. It produced two other major flares before it reached its greatest size. It is typical of most regions that the most energetic flares occur prior to the region reaching its greatest size, or during a phase of rapid decay.

The flare was large and intense at optical wavelengths. Parallel ribbons, a loop prominence system, and umbral coverage were all reported. The optical flare had a slow rise (31 minutes) and a long duration (over 9 hours). The x-ray burst was long and intense. The soft x-rays took 29 minutes to climb to maximum intensity of X4, and the hard x-rays exceeded X6. The radio bursts took approximately 15 minutes to climb to their peak intensities on all frequencies. A Castelli-U occurred, as did a type II and type IV (plus type III) swept frequency radio bursts.

## 8.7 Flare Origin and Emission

Various models of solar flares have been proposed, but none explains all flare observations. We will now look at one solar flare model which explains many of the major features.

Solar flare energy is stored energy which is released in a short period of time. The energy is probably stored by the twisted and kinked magnetic field lines of the active region. A process called magnetic reconnection has been proposed as the release mechanism for the stored energy. The amount of energy released by an average flare is about $10^{30}$ ergs over 100 to 1000 seconds. This equates to only a 5% decrease in stored magnetic energy, so the magnetic field could be a storage device for flare released energy. In magnetic reconnection, a stretched, twisted magnetic field line instantaneously "breaks" and "reconnects" to shorten itself. The shorter distance means less stored energy, and the released energy is available to the flare. The reconnection occurs along the inversion line, so the energy release (and the flare) occurs along it. Part of this energy may be imparted to the ambient plasma to accelerate it to escape velocity.

A solar flare releases energy across the electromagnetic spectrum. Each flare has its own distribution of energy by frequency; that is, if flare A is twice as intense in x-rays as is flare B, flare A will not necessarily be twice as intense as flare B at optical or radio frequencies. Flares which release a greater portion of their energy at higher frequencies, "hard" spectrum flares, cause a greater disruption of the near-earth environment.

On any given frequency, the typical flare shows a near-exponential rise in intensity during the flash phase and a slow decay to pre-flare level. This pattern occurs again and again in solar emission. An intense solar flare gives a significant portion of its energy to solar plasma particles, primarily electrons, protons, and a few alpha particles. These particles are accelerated to high velocities. Electrons are easily accelerated to near light speed, and protons can be accelerated to near-relativistic speeds by an intense flare. Some accelerated particles may immediately exit the lower solar atmosphere. Those which move downward cause a white light (photospheric) flare, while those which move upward may escape into the corona and solar wind. Other accelerated particles will be trapped by the lower coronal and chromospheric magnetic fields, where they slowly diffuse across the magnetic field to escape or give up their energy to the coronal gas (and produce radio frequency emissions).

The particles which escape the sun show a sudden jump from background to peak flux (the particles which get out immediately) followed by a slower decline to background (the particles which diffused out of the trapping field). At the earth, we usually see harder (higher energy) particles first. The higher energy particles have larger kinetic energies than their lower energy counterparts. This means a high energy proton moves much faster than does a lower energy one. Since the fast protons outdistance the slower ones, they arrive at the earth first.

## 8.8 Summary

From a distance many stars, including our sun, appear stable. Yet, closer inspection reveals that most, if not all stars are variable to some degree. Of particular interest are the rapid, seemingly random, variations in emission. It is these which generate serious terrestrial effects. We are only now beginning to understand what questions to ask and where to look for

the answers. We have seen the level of activity, indexed by the ubiquitous sunspot number, varies with a semi-constant period. The more complicated the structure of a given sunspot group, the more likely it is to be the site of sudden, explosive distruptions of military and civilian systems. Both electromagnetic radiation and emitted particle fluxes are of concern, but for different reasons. To reach us, both must travel through the interplanetary medium.

# CHAPTER 9

# THE INTERPLANETARY MEDIUM

Our understanding of the structure and dynamics of the interplanetary medium has developed very recently. Until the decade of the 1950's, researchers thought of interplanetary space as a vacuum only occasionally traversed by stellar material. The development of space probes has changed our concept considerably. We now know that the interplanetary region consists of solar material traveling at high velocities and interacting to produce shock waves and discontinuities. In order to understand the structure and dynamics of interplanetary space, we need to consider the origins and dynamics of the solar wind and the interplanetary magnetic field (IMF). These two elements provide a vital link in the solar-terrestrial relationship.

## 9.1 The Solar Wind

Evidence of "something" moving between the sun and earth existed for many decades before the discovery of the solar wind. In the 1930's, Chapman and Ferraro noted that aurora and fluctuating magnetic fields occurred on earth following intense solar activity. These observations led to the theory of intermittent plasma streams flowing from the sun to earth during periods of strong solar activity. During the 50's, an astronomer named Bierman, while studying comet tails, speculated that the tails always pointed away from the sun due to a steady stream of solar material moving out through space. This was the first suggestion of a continual solar wind. In 1958, Eugene Parker worked out the theoretical details and concluded that the solar corona is continually expanding outward from the sun. This outward flow of matter is called the solar wind. Parker's predicted velocities of several hundred kilometers per second have since been confirmed by spacecraft observations.

### 9.1.1 Solar Wind Plasma

About 99% of the universe exists in the plasma state. The solar wind, like other stellar winds, is a plasma. Its average speed of 400 km/sec is roughly 10 times the speed of sound. The source of these particles is atomic hydrogen ionized in the lower regions of the solar corona. Present models assume that solar wind particles originate in the lower corona on open field lines. Similar particles which exist on closed field lines are trapped in the lower corona and do not escape to become part of the solar wind. As a result, not all coronal regions contribute equally to the solar wind plasma.

Certain areas of the photosphere contain mostly open field lines, while others contain primarily closed ones. Solar active regions occur where the photosphere is penetrated by a strong, mixed-polarity magnetic field. The field lines in this region are mostly closed, or connect on both ends into the photosphere. Unipolar areas of the sun are the primary sources of open field lines. Thus, unipolar regions of the photosphere generally underlie regions of open field lines along which charged particles may readily escape. In the corona, such regions are called coronal holes. On x-ray and ultraviolet (coronal) photographs, coronal holes appear as vast dark areas, or holes in the coronal emission. The discovery of coronal holes has helped explain many of the dynamic features of the solar wind. The interaction of higher velocity plasma streams with neighboring slower speed wind streams can produce many of

the discontinuities and shocks which propagate outward to the earth. As these move outward, they modify the density, temperature, and magnetic field of the ambient medium.

Satellites in the near-earth orbit have measured several properties of the solar wind. These include:

(a) The wind blows continually with velocities between 300 and 700 km/sec;

(b) The wind is "gusty" over a period of minutes and displays cyclic variations with seasonal and solar cycle periods;

(c) Particle density varies between 3 and 40 $cm^{-3}$; the most typical value being 8 $cm^{-3}$;

(d) Temperatures range between 10,000°K and 100,000°K. (The plasma in a fluorescent tube is over 20,000°K.)

Properties of the solar wind have been observed well beyond earth's orbit by Pioneer and Voyager spacecraft. Beyond earth's orbit, solar wind speeds remain relatively constant. The flow, however, becomes more turbulent probably due to substantial wave generation and the resultant formation of shock fronts. How far the solar plasma penetrates into the interstellar medium has not yet been determined. It is theorized that the interstellar gas pressure (other stars have plasma winds also) eventually balances the dynamic pressure of the solar wind to form a boundary known as the heliopause. This spherical and static boundary is at least 100 AU from the sun.

### 9.1.2 Influence of the Magnetic Field

In the presence of a magnetic field, plasma particles spiral around each field line (lines of magnetic force). The extremely high conductivity of the plasma constrains the particles to a given field line. The particle is free to stream along a field line, but cannot cross it. This property, known as the frozen-in condition, has several important consequences. First, there can be no slippage of plasma particles across magnetic field lines. Both the plasma and the field will move together. Secondly, two magnetized plasmas, when brought together, cannot mix except under special circumstances (when the field lines "merge" together). These two properties are fundamental to understanding the ongoing processes within the interplanetary medium.

Plasma exerts pressure (recall that plasma is a gas) which is so small that it is measured as energy density, i.e. the amount of particle energy per unit volume. Magnetic fields also exert pressure (energy density) in the tension of the field lines (the repulsive force between two like magnetic poles when pushed together is an example). If the plasma energy density is greater than the magnetic field energy density, the movement of the plasma will force the field to move. Conversely, if the magnetic field energy density is greater, it will control the plasma movement.

Within the solar atmosphere, regions exist where the plasma exerts control and others where the magnetic field dominates. Close to the solar surface, where the horizontal magnetic fields are strongest, the plasma energy density

is less than the magnetic field energy density. Here, the field can direct and retard the plasma flow. Trapping of the plasma by the field can exist out to almost one solar radius. Flare loops are visible evidence of the process. Unable to diffuse outward, the condensed solar plasma flows downward along the field lines. Moving further out in the corona, the magnetic field energy density decreases faster than that of the plasma. Near 3 solar radii, the plasma energy density exceeds that of the magnetic field, and it carries the field outward into the interplanetary medium, as it escapes, forming the interplanetary magnetic field (IMF). The transition region (.6 to 3 solar radii) between field control of the plasma and plasma control of the field can be considered as the effective source region for the solar wind (Figure 9.1). One visible structure which spans this area is the helmet streamer. Near its base, the closed field lines confine the plasma. At greater distances, where the plasma pressure exceeds the magnetic pressure, the plasma carries the field outward as it escapes. This forms the "open" field lines. The plasma continues to drag the field lines outward through space, possibly to the heliopause.

## 9.2 The Interplanetary Magnetic Field (IMF)

### 9.2.1 Formation of the IMF

As the solar wind particles move radially outward, dragging the field with them, the sun continues to rotate. One end of the field line remains embedded in the photosphere. The solar wind carries the other end outward. In this way, field lines are bent into Archimedian sprials. The effect is similar to a rotating garden sprinkler. The jet of the water follows a spiral path, although the trajectories of individual drops are radial. The magnetic field lines connect all plasma originating in the same location of the lower corona. This process is sometimes referred to as the garden hose effect. The spiral shape of the field lines is related to the outward velocity of the plasma. Variations in the solar wind speed produce different amounts of "bend" (different pitch angles) in the field spirals as viewed from above the ecliptic plane.

For a uniform flow, spirals would be symmetric about the sun (Figure 9.2). Low wind speeds allow the spiral to become tightly wound, while faster wind streams (such as that leaving coronal hole regions) will produce a more open spiral shape. Since the wind speed varies with time and location, interaction regions are generated as fast streams, with "loose" spirals, overtake and compress a slower stream. Fast streams moving ahead of slower streams generate areas of rarefraction. Compressions in the medium generate stronger magnetic fields (field lines closer together) and increased densities of the solar material (recall that two plasmas, in this case a slow and a fast one, don't mix). As the field lines are shoved together, the particles must also move closer. If conditions are right, shock waves can be generated. Near the earth, the spiral angle is approximately $45^{o}$ (garden hose angle) to the solar west of the direct earth-sun line. It takes about four days for a portion of a field line to be carried from the solar surface to the vicinity of earth (93 million miles at about 400 km/sec). Variations in these values are common, however, due to the gusty nature of the solar wind.

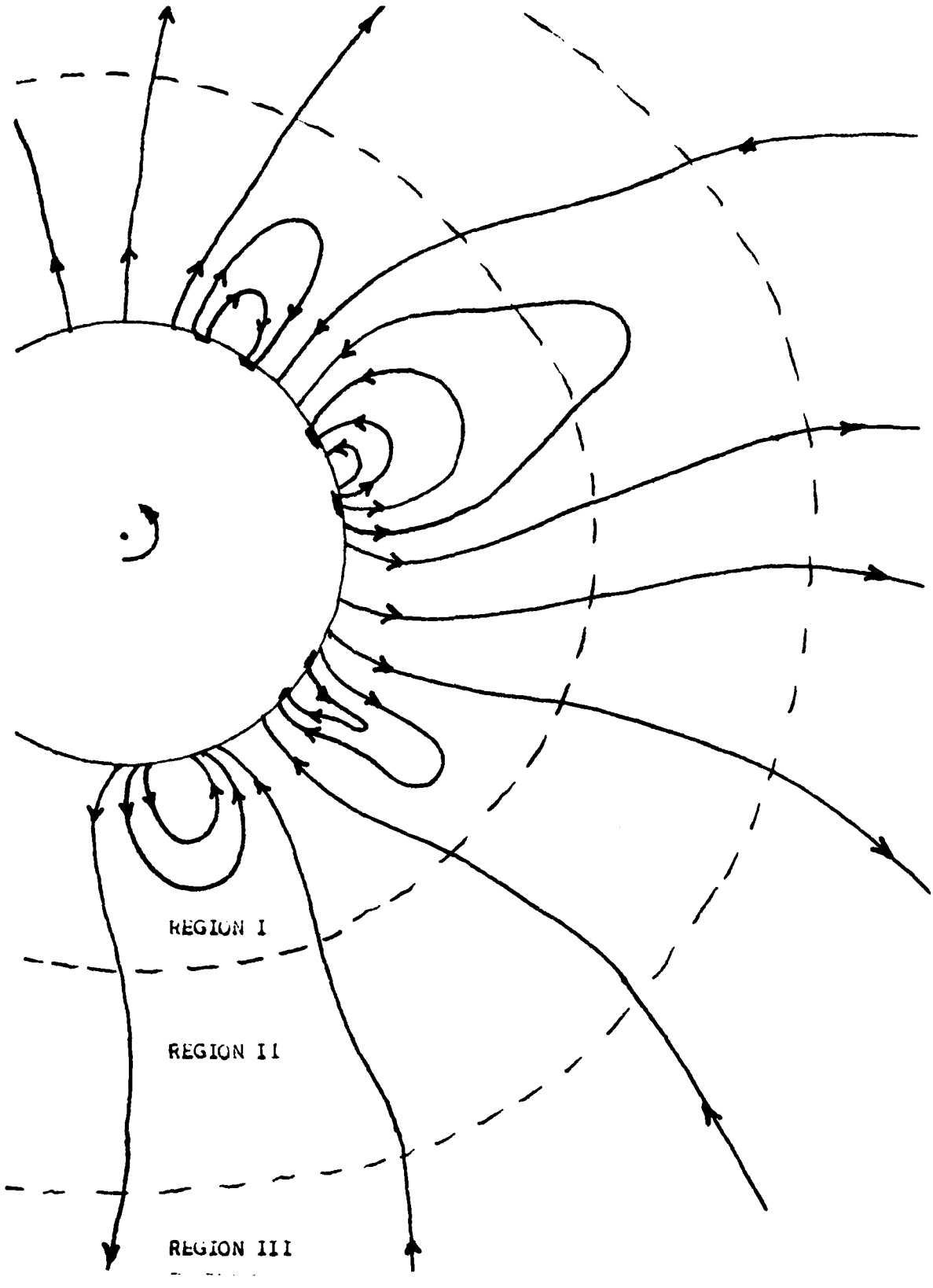



Figure 9.1 Magnetic field line orientation near the sun as viewed from above the ecliptic plane. Field controls plasma in Region I. Plasma controls field, stretching it out in Region III. Region II is the effective source of the solar wind.

### 9.2.2 Sector Structure

The dipole nature of the background solar magnetic field adds latitudinal structure to the interplanetary medium. As the dipole field lines are drawn out in space, they form a magnetically neutral sheet as viewed in the meridional plane. This neutral sheet, also called a current sheet (a current of nearly a billion amperes flows in this region), exists in the region where the field lines reverse direction (polarity). This current sheet can be thought of as an extension of the heliomagnetic equator into space. It is normally located close to the ecliptic plane (Figure 9.3). During the past solar cycle (cycle 21), the field was directed towards the sun (negative polarity) when earth was below (south) the current sheet. Above or north of the current sheet, the field was organized away from the sun (positive polarity). Regions of space where the field is predominantly towards or away from the sun have been termed sectors.

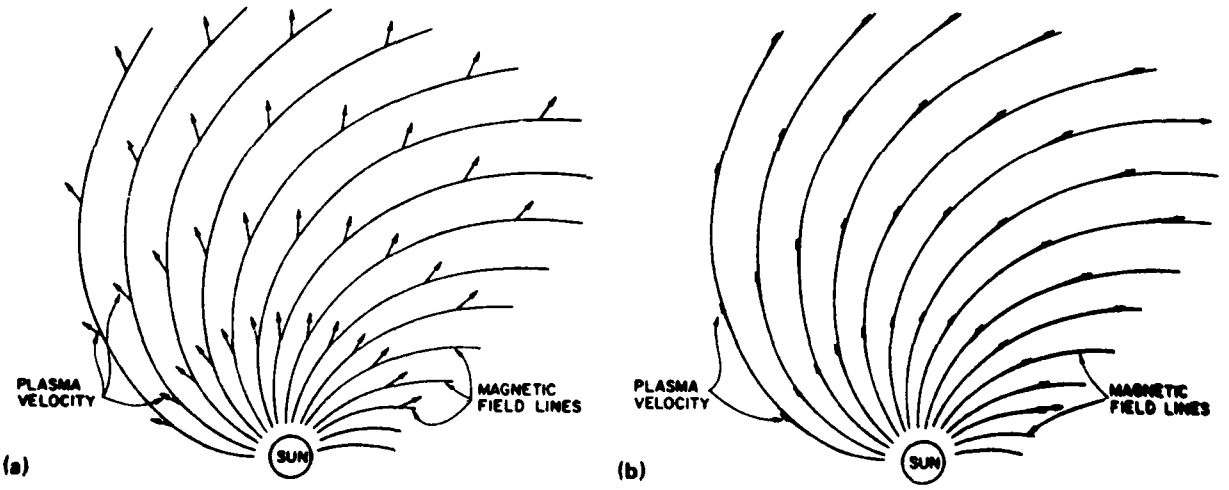



Figure 9.2 Archimedian spiral structure of the IMF in stationary reference plane (a) and rotating with the sun (b) showing plasma flow and magnetic field direction (from Hundhausen, 1972).

Two sectors are readily indentified, one above and the other below the current sheet. The current sheet is apparently flat near the sun and inclined to the solar rotational equator. The sun's rotation causes the current sheet to wobble, producing a wavy structure similar to that shown in Figure 9.4. The Parker spiral structure prevails in both sectors, and the wavy structure is frozen into the solar wind out to at least 8 AU (Thomas and Smith, 1981) for heliographic latitudes within $10^{o}$ of the current sheet. Pioneer 11 observations suggest that these waves do not extend much beyond $16^{o}$ heliographic latitude above or below the current sheet. Finally, the current sheet's inclination to the solar equatorial plane ranges from a minimum of $15^{o}$ near solar minimum to as much as $30^{o}$ near the time of solar magnetic field reversal. The latitudinal extent of the sheet warpage is least near solar minimum.

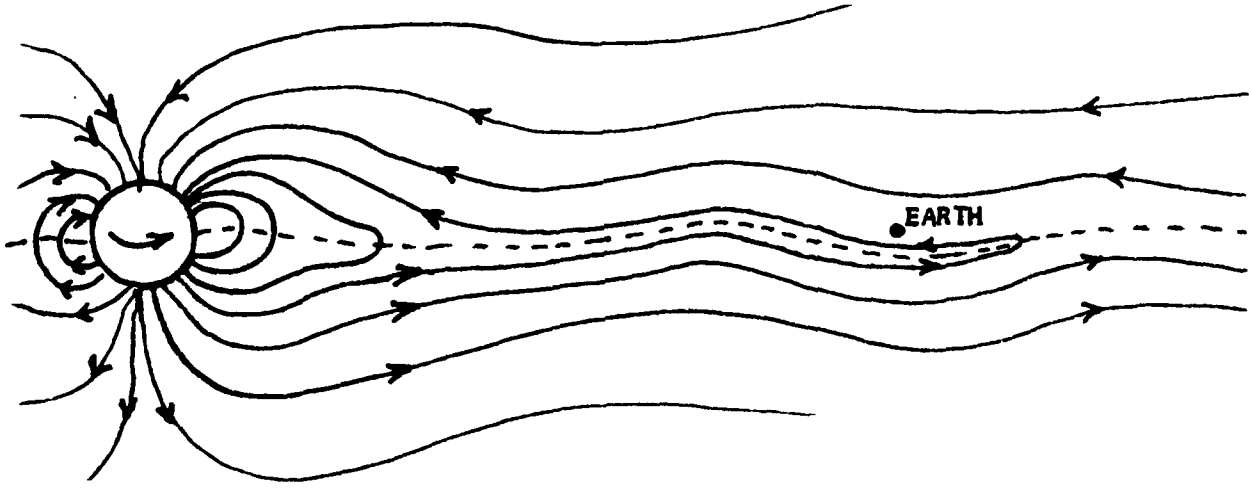



Figure 9.3 Formation of a current sheet (meridional view). Current flows out of the plane of the paper along the dashed line. The earth is pictured above the current sheet.

Solar wind speeds are found to be a minimum in the immediate vicinity of the sheet, and increase with latitude above and below the sheet. This seems to confirm the low latitude origin of the sheet (since solar wind speeds are known to increase with heliographic latitude at about 15 km/sec/deg) as does the 26.1 ($\pm$ 0.2) day rotation rate of sheet structures. The corrugations in the current sheet seem to have a wavelength of about 0.1 AU (9 million miles), and the sheet averages $3 \times 10^{4}$ km in thickness. Passage through the sheet is relatively rapid.

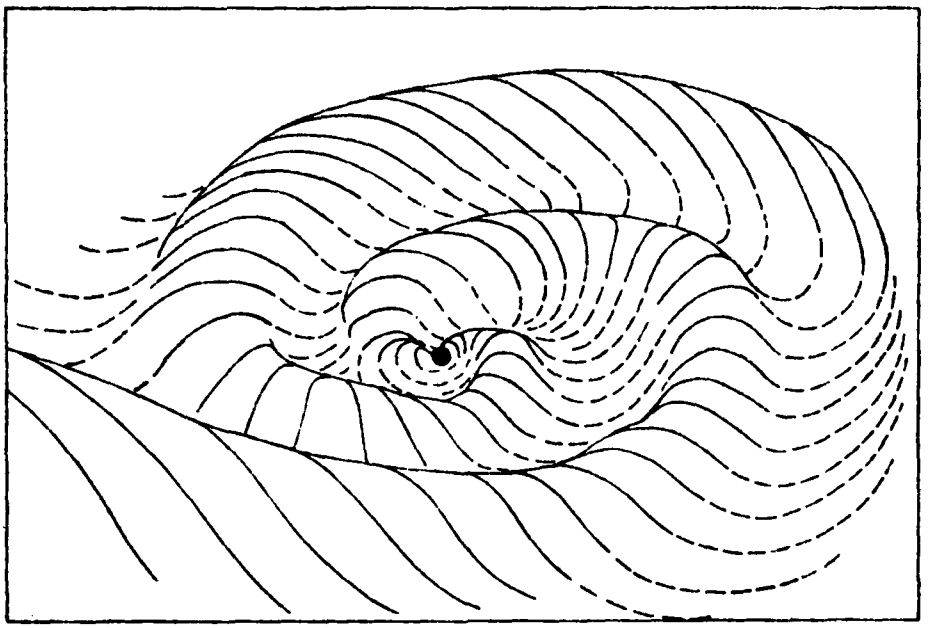
The inclination of the current sheet to the solar equator (itself inclined to the ecliptic) almost ensures that the earth will cross the current sheet at least twice during each solar rotation. These crossings are termed sector boundaries and may produce geomagnetic disturbances on the earth. This means a two sector structure should be common during at least a portion of each solar activity cycle, and it is. Solar activity (particularly near solar maximum) can produce low latitude variations in the current sheet structure and the consequent appearance of more than two sectors.

The current sheet structure is also modified by corotating interaction regions (CIRs). CIR is the term applied to the interaction between slow and fast solar wind streams and may be flare associated. Abnormally high plasma densities and IMF strengths are usually associated with CIRs because of the localized compression they produce. CIRs may be continuous or transient phenomena. The associated magnetic field and high density plasma ridge may force the current sheet to higher than normal heliographic latitudes in quiet regions. As the CIR propagates outward from the sun (Figure 9.5), it can deflect or actually absorb the current sheet field at its leading edge. The CIR structure thus modifies the current sheet by producing a high density wave in which the current sheet field is locally perpendicular to the mean current sheet plane. This may result in a sector boundary crossing causing a nearly perpendicular IMF component combined with a sudden jump in wind speed and density. Exiting the backside of the CIR produces a reverse shock. Note that the CIR cannot penetrate the current sheet in an organized manner, though it may eventually diffuse through and become indistinguishable as a well-defined shock. This also accounts for the shallow irregularities observed in the current sheet.

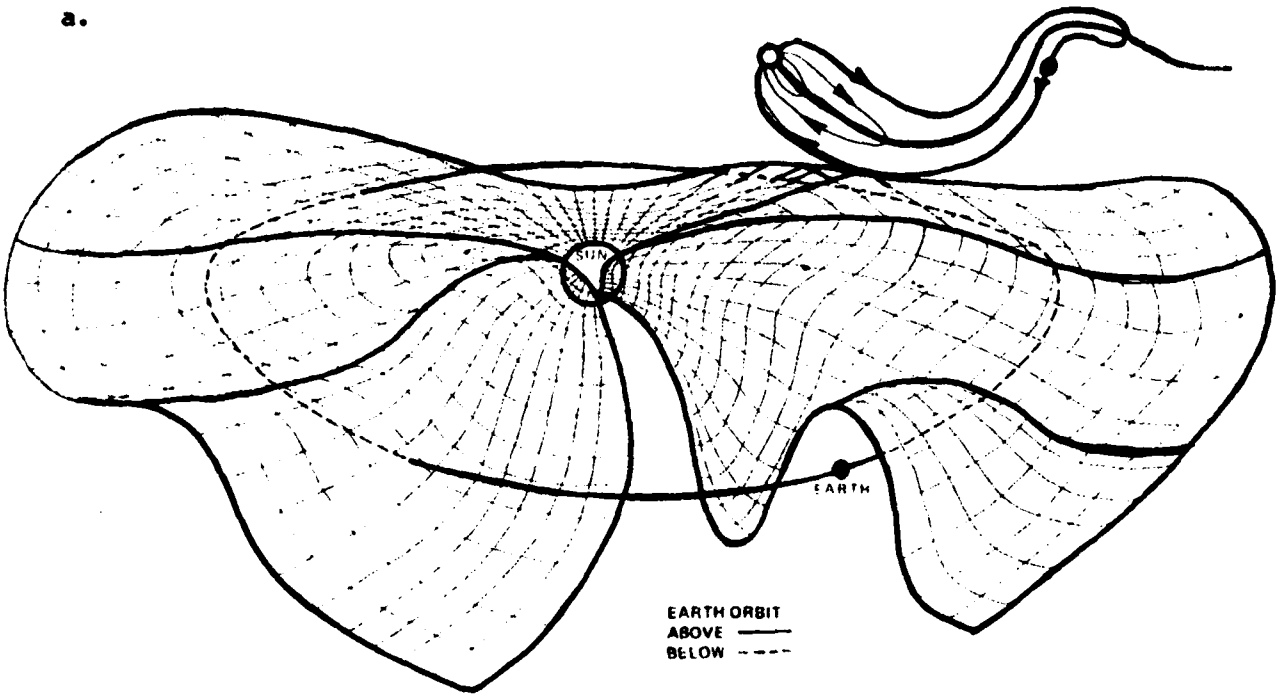
CIRs, or the irregularities which they create, probably play an important role in modifying the geophysical effects of solar flares. They may block or redirect flare plasma. Successive flares from a given region may produce a number of CIRs in the IMF. These may, in turn, produce a series of short-lived geomagnetic disturbances. By briefly trapping higher (MeV) energy protons, they may also account for corotational proton enhancements.

## 9.3 Flare Modification of the IP Medium

In addition to the continual release of low energy solar wind (.25 eV for electrons, 500 eV for protons) plasma, the sun may release high energy particles during intense solar flares (MeV electrons and protons). Although they are individually high in energy (as much as 500 MeV), their numbers are generally much lower. Their energy density is insufficient to significantly alter the spiral IMF established by the quiescent solar wind. Consequently, their outward flow is guided by the IMF. Traveling faster than the ambient solar wind plasma, these high energy particles (protons, electrons, and ions) stream along the field lines arriving at earth along an unmodified field direction. The time required to reach earth depends on the flare location, particle energy (velocity), and the shape of the field line emanating from the flare region. Detection at the earth also appears to be dependent on the earth and the flare plasma being on the same side of the current sheet when the plasma arrives at 1 AU.

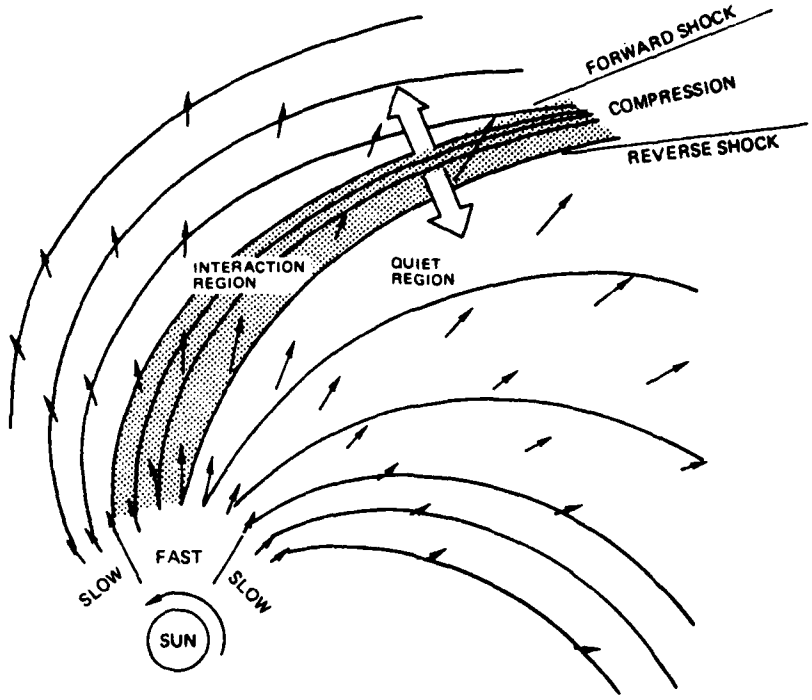


a.

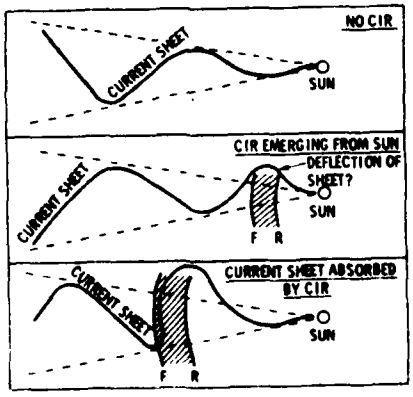



b.

**Figure 9.4 (a) Wavy Structure of the Heliomagnetic Current Sheet (after Bruzek and Durrant, 1977) and (b) combined meridional and obligue views (from National Research Council, 1981).**




(a)




(b)

**Figure 9.5 Interaction of a CIR with the Current Sheet (a) from above the current sheet (from Hundhausen, 1972) and (b) a meridional view (after Thomas and Smith, 1981).**

Although western hemisphere flares usually produce the largest energetic proton flux at earth due to their optimum geometry, the IMF spiral varies considerably and may allow protons released east of central meridian to arrive at earth, especially if the field were loosely wound due to higher than normal wind speeds preceding the flare event. Likewise, higher than normal densities of flare plasma permit some reshaping of the ambient IMF. This is particularly true if the same region produces successive energetic flares.

For an average field spiral, a 50 MeV proton will take 1/2 to 2 hours to reach the earth. During the release of high energy particles, the flare may also release a dense "cloud" of low energy (10's to 100's KeV protons and electrons) material which travels outward, arriving at earth's orbit after 40-100 hours (Figure 9.6). This material has sufficient energy density to restructure the IMF and has been implicated in the generation of strong geomagnetic storms.

Solar wind speeds also show a cyclic variation with respect to sector structure (Figure 9.7). The abscissa represents time since SSB passage. Wind speed is strongest and density highest just after SSB crossings. Because the boundary represents a barrier to the flow of solar wind particles, the plasma tends to "pile up" just behind the SSB (i.e. it lies near the current sheet ripples) in CIRs. CIRs and the associated sector boundary crossing have been associated with geomagnetic activity (Figure 9.8). CIRs show a great deal of variability. High densities (greater than 10 particles/$cm^3$) observed at 1 AU are usually the signatures of some type of traveling disturbance in the solar wind. These discontinuities may be caused by solar flares or originate in coronal holes, disappearing filaments, or other coronal features. The average particle energy is small, less than 1 KeV. Principle constituents of the plasma are hydrogen ions and electrons. Heavy ions, mainly $He^{++}$ concentrations, tend to increase during solar wind disturbances and can normally be found at highest concentrations just behind interplanetary shock waves.

## 9.4 Long-Term Variations in the Solar Wind

Satellite observations of the solar wind since 1959 have shown that the average flow speed of 400 km/sec is subject to a great deal of variability (Figure 9.9). During the solar cycle, wind speeds and densities are lowest near solar maximum and higher during the solar minimum (particularly just prior to minimum when polar magnetic fields are a maximum). The extreme variation in these quanitities is about 30%, resulting in mass flux variations of a factor of two. This cyclic variation may result from a variation in the middle and low latitude coronal magnetic field structure. Near sunspot maximum, strong closed magnetic fields associated with active regions dominate the low latitudes. Equatorial coronal holes are small and short-lived. Near solar minimum, the low latitudes are dominated by large, semi-permanent coronal holes. These features spread high speed solar wind over large longitude ranges and eliminate many of the small current sheet irregularities thought to result from solar flares.

Whether a given coronal hole will influence the IMF near earth depends on its heliographic location and field orientation. Large, low latitude holes are, in general, more influential at earth's orbit. The field line

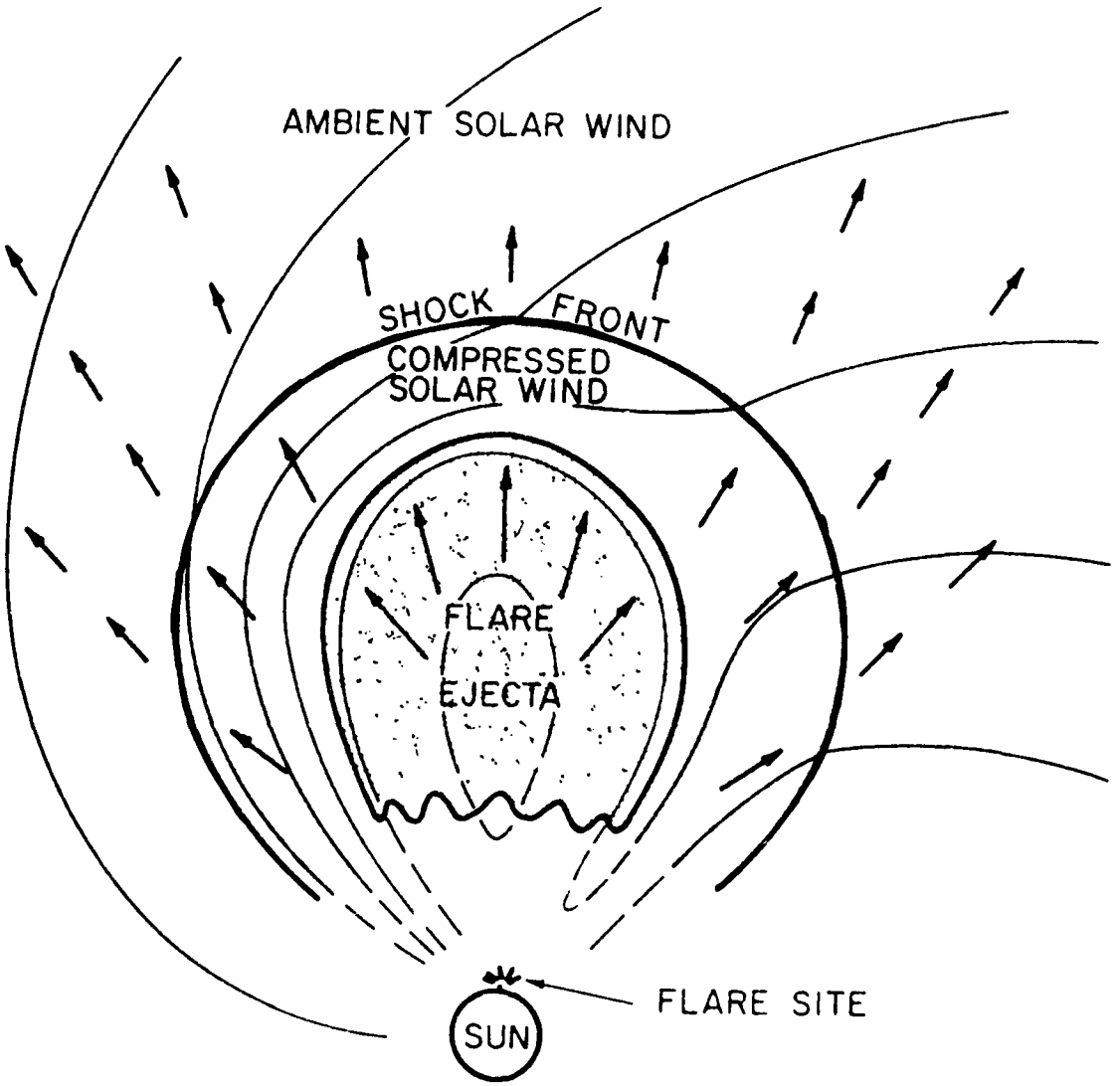
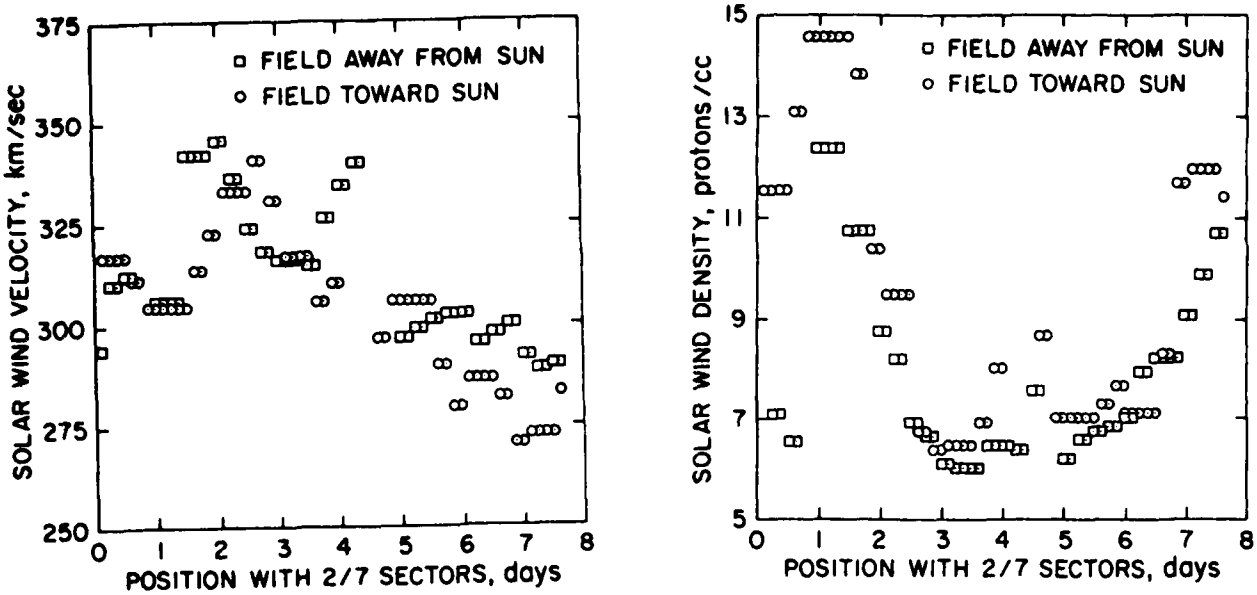



Figure 9.6 A CIR from above the ecliptic showing flare related plasma cloud and IMF compression. A qualitative sketch, in equatorial cross section, of a flare-produced shock wave propagating into an ambient solar wind. The arrows indicate the plasma flow velocity, and the light lines indicate the magnetic field (from Hundhausen, 1972).



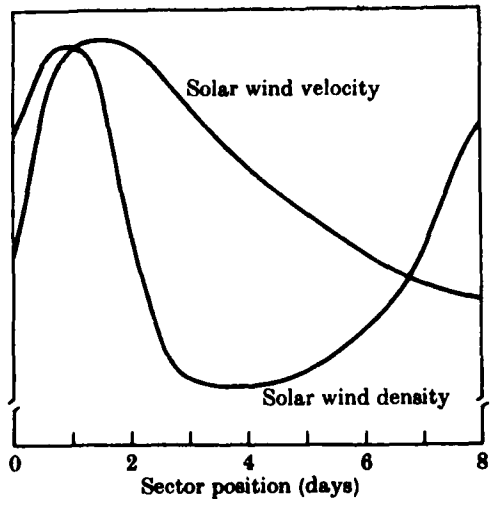




**Figure 9.7 Solar Wind and Density Variations During a Sector Transition** The distribution (along the earth's orbit) of the solar wind velocity and density within a sector. The abscissa is reckoned from the time of the crossing of a sector boundary (from Wolfe, 1972).

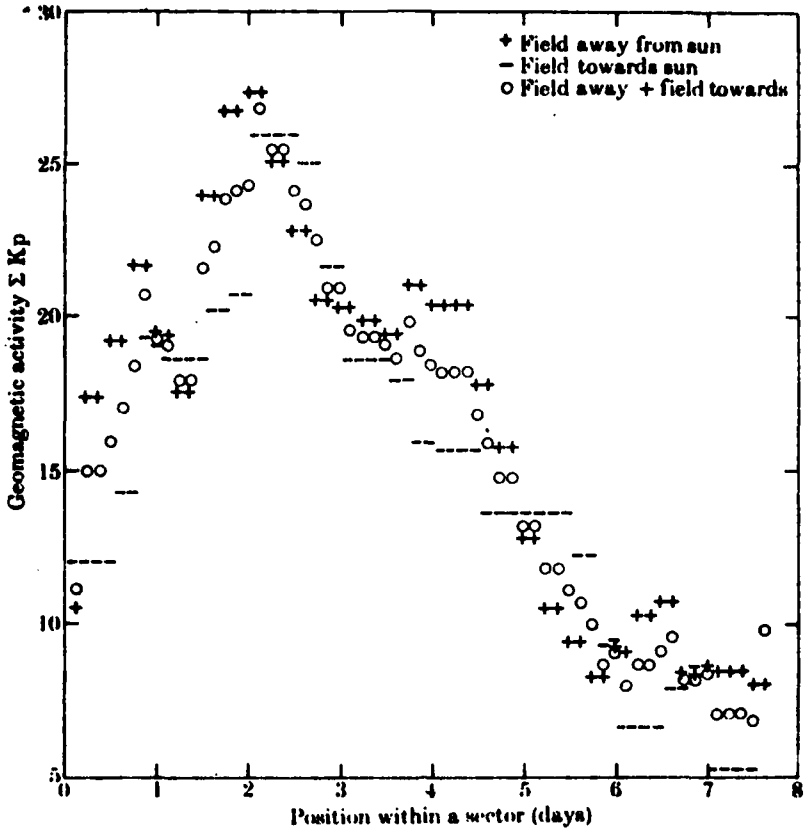



**Figure 9.8 Geomagnetic conditions during a sector crossing. Highest values of geomagnetic activity usually occur 1 to 3 days following SSB crossings (from Wilcox and Ness, 1965).**

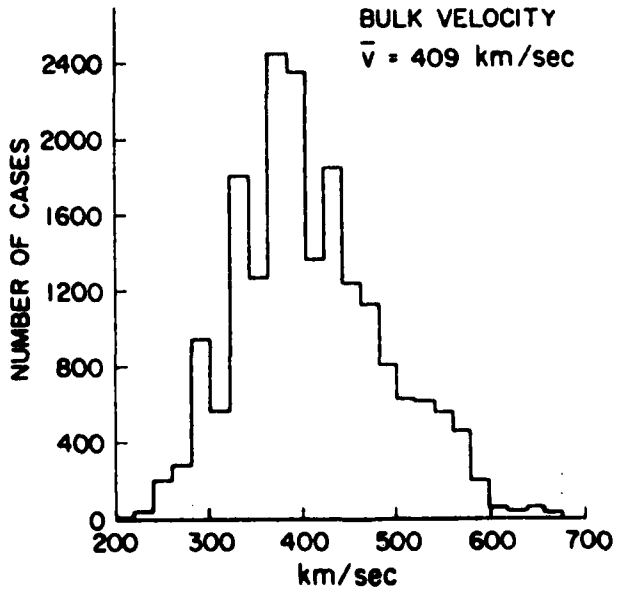



**Figure 9.9 Speed of the Solar Wind, 1962-1970 (from Wolfe, 1972).**

orientation close to the sun's surface influences the direction of the plasma stream to some degree by focusing the stream into a narrow "jet" of high speed plasma or a diffuse fan.

## 9.5 IMF Analysis

In discussing the IMF, it is common to decompose the magnetic field vector, B, into its components, $B_r$ (radial) and $B_z$ (north-south). The radial component of the IMF is that part of the total field which lies in the ecliptic plane (Figure 9.10). It has strength, measured in gauss or gamma, and direction (polarity), directed towards (-) or away (+) from the sun. Similarly, that portion of the IMF which is perpendicular to the ecliptic plane is designated as $B_z$. $B_z$ has strength and direction. In the conventional X-Y-Z coordinate system, the X axis points towards the sun. The Y axis is oriented perpendicular to the sun - earth line in the ecliptic plane, and the Z axis is considered positive upwards (north). $B_z$ is always along the Z axis, with $B_r$ existing somewhere in the X-Y plane. The angle phi is used to denote the angular displacement of $B_r$ from the X axis. It is measured eastward (counterclockwise when viewed from north of the ecliptic) from the X-axis.

Solar sector structure and the related current sheet crossings have been observed by spacecraft and inferred by ground based data. Typically, 2 to 8 SSB crossings occur each rotation. The number of sectors seems to be related to the solar cycle, with maximum solar activity producing the greatest number of sectors. This is probably due to the larger number of ripples in the current sheet from flare plasma streams. Owing to the 7 1/4° tilt of the solar equator with respect to the ecliptic, the heliomagnetic equator (current sheet) comes in close proximity to the earth twice each year. During these periods (near the solstices) repeated SSB passages may be observed. Here, even slight warpage of the nearby current sheet causes movement across the earth. Similarly, when the solar equator is at its maximum distance from the ecliptic (during the spring or autumnal equinox) the magnetic polarity corresponding to that of the nearest polar field dominates. SSB passages are less common during these times. Shortly after sunspot maximum, this relationship breaks down, as the dipole field of the sun weakens. A very irregular and poorly defined current sheet results.

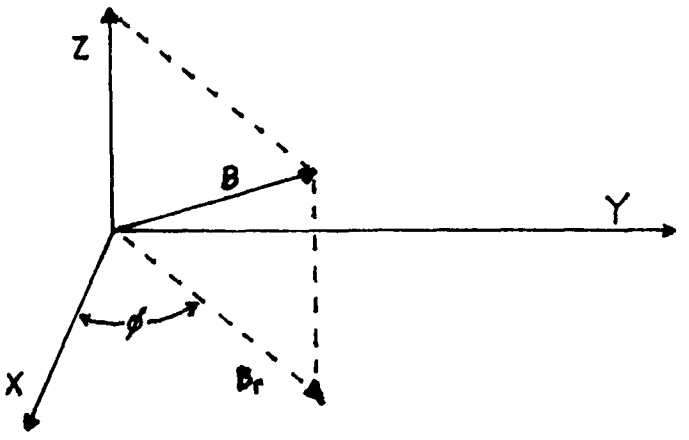



Figure 9.10 Decomposition of the IMF Vector B.

Sector boundary crossings have been observed by spacecraft and are relatively thin, being convected past the spacecraft in minutes, as compared to the width of a typical sector which takes a week or more to rotate past. In addition to the wind and density changes described earlier, the radial and perpendicular components of the IMF also change. The important consideration of SSB's is that their proximity to earth can increase the probability of geomagnetic disturbances due to solar wind enhancements and variations in the value of $B_z$. $B_z$ controls the amount of energy available to the magnetosphere for magnetic storm production. Since flare related material cannot cross the current sheet (although it may alter the sheet's position), it becomes important to know where the earth and flare source regions are with respect to the current sheet. Flare material originating on the same side of the current sheet as the earth is more likely to influence earth than material originating on the opposite side. For this reason, attempts are made to trace the sector boundary back to the solar surface.

An examination of the magnetic field configuration of a sector boundary reveals that it is an extension of the magnetic inversion line which separates major areas of opposing polarity. Since the earth orbits near the solar equator, we are interested in equatorial crossings of inversion lines. Inferred sector boundary positions (essentially photospheric footprints of the current sheet) are reported by solar optical observatories using the DALAS code. Information used in the analysis includes the following:

a. Long magnetic inversion lines which cross the solar equator in a north-south direction and divide large opposing polarity regions on <u>both</u> sides of the solar equator, and which persist for several rotations are good indicators of sector boundaries;

b. Large filaments and well defined long-lived inversion lines (i.e., they are good SSB indicators);

c. The disappearance of a filament which marked a previously well defined SSB does not necessarily indicate the disappearance of the SSB;

d. Active regions are most often concentrated in the leading (western) portions of solar sectors, while the trailing portions are often nearly void of active regions; and

e. During solar minimum, the major areas of opposing polarity which a SSB separates have an average width in solar longitude of approximately 100 degrees.

An additional aid in locating solar footprints of the current sheet is coronal emission. Enhanced green line emission (5303A) extending from N20° to S20° usually marks a sector on the limb or within one day of the limb. This analysis is performed by Mount Wilson and available through Boulder. Stanford Mean Field data measures day to day changes in the magnetic polarity of the majority of the visible disk. These observations may be used to infer the existence of a sector boundary. They are reported in the code SOLMF. When a mean field reversal occurs, a SSB is probably near or just west of solar CM. Figure 9.11 compares two successive days on which SOLMF data should indicate a polarity reversal, since the polarity of the majority of the visible has

reversed as the current sheet footprint (SSB) crossed central meridian. (The current sheet always crosses central meridian. Here, we are concerned with a north-south crossing which might cause a SSB crossing at earth.)

Due to the spiral structure of the IMF, solar sector boundaries arrive near earth approximately 4 to 5 days following CM passage of the photospheric footprint. Satellite data from spacecraft such as ISEE-3 (International Sun Earth Explorer) can be used to confirm the passage of a SSB by examining the data for the characteristic density, $B_z$, and wind speed variations. Field direction (phi angle) may also be used to infer field polarity and, therefore, the location of the current sheet with respect to the satellite location. On earth, inferred field polarities can be made using the Z-trace of magnetometer observations from Thule, Greenland and Vostok, Antarctica. These observations come to AFGWC in the code SSIMF. Care must be used not to confuse a temporarily reversed field polarity caused by some disturbance in the medium with the predominant field polarity which yields the true location of the current sheet. False indications of a SSB crossing maximize near solar

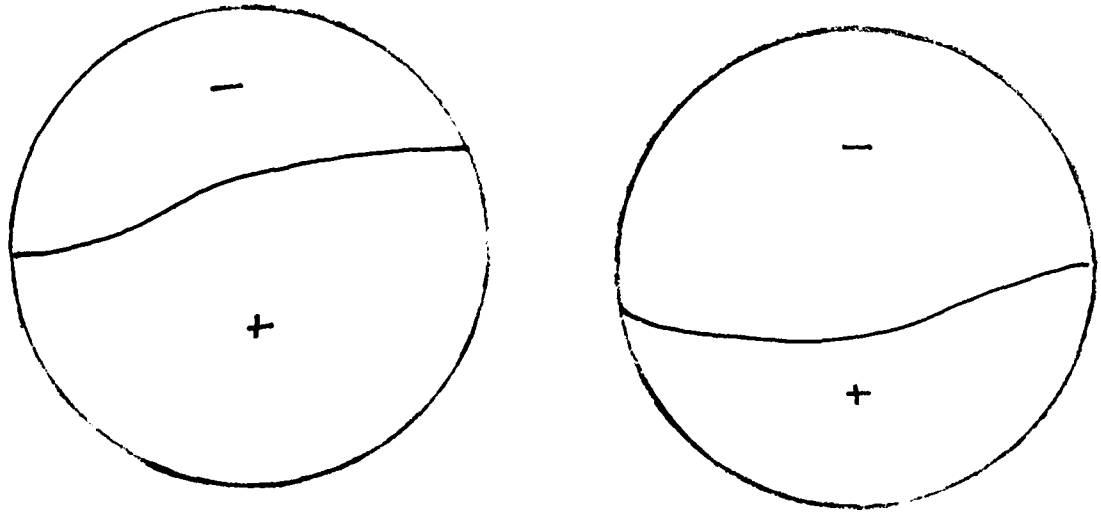



**Figure 9.11 A current sheet warpage crossing central meridian causes a reversal in the dominant polarity (plus or minus) of the visible solar disk.**

maximum when the field polarity seems to have no predominant direction. Note that ISEE-3, SOLMF, and SSIMF values should all agree for a given day after appropriate delays are taken into account. Satellite data provides the only direct measure of IMF polarity and is, therefore, the most reliable source.

## 9.6 Interplanetary Spacecraft

One of the most important sources of information on the state of the interplanetary medium is an interplanetary satellite. The Pioneer spacecraft and, more recently, ISEE-3 are such spacecraft. The ISEE-3 interplanetary orbit of interest is a halo about the sun-earth libration point. In plain language, the satellite orbits in an ellipse (semi-major axis 640,000 km, period 178 days) about the point where the gravitational pull of the sun is just equal to (and oppositely directed from) the earth's gravitational pull (about 1.5 million kilometers from the earth). This puts the vehicle roughly in the ecliptic plane on the line between the sun and earth. The orbit is described as a "halo" about the point where the gravitational pulls are equal and opposite. The spacecraft is not located precisely at that point, because it would be directly between the earth and sun. A solar radio burst could interfere with satellite-to-earth communications if the orbit were along the earth-sun line. The coordinate system used aboard ISEE-3 is shown in Figure 9.12.

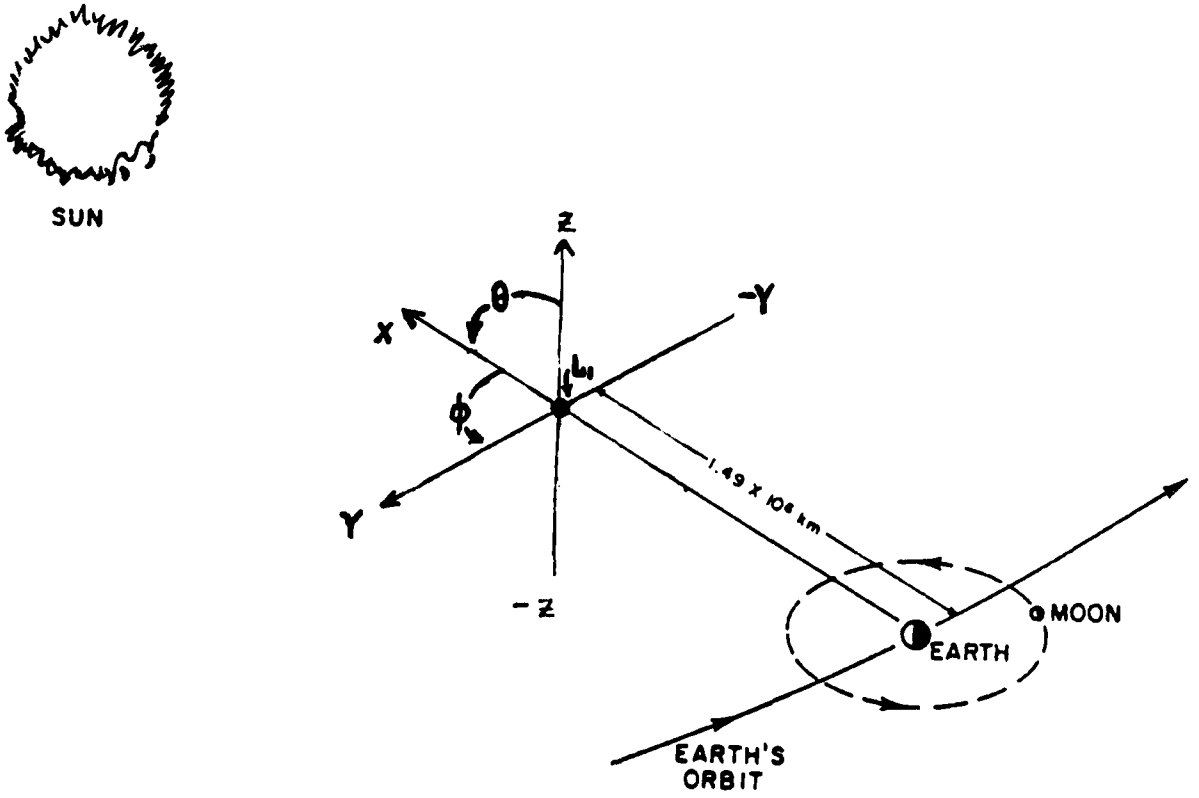



Figure 9.12 ISEE-3 Coordinate System.

ISEE-3 interplanetary data includes:

(a) Solar Wind: Obtained indirectly by measuring electron flux. Solar wind data includes density and speed;

(b) Interplanetary Magnetic Field Data: This includes total field strength ($B_{mag}$). Phi is the angle between the sun-earth line and the radial component of the IMF. Angles 270°-0°-90° indicate a negative (towards the sun) polarity. Angles 90°-180°-270° indicate a positive (away) polarity. $B_z$ is the component of the IMF which is normal to the ecliptic. For $B_z$ greater than zero, the component is directed northward. $B_z$ less than zero indicates a southward directed field;

(c) X-ray Data: Both average and peak flux is available for the following channels:

| | | |
|---|---|---|
| .6 - 1.0Å | = | 12 - 20 keV |
| .3 - .6Å | = | 20 - 36 keV |
| .2 - .3Å | = | 36 - 84 keV |

The location of the satellite upstream in the solar wind allows us to monitor conditions approximately 1 to 2 hours before they reach earth's orbit. The delay caused by transit from the satellite to earth's orbit will be somewhat less during high solar wind speed conditions.

Traveling disturbances such as shocks and discontinuities can be inferred by close analysis of such data. Jumps in wind velocities and densities, either up or down, by 30% or greater usually indicate some type of disturbance is present. Strong disturbances will also display a sharp change in the IMF strength ($B_{mag}$) as the disturbance compresses or stretches the field lines. The phi angle and Z component of the field can be used to monitor sectors and SSB crossings. In addition to these uses, the values can help assess the degree of turbulence present in the solar wind. Rapidly varying phi angle values and continual switching of the $B_z$ direction suggest a highly irregular, turbulent flow and current sheet proximity.

Akasofu's parameter, Epsilon, (Akasofu, 1978) describes the amount of energy supplied by the solar wind to the magnetosphere. Epsilon is defined by the relationship:

$$E = VB^2 \, Sin^4(\theta/2)L^2 \quad \text{ergs/sec},$$

Where

V = solar wind speed
B = IMF strength,
$\theta$ = angle between Z axis and component of IMF perpendicular to the ecliptic,
L = geometrical factor used to simulate the size of the magnetosphere presented to the solar wind; about 7 earth radii.

Quiet time values of Epsilon normally are less than $10^{18}$ ergs/sec, while major geomagnetic storms usually occur with values above $10^{19}$ ergs/sec.

Data from ISEE-3 or similar spacecraft can be very useful, but it does have some limitations. The conditions being sensed are over 1.5 million miles

from earth. We assume those conditions will be transported to earth, but this is not always the case. There will be times when, due to the Archimedian spiral, the conditions sensed by the spacecraft will pass the earth's orbit "ahead" of or "behind" the earth. This problem is acute when the satel- lite is east of the sun-earth line during low solar wind speeds. The tightly wound Archimedian spiral means conditions sampled by the satellite may arrive at 1 AU behind the earth. Likewise, when the satellite is out of the ecliptic, it is possible for the vehicle to be on the opposite side of the current sheet. Epsilon does not reflect seasonal influence brought about by the inclination of the earth's equator ($23.5^{o}$) to the ecliptic. Periods will exist when Akasofu's parameter calculated from spacecraft data will underestimate or overestimate the energy input into the magnetosphere. With these limitations in mind, the data can best be used for short term assessments of the interplanetary medium and in combination with other information to assess the current position of the neutral sheet and the solar longitude now connected to the earth. Such information can be vital in forecasting flare effects or determining the validity of optical sector boundary reports.

## 9.7. Summary

The interplanetary medium is the medium through which solar particle emission influences the earth. Its structure changes, sometimes rapidly. Our lack of observations of this medium necessitates full use of all available data and theory. Failure to do so reduces an otherwise valid analysis to pure guess work.

# CHAPTER 10

# THE MAGNETOSPHERE

The magnetic field of the earth produces a semi-permeable obstacle to the solar wind. This obstacle is the magnetosphere. The magnetosphere is defined as that region about the earth where the geomagnetic field has an important role in physical processes. The magnetosphere protects the earth from most of the direct effects of the solar wind. However, the size of the magnetosphere is related to changes in the solar wind density and velocity, and to variations in the strength and orientation of the interplanetary magnetic field. The magnetopause is the outer boundary of the magnetosphere.

To a first approximation, the geomagnetic field is a simple dipolar magnetic field (see Figure 10.1). Geomagnetic means the magnetic field of the earth, and dipolar means one north and one south magnetic pole. The word simple refers to the magnetic field being symmetric about the earth. This means the magnetic poles would have to be nearly opposite each other on the globe. The magnetic field lines, lines which show which way a compass needle would point at any location, are nearly symmetric about the earth.

## 10.1 Structure

As the solar wind leaves the corona, the energy of its constituent particles (manifested as temperature and pressure) exceeds the IMF energy density (as measured by field strength, or tension). Consequently, the solar wind plasma carries the IMF with it. IMF and geomagnetic field lines cannot cross when they encounter each other near the earth. Since the wind plasma has IMF lines frozen into it, it is excluded from the region of the geomagnetic field lines. As the plasma approaches the earth, it initially encounters a very weak geomagnetic field, which the solar wind compresses towards the earth. Nearer the earth, geomagnetic field strength increases due to compression by the solar wind and proximity to the source of the field. At the magnetopause, the outward force of the compressed geomagnetic field (plus the gas pressure of the magnetospheric plasma) is balanced by the inward force of the solar wind plasma (plus the IMF pressure). The solar wind plasma is diverted around the magnetosphere and does not approach the earth any closer than the magnetopause. The magnetopause, then, is the "surface" where this balance of forces occurs. Behind the earth (anti-solar side), the solar wind drags the geomagnetic field out into a long "geomagnetic tail", bounded by a less distinct magnetopause than on the solar (sunward) side of the earth. The distorted magnetosphere is shaped like a bullet, fairly blunt on the sunward side and nearly cylindrical for a long distance in the anti-solar direction (see Figure 10.2). The magnetopause occurs approximately 10 earth radii ($R_E$) on the sunward side of the earth and has been observed to vary between 7 and 14 $R_E$ during geomagnetic disturbances. The magnetotail extends well past the orbit of the moon (at 60 $R_E$) to possibly one thousand $R_E$ or more.

The solar wind is supersonic in interplanetary (IP) space. On encountering the magnetosphere a shock wave is formed. This shock is much like the aerodynamic shock wave formed by a blunt obstacle in the supersonic flow of a wind tunnel. The shock may be modeled by considering a speeding boat in a calm lake. Just ahead of the bow of the boat, the water particles pile up (if you were to look carefully you would observe that the water

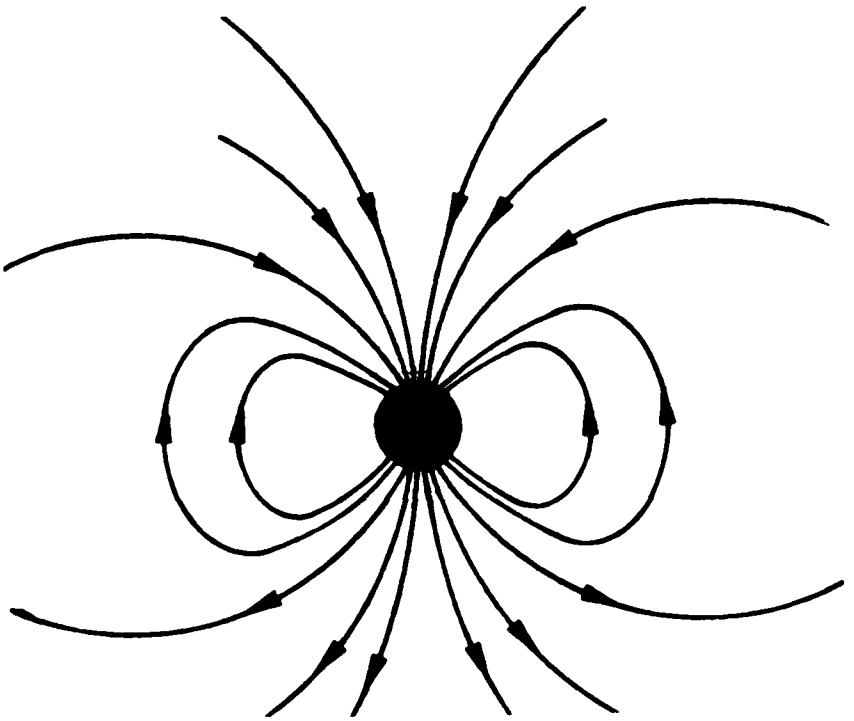


Figure 10.1 Simple Dipole Geomagnetic Field.

surface is slightly higher just in front of the bow). These "piled-up" water particles are being pushed aside by the boat. They cannot pass through the boat, so they force their way to the side of the boat. In doing so, they force the water molecules already on that side of the boat to move even further from the boat. This sets up a wave, centered on the region just ahead of the bow of the boat and streaming out on each side. This wave is called the bow wave, and anyone who has water skied has encountered one. Similiarly, the supersonic solar wind piles up ahead of the "solid" magnetosphere. As with the water molecules, the plasma particles are forced to deviate around the magnetopause. Particles pile up and a shock wave, called the bow shock, forms ahead of the magnetopause and extends outward behind the earth. In this region, solar wind speeds fall into the subsonic range.

Geomagnetic lines of force can be pictured as emerging from near the south (rotational) pole and returning to earth near the north (rotational) pole. This means that the field north of the earth's equator contains a component directed toward the earth, while that south of the equator contains a component away from the earth. Since the fields in the northern and southern halves (lobes) of the geomagnetic tail are oppositely directed, a region of field reversal, the neutral sheet, must exist. The neutral sheet is the place where the geomagnetic field switches from a component away from the earth (southern lobe) to a component toward the earth (northern lobe). The neutral sheet is a consequence of the greatly stretched field lines in the geomagnetic tail. It exists only on the anti-solar side of the earth. The neutral sheet

# MAGNETOSPHERIC MODEL

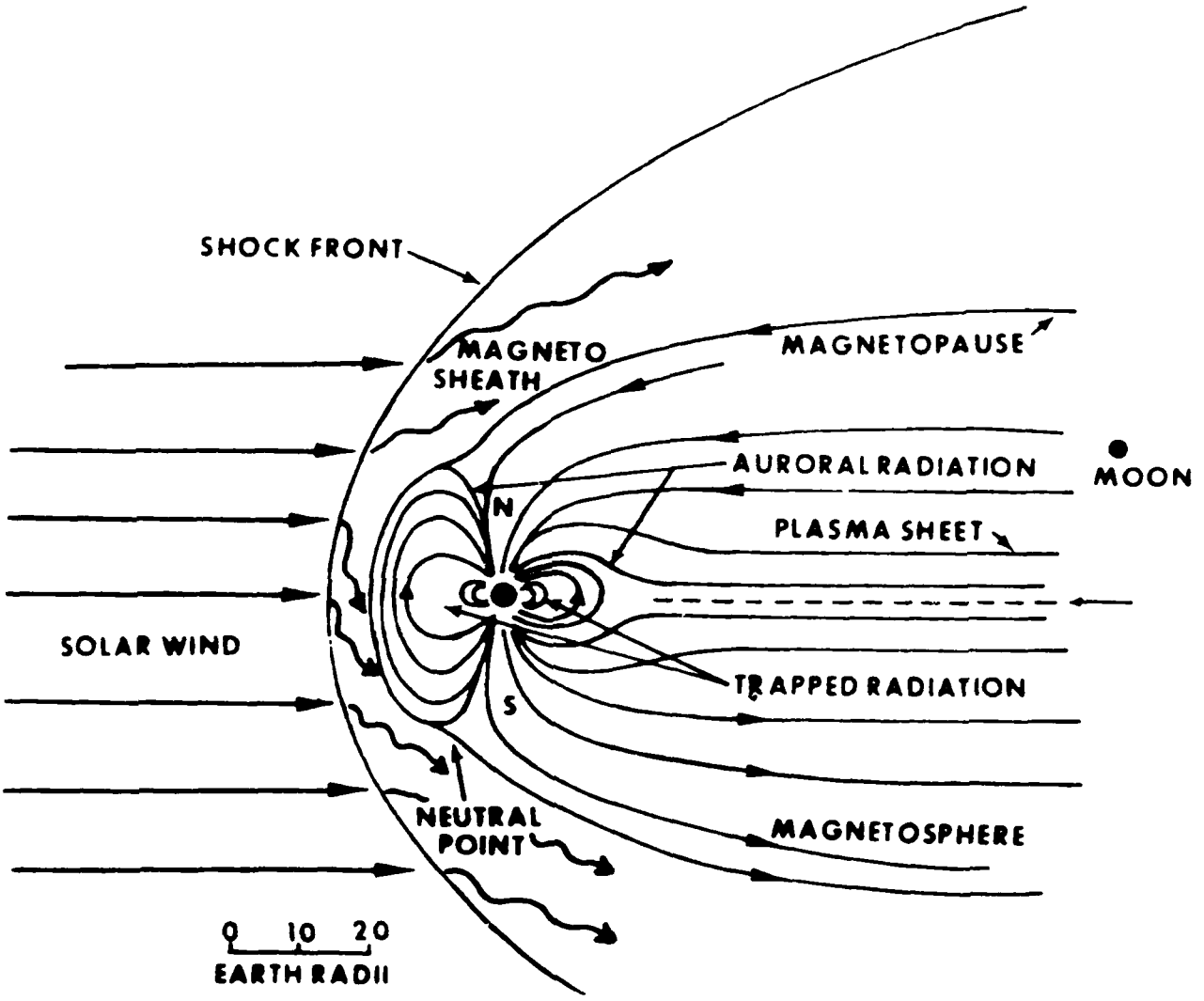



Figure 10.2 Magnetospheric Cross Section.

appears to begin above the geomagnetic equator (roughly) at a distance of approximately 10 $R_E$ and continues back through the magnetotail. The distance of the closest point of the neutral sheet to the earth varies from 7 to 13 $R_E$, approximately the same range as the point on the magnetopause directly between the sun and earth. There is a fairly distinct, near-earth boundary to the neutral sheet, and it connects to a particular set of magnetic field lines. The nearest approach of this neutral sheet boundary to the earth is associated with geomagnetic disturbances known as magnetic bays or polar substorms. The magnetic field lines just outside or, poleward of the set of field lines which define the inner boundary of the neutral sheet pass through a magnetospheric region known as the plasma sheet.

The plasma sheet is a region of relatively high energy plasma which surrounds the neutral sheet. The plasma sheet contains approximately 0.3 to 1.0 charged particles per cubic centimeter, with mean electron energies between 0.5 and 2.0 keV and mean proton energies between 2 and 10 keV. The plasma sheet has an extension earthward of the nearest approach of the neutral sheet. The magnetic field lines which pass through the plasma sheet reach the earth's surface in the area known as the auroral oval. The particles trapped on these field lines have ready access to the earth's lower atmosphere.

Equatorward of the auroral oval/plasma sheet field lines are field lines which are only slightly distorted from the simple dipolar approximation. (They are not dragged out into the tail by the solar wind.) These field lines extend through two plasma regions: the trapped radiation belts and the plasmasphere. When the first U.S. spacecraft was launched in 1958, it carried a Geiger-counter, an instrument which counts energetic particles. This instrument detected a very high density of high energy particles (actually high energy electrons). The region where they were detected was named the Van Allen belt, for the scientist who discovered them. Recently, the more descriptive name "Trapped Radiation Belts" has been used for the entire region of trapped energetic particles. The second region of plasma through which the near dipolar geomagnetic field lines extend is known as the plasmasphere. The charged particles in the plasmasphere have a relatively high density, but each particle carries relatively low energy. The plasmasphere is considered to be the upward extension of the ionosphere. It is thought to be the highest (above the earth's surface) portion of the ionosphere which co-rotates with the earth.

Two interesting features of the magnetosphere are the triangular, tongue-like regions extending from the near polar regions to the magnetopause on the sunward side. They are called the polar cusps (or clefts) and are depressions or slots in the magnetopause (see Figure 10.3). The field lines poleward of a cusp are swept far back into the geomagnetic tail by the solar wind, and those equatorward of the cusp pass through the outer portion of the radiation belt on the solar side. Those field lines which pass through the cusp exit the magnetosphere at a low altitude and connect to the interplanetary magnetic field.

The main features of the magnetosphere are (see Figure 10.4):

(1) The bow shock, where the solar wind is deflected;

(2) The magnetopause, the outer boundary of the magnetosphere which divides magnetospheric and solar wind plasma;

(3) The neutral sheet which extends back in the antisolar direction, dividing the geomagnetic tail into lobes with magnetic field lines pointing toward and away from the earth;

(4) The plasma sheet, a region of high particle density surrounding the neutral sheet;

(5) The trapped radiation belts and plasmasphere, regions of high density plasma near the earth; and

(6) The cusps, which divide the magnetic field lines which extend on the sunward side from those which are stretched out into the tail.

If the simple bipolar geomagnetic field were the entire magnetic field of the earth, we would not be concerned with it. This stable, unchanging magnetic field would always affect the near-earth environment in the same way. After we measured its effects once, they would be known for all time. There are, however, other magnetic fields affecting the earth.

## 10.2 Interaction with Interplanetary Space

The size of the magnetosphere varies with the pressure of the solar wind. This is not surprising since the location of the magnetopause is based on a balance between forces outside and those inside the magnetosphere. Higher velocity and/or higher density solar wind will compress the magnetosphere. Lower velocity and/or density solar wind allows the magnetopause to expand. Likewise, heating the magnetospheric plasma will cause it to expand. Observations of variability of the solar wind speed and density can be used to infer magnetospheric size variability.

The original (and simplest) magnetospheric model is known as the closed magnetosphere. This model requires very little interaction between the interplanetary magnetic field (IMF) and the geomagnetic field. The general opinion is that the magnetosphere behaves like the closed model when the IMF has a northward directed component ($B_z$ is positive). A northward (positive) component means the solar wind north-south component points toward ecliptic north. A northward IMF component does not easily connect into the geomagnetic field and leads to very little IMF-geomagnetic field interaction or plasma interchange. A northward IMF component produces a closed magnetosphere.

A southward IMF component allows extensive interconnection of magnetic field lines between the IMF and magnetosphere. The connection between the two is possible, because the magnetic fields (IMF and magnetospheric) are oppositely directed on the sunward edge of the magnetosphere. The magnetic field lines are thought to experience reconnection in which an IMF and a magnetospheric field line merge together and exchange their leading and trailing ends. Solar wind particles on the reconnected IMF field lines have ready access to the magnetosphere. The occurrence of this open magnetosphere is dependent on the existence of a southward IMF component.

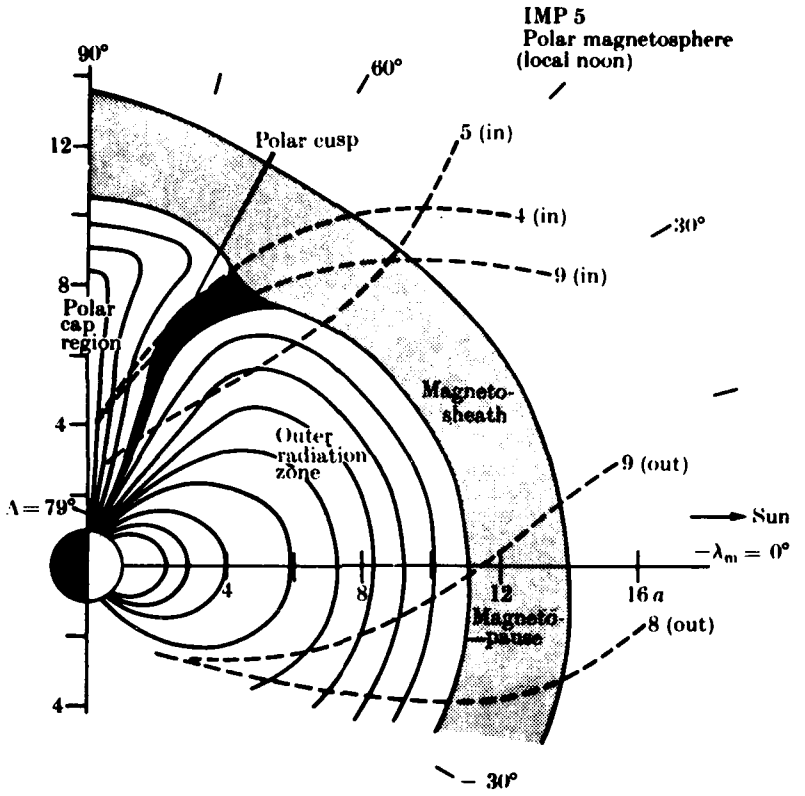



Figure 10.3 Polar cusps looking east from sunset meridian (from Akasofu and Chapman, 1972).

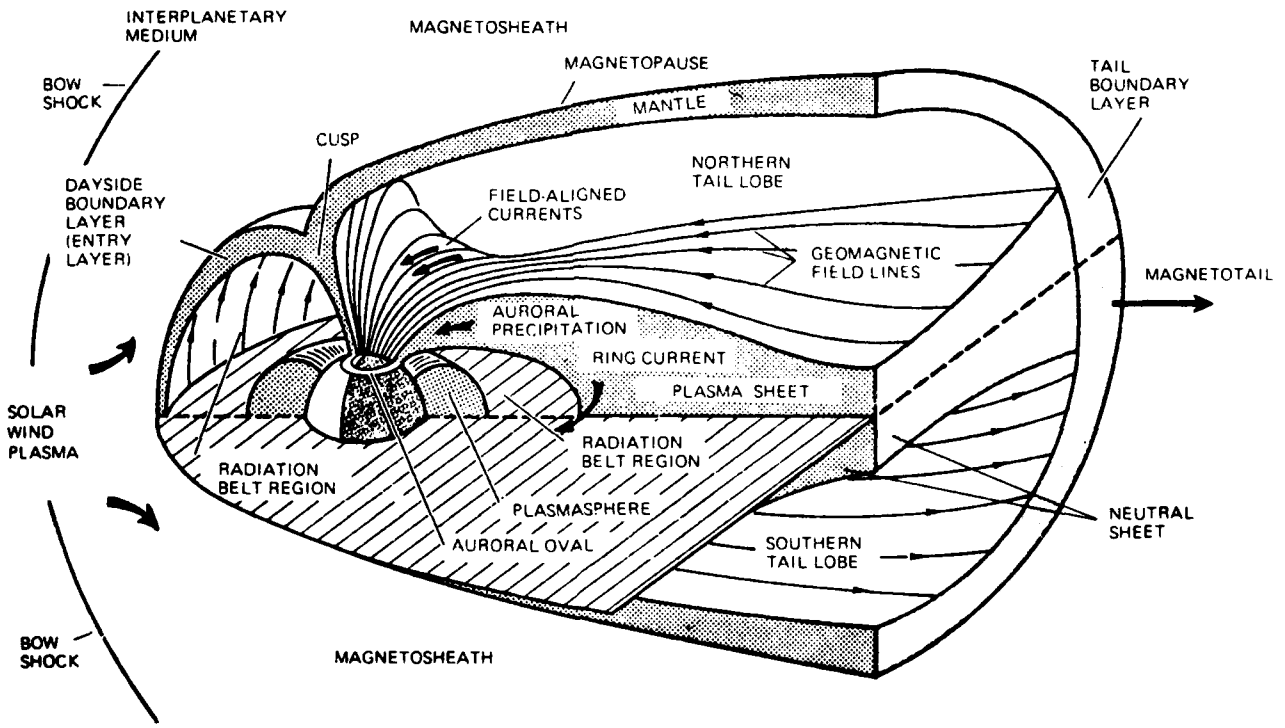



Figure 10.4 The Magnetosphere (from National Research Council, 1981).

We divide the open magnetospheric model into three regions based on the type magnetic field line each contains. Close to the earth, the magnetic field lines are closed (both ends connect back to the earth), symmetric, and very close to the dipole approximation. These closed field lines reach the earth at middle and lower (less than 60°) latitudes. Poleward of the symmetric field lines are the penetrations of closed but distorted field lines. On the anti-solar side, they pass through the plasma sheet. On the sunward side, they extend out along the magnetopause. On the night side, they are separated from the symmetric field lines by a region called the plasma cusp (different from the polar cusps!). Even further poleward are the open field lines of the magnetotail. These field lines are all swept back into the tail and connect to the IMF lines. The magnetosphere, then, is divided into open (polar) field lines, closed symmetric (lower latitude) field lines, and closed, distorted (auroral) field lines.

Solar wind particles are able to enter the magnetosphere. This entry is due to three effects. First, the open magnetosphere allows particles frozen on reconnected field lines free entry into the high latitude magnetosphere along the open polar field lines. Second, plasma particles may diffuse across field lines (see Figure 10.5). Finally, the cusps provide a direct access into the dayside auroral oval for the solar wind plasma. Since polar field lines extend from interplanetary space into the magnetosphere, the polar regions are open to the solar wind plasma. A stronger connection between the two magnetic fields will allow more solar wind particles access to the magnetosphere. This reconnection between polar and IMF lines permits polar magnetometers to sense the polarity of the IMF, and Thule and Vostok magnetometer data are used to infer the position of the earth above or below the current sheet. This information is routinely encoded in the SSIMF messages. The open polar field lines also provide ready access for high energy solar protons (above 5 MeV) to the polar ionosphere.

Diffusion of particles across field lines occurs, because a charged particle is not truly frozen to a particular magnetic field line. Some particles can diffuse (or move) across field lines. The ease and speed of the diffusion is dependent on the strength of the magnetic field, with more diffusion possible for a weaker field, and for more energetic particles. Most diffusion of particles across the magnetopause occurs where the geomagnetic field is weakest, in the magnetotail. Most low energy protons and electrons which populate the outer magnetosphere probably originated as solar wind protons which diffused across the magnetic field lines along the magnetopause in the tail.

## 10.3 Trapped Radiation Belts

Magnetic field lines which reconnect at latitudes below about 65° are generally closed and moderately symmetric about the earth. Charged particles injected into this region are trapped, at least briefly, by their interaction with the geomagnetic field. The more energetic the particle, the lower the altitude at which it is trapped, since energetic particles will continue to descend until the magnetic field strength is sufficiently high to deflect their motion. The trapped particles are loosely separated by their energy and the magnetic field into the so-called belts shown in Figure 10.6. The trapping structure is not discrete, but rather is continuous with slightly higher <u>average</u> concentrations defining the belt structure. Magnetic disturbance and

solar cycle variations modify the trapped radiation environment. Some portions of the magnetosphere are capable of only temporary trapping, while others manage trapping over months or years.

### 10.3.1 Trapping Mechanisms

Three interactions combine to account for the trapped radiation belts. Spiral or gyroscopic motion "attaches" a particle to a particular geomagnetic field line. Bounce motion (magnetic mirror effect) prevents the particle from reaching the earth's surface. The altitude gradients (in field strength) of the geomagnetic and gravitational fields cause the trapped particles to drift from one field line to another, gradually creating a nearly symmetric distribution of trapped particles about the earth.

A charged particle injected into a magnetic field (of sufficient strength) will spiral about the field line with a frequency (gyro frequency, $f_g$) given by:

$$f_g = \frac{q\ B}{(\text{const})\ m,}$$

where q = electric charge of particle,

B = magnetic field strength,

and

m = mass of particle.

The radius of the spiral is proportional to the particle mass, and the motion is a consequence of the Lorentz Force. Since the electron is much less massive than the proton, the electron is more tightly bound to the field line and spirals with a higher frequency (about 7 microseconds vs 4 milliseconds for the proton). If the particle is <u>not</u> injected at right angles to the ambient magnetic field lines (pitch angle is thus less than $90^o$) it will spiral along a field line in the direction of initial motion. Eventually, it will encounter atmospheric particles. It may recombine with these particles, perhaps after giving up energy in several collisions (indeed, this is the origin of aurora).

As the field line about which the particle is spiraling enters the atmosphere, the ambient magnetic field strength is increasing. This is a consequence of the convergence of the field lines near each magnetic pole. If B increases, so must $f_g$. For each unit of forward motion, the particle makes more and more spirals. For particles of sufficiently low energy, the convergence of the magnetic field lines will eventually halt the particle's descent. At this point, the particle may be lost (by collisions/recombination) if it is stopped at too low an altitude (where the atmospheric density is higher). Higher energy particles will penetrate to lower altitudes before the magnetic field attains sufficient strength (via convergence) to halt the particle's progress. In the auroral oval, the combination of field strength and particle energy is insufficient to stop the particle's descent before it is lost. These particles provide a steady "drizzle" which creates the quiet

(or diffuse) aurora. At lower latitudes, most particles are stopped before they are lost (most of the time). The charged particles don't just pile up here, however.

As the field lines converge, they are no longer parallel to one another. This means that in the process of orbiting a field line (even perpendicular to the field line) the particle will have a component of motion upward (or backwards along the field line) but none downward. The particle halts its descent and reverses its motion along the field line until it is again stopped at the conjugate point (same magnetic latitude and longitude, but opposite hemisphere). The effect is magnetic mirroring; the result is bounce motion. The bounce period is about 0.1 second for an electron and 2 seconds for a proton. The magnetic mirror effect does not work for particles initially injected parallel to the magnetic field line (pitch angle of $0^o$) as these particles will not spiral about the field line ($v \times B = 0$). It is similarly ineffective for a range of low pitch angles determined by the field strength along the particular line and the particle's energy.

Those particles which successfully bounce will also drift slowly about the earth, protons to the west and electrons to the east. Drift motion is a consequence of the particle spiraling about the field line and not precisely on it. Since both gravity and magnetic field strength increase as one moves closer to the earth's surface, the particle will feel different forces when it is below a field line about which it spirals than is the case when it is above the same line. The larger the radius of the spiral, the greater the variation between the above and below forces. Consequently, electrons are the least affected by drift forces and have the longest period to drift about the earth (50 minutes versus 30 minutes for protons, on the average). The actual drift motion is a combination of the effects of gravity and magnetic field gradients.

Of the three motions, drift is the most susceptible to variations in the geomagnetic field strength. Dayside compression of the magnetosphere (by solar wind pressure) means that a given particle is initially trapped at a higher altitude than would be the case on the nightside. The particle may then be lost (become untrapped) as it drifts into the nighttime hemisphere and the ambient magnetic field is insufficiently strong to trap the particle. This is one reason for the existence of the psuedo-trapping region shown in Figure 10.6.

Magnetic anomalies also exist, because the earth's magnetic axis is offset from the geometric center of the earth towards the Southeast Asian area. This deforms the low altitude trapped radiation belts and results in an unusually large amount of particle precipitation over the South Atlantic (opposite Southeast Asia). At a given altitude, the magnetic field is weaker here than elsewhere. This means that particles of a given energy are trapped at a lower altitude than elsewhere and are thereby more susceptible to collisional loss in the atmosphere.

The trapped radiation belts overlie the plasmasphere and differ from it mainly in being a hotter, lower density plasma than the component which defines the plasmasphere. The trapped radiation environment nominally extends at least to geosynchronous altitude and is, for convenience, usually considered as separate belts.

### 10.3.2 Proton Belts

Below geosynchronous orbit (6.6 $R_E$), trapped protons trace out shells which are fairly symmetric about the earth. Local time effects are generally minor (and are ignored for most uses). However, factor-of-four variations in quiet 0.6 to 3.3 MeV proton fluxes are occasionally observed between noon and midnight at geosynchronous orbit.

There is essentially only one energetic trapped proton belt at solar minimum. Higher energy particles are trapped at lower altitudes in this belt. The density of protons with energy above 400 MeV, for example, peaks less than $1 R_E$ above the earth's surface, while those with energy above 2 MeV peak near 1.5 $R_E$ above the surface. All those with energy above 0.1 MeV peak above 3 $R_E$ above the center of the earth. Protons with energies above 30 MeV result from the decay of neutrons produced in the atmosphere by incident cosmic rays. An incident cosmic ray (high energy charged particle) may collide with a neutral atmospheric particle and split apart its nucleus to release (among other particles) energetic neutrons. These neutrons will exist for only a few minutes (mean of 11 minutes) before they decay into an energetic proton and an energetic electron. The high energy protons are quite stable and are thought to have average trapped lifetimes which vary from years to centuries.

Lower energy protons trapped in the radiation belts are produced by (inward) radial diffusion from the outer magnetosphere. The diffusion process relies on multiple geomagnetic variations to move particles in to fill the trapped radiation belts. These protons have average trapped lifetimes of days to years, depending on location.

The solar cycle variation in high energy particle density in the trapped radiation belts is reasonably well understood. A decrease of a few percent during solar maximum at low altitudes is due to the solar cycle variation in neutral atmospheric density. Higher density means a higher probability of collisional loss of a trapped particle at a given altitude. Other variations of particle fluxes are also seen.

The variation best understood is the geomagnetic storm response. Low altitude trapping is insensitive to geomagnetic disturbances. Below 2 $R_E$ (primarily protons over 25 MeV), there is little response to even severe storms. Above 2.2 $R_E$ (lower energies), there is some response to major storms. At lower energies, there is an observable response to even minor storms, with decay to normal conditions requiring as long as months or even years. Near 4 $R_E$, the trapped proton flux shows a strong correlation to the Dst index. Dst is a measure of the magnetic field changes due to changes in the total energy of trapped particles in the ring current. It is generally believed that the source of low energy protons is the dumping of plasma sheet protons into the radiation belts.

The proton flux becomes very unstable above 5 $R_E$. Order of magnitude variations on time scales as short as ten minutes have been observed. At and beyond geosynchronous altitude, the fluxes are very dynamic. On rare occasions (Opp, 1968), the earth's magnetopause has been as close as 6.6 $R_E$ on the noon meridian, resulting in the complete loss of the trapped particles in that region. Solar flare produced keV protons have easy access to the outer magnetosphere during geomagnetic disturbances.

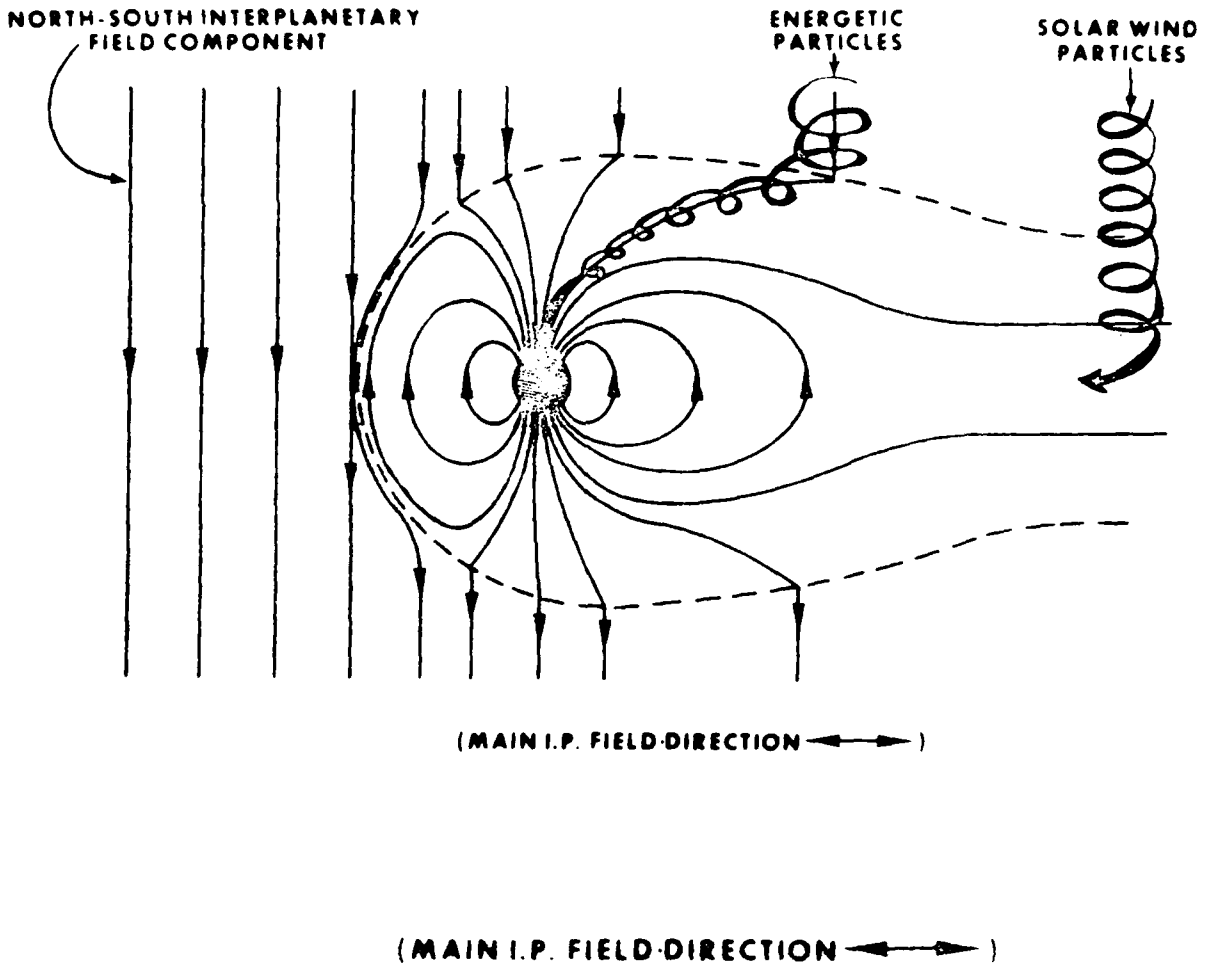



Figure 10.5 Particle Access to the Magnetosphere.

10.3.3 Electron Belts

The high energy (greater than 40 keV) electron belts exist in two zones. These are the classical Van Allen belts. The inner electron belt is quite stable to geomagnetic disturbances. Below 2 $R_E$, only very severe geomagnetic disturbances will cause measurable variations. The outer electron belt shows geomagnetic storm related enhancements at all levels.

Other effects complicate the outer zone electron density variations. The diurnal effect at geosynchronous altitudes results in a typical noon to midnight ratio of electron densities (for energies greater than 1.05 MeV) of 2.8. The IMF structure is thought to cause changes in the outer belt trapped electron population, with enhancements when the earth is in a (+) sector in northern hemisphere autumn and when the earth is in a (-) sector during northern hemisphere spring (both situations usually produce a southward-pointing IMF).

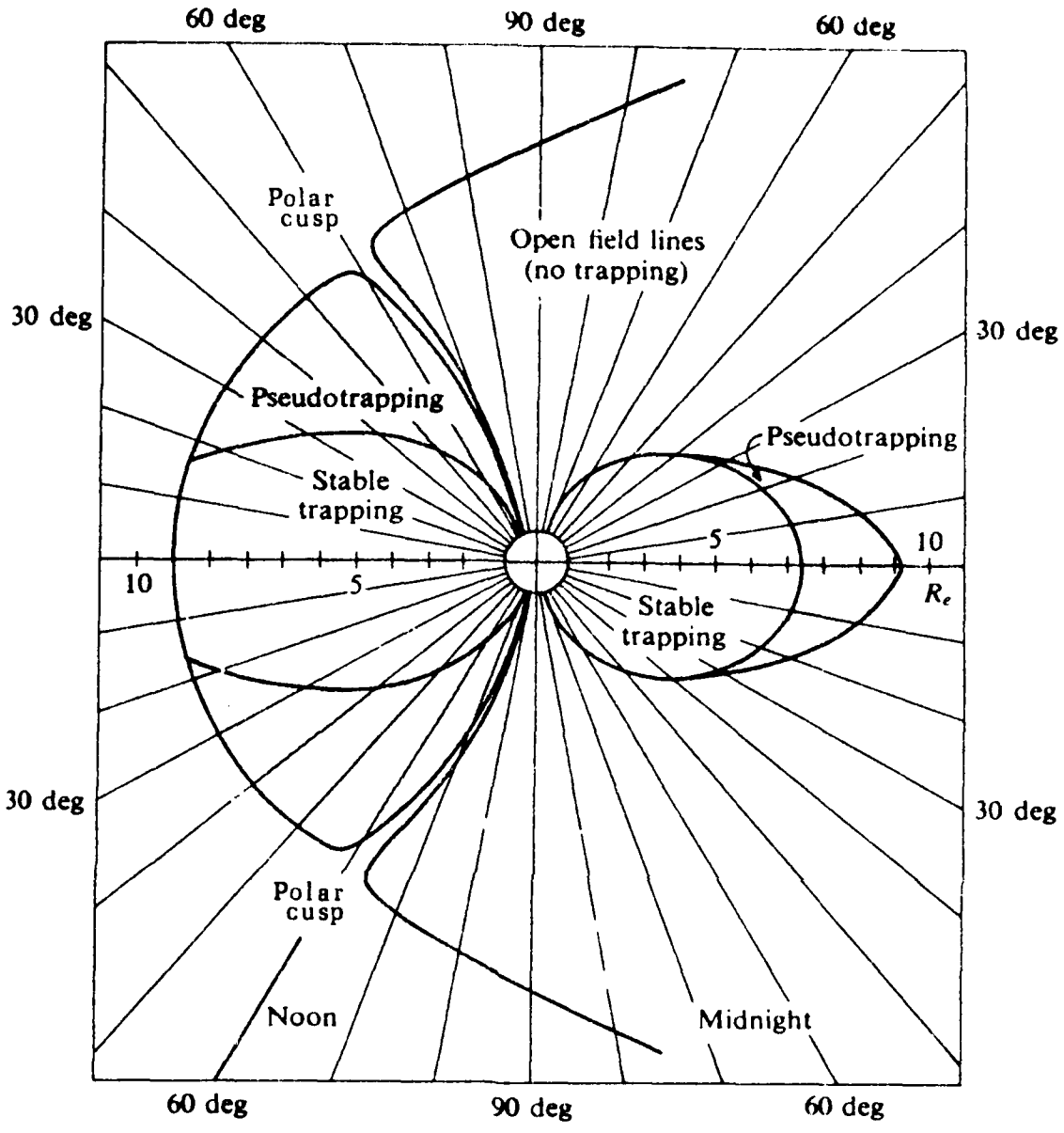



Figure 10.6 Magnetospheric Trapping Regions (Cladis, et. al. 1977).

AFGWC receives trapped particle observations from two separate sources, EP and GOES/SMS. Both are geostationary measurements and include trapped electron flux in various energy ranges between 30 keV and 2 MeV and trapped proton flux between 50 keV and 150 MeV. The GOES/SMS (Geostationary Operational Environmental Satellite/Stationary Meteorological Satellite) spacecraft observations are of electrons with energy above 2 MeV, protons in various steps between 0.8 and 500 MeV, and alpha particles in various steps between 3 and 412 MeV.

## 10.4 Main Earth Field

The magnetosphere is a consequence of the geomagnetic field, yet the geomagnetic field is also important for other reasons. Variations in the

geomagnetic field are responsible for important perturbations in the neutral density at satellite altitudes. Finally, geomagnetic disturbances are caused by physical processes that can produce drastic changes in the trapped radiation environment and in the ionosphere. Since many communication systems are linked to the condition of the ionosphere, it is very important to be able to relate geomagnetic disturbances to ionospheric disturbances.

The existence of the geomagnetic field has long been recognized. The usefulness of a magnet as a directional reference was probably known in China more than 1000 years ago and in Europe at least 800 years ago. As early as the 15th century, the notion that a compass needle points true north was discarded. Recorded measurements of magnetic declination (the deviation of the compass from true north) at various locations on the earth date back to the early 16th century, which also saw the discovery of magnetic dip (the deviation of an unconstrained compass needle from horizontal). Although experiments with magnets had been carried out since the 13th century, the concept that the earth itself is a magnet was not advanced until the end of the 16th century. It was first assumed that the magnetism of the earth was like that of a solid, permanent magnet, and it was expected to be constant in the absence of major geological changes. This view was soon proven wrong; the variation of the field over years or centuries was discovered in the 17th century. Transient variations of the field (geomagnetic disturbances) were first observed during the 18th century, and geomagnetism was increasingly appreciated to be a dynamic phenomenon. By the early 19th century, a large number of magnetic observatories had been established, both in Europe and in the distant colonies. Through coordinated measurements by many stations, the geographic dependence of some geomagnetic phenomena was discovered, and the worldwide nature of a major disturbance was established. The increasing volume and precision of accumulated data made discouragingly clear how complex were the phenomena being studied. Increasing international cooperation included coordinated investigations during the first International Polar Year (1882-1883). By this time, the correlation between the 11-year periodicities of sunspot occurrence and geomagnetic phenomena had been noted. Early in the 20th century, the intimate connection between solar and geomagnetic phenomena was further established by the correlation of recurrent magnetic disturbances to the 27-day solar rotation and later, by correlation of certain magnetic storms to solar flares. However, the most important connection, the fact that the geomagnetic field interacts with the solar wind, was established only within the past two decades. As a result of satellite investigations, recent years have seen major revisions in many of the fundamental concepts of geomagnetism.

The earth's magnetic field can be represented by a sphere uniformly magnetized in the direction of the centered dipole axis. The centered dipole axis cuts the surface of the earth at two points, A and B, known as the south and north dipole poles, as shown in Figure 10.7. The best fit between the earth-centered dipole and the actual magnetic field is obtained by taking A at 78.3°S, 111°E and B at 78.3°N, 69°W. It is obvious that the dipole axis and the axis of rotation do not coincide. The plane through the center of the earth, perpendicular to BA (the dipole axis) is called the dipole (or geomagnetic) equator. Dipole latitude is reckoned relative to this equator. The semicircles joining B and A are called the dipole meridians. The one passing through the dipole and south geographic pole is designated zero geomagnetic longitude (Figure 10.8).

Magnetospheric and ionospheric phenomena are ordered in geomagnetic coordinates. For example, the earthward extension of the plasma sheet is on magnetic field lines which intercept the earth at approximately 67 degrees geomagnetic latitude. Also, the ionosphere contains a current (the equatorial electrojet) which flows along the dayside geomagnetic equator. Normally, references to latitude and longitude when discussing ionospheric and magnetospheric activity are in geomagnetic coordinates.

The magnetic field at any point can be represented by a vector which is described by its magnitude and direction relative to a selected coordinate system (see Figure 10.9). Commonly measured magnetic field elements include (1) the total field strength, F; (2) the inclination of the total field to the horizontal plane, I; (3) the portion of the field in the horizontal plane, H; (4) the declination or angle of H from true north, D; (5) the vertical component, Z; (6) the north-south component, X; and (7) the east-west component, Y. The magnetic field vector can be described completely by a set of any three independent elements. Surface observations commonly measure either H, D, and Z, or X, Y, and Z. Note the positive directions of each element shown in Figure 10.9.

The geomagnetic field is composed of a number of fields due to various currents. These are different from the secondary field which is related to geomagnetic disturbances. At present, the terrestrial and extraterrestrial sources known to contribute appreciably to the geomagnetic field are the following:

a. Core Motion: Convective motion of the conducting fluid core of the earth;

b. Crustal Magnetization: Residual permanent magnetism which exists in the crust of the earth;

c. Solar Electromagnetic Radiation: Atmospheric winds, produced by solar heating, move charged particles produced by solar ionizing radiation (these produce ionospheric currents which generate a field);

d. Gravitation: The gravitation fields of the sun and moon produce tidal motion of air masses which generates a field in the same way as air motion from solar heating;

e. Solar Corpuscular Radiation: A number of field contributions arise directly or indirectly from the interaction of solar plasma with the main field; important effects include compression of the main field by external plasma pressure, the intrusion of solar plasma into the main field, and the heating of plasma already within the field;

f. Solar Interplanetary Field: This field is relatively weak, but at large distances from the earth it has a major effect on the interaction of the solar plasma with the geomagnetic field.

Figure 10.10 shows contours of constant F (total field strength, in gauss). Notice the anomalies in the South Atlantic, Siberia, and the Western Pacific. This field results primarily from convective motion of the core and is approximately a dipole configuration. About ten percent of the main field,

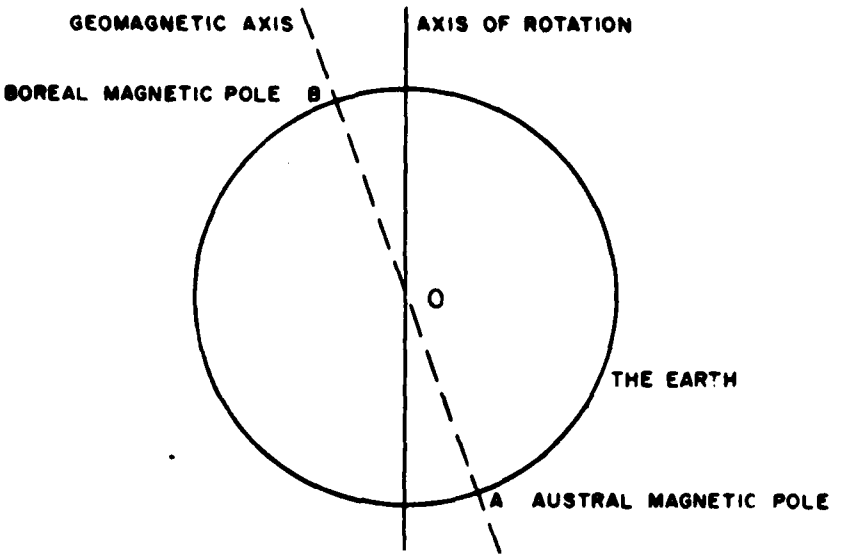



Figure 10.7 Geomagnetic Axis.

called the residual field, is not dipolar; it consists the large-scale anomalies noted above. These are believed to be generated by eddy currents in the fluid core and small-scale irregularities (not visible in this figure), arising from residual magnetism in the crust. These result in an effective displacement of the magnetic axis towards 150°E, 15°N. The main field is actually changing very slowly. This secular variation is attributed to four factors: (1) a decrease in the strength of the centered-dipole part of the field, (2) a westward drift of regional anomalies (residual field), (3) northward movement of the centered dipole, and (4) residual nondrifting variations of regional or worldwide extent.

The earth, with its atmosphere and main geomagnetic field, rotates in the interplanetary environment and moves along its orbit so that any point stationary in geographic coordinates experiences periodic variations in gravity, solar illumination, and solar-wind effects. The field contributions which vary this slowly and regularly and do not result from disturbances in the interplanetary environment are known as quiet variation fields.

Quiet variation fields include several contributions. The so-called $S_q$ (solar quiet) variation field results almost entirely from solar electromagnetic radiation, which heats and ionizes the atmosphere, producing convective flow and high conductivity in the ionosphere. This motion of charged particles generates currents which produce the field. The $S_q$ field at most surface locations has a peak-to-peak amplitude of several tens of gammas. Similarly, the tidal flow of the atmosphere arising from the luni-solar gravitational field generates currents which produce the so-called L (lunar daily) variation field, which has an amplitude at the surface of a few gammas (about a tenth that of the $S_q$ field). Another contribution results from the confinement of the main field by the quiet solar wind. Since this compression is stronger on the day side than on the night side, there is

a diurnal variation; its amplitude is also a few gammas at the surface. In the more distant regions of the magnetosphere, this effect is dominant, completely altering the field configuration.

Figures 10.11 and 10.12 show the worldwide average of the quiet daily variations. Notice that the largest variations in magnitude are near the magnetic equator. In addition, the diurnal trace varies with the season and the solar cycle. Moreover, variations are opposite direction in opposing hemispheres due to the prevailing field directions. The variations due to the moon and magnetospheric compression are much smaller and will not be discussed here. The trace of the quiet variation of a field component is known as the quiet day curve for that component.

## 10.5 Secondary Geomagnetic Field

The secondary magnetic field of the earth is produced by electric currents due to mass motions of charged particles in and near the magnetosphere. These currents produce a variable component in the magnetic field. We are primarily concerned with this variable component of the magnetic field when we specify the geomagnetic field variability. The secondary field variations are measured as deviations from the quiet day curves shown in Figures 10.11 and 10.12.

There are several currents in the magnetosphere (Figure 10.13). Indeed, organized flows of charged particles are common occurrences in the geomagnetic field, and these currents, in turn, modify the magnetic field surrounding the earth. The ring current and auroral electrojet are among the most important. The ring current is a flow of low energy charged particles trapped in the geomagnetic field. Protons with energy less than 100 keV are trapped in a toroidal (doughnut) shape which encircles the earth. These protons drift westward (clockwise as seen from above the north pole). Changes in the ring current are thought to be the cause of decreases in the geomagnetic field strength during the main phase of a geomagnetic storm. Drift of heavier ions may also contribute to this effect. During a geomagnetic substorm, there is a complex current system low in the earth's atmosphere near the auroral oval. These currents are collectively known as the auroral electrojet. The auroral electrojet is closely associated with geomagnetic bays and with auroral activity seen optically or on radar. Many other magnetospheric currents are known.

## 10.6 Geomagnetic Disturbances

A variation in the geomagnetic field which does not have a simple periodicity and which appears to result from changes in the interplanetary environment is called a geomagnetic disturbance. Large disturbances of relatively long duration are termed geomagnetic storms. The sun is responsible for all significant disturbance effects recognized at present. It is the solar wind, with frozen-in solar magnetic field, which transmits the disturbance to the vicinity of the earth. The geomagnetic field and the plasma it contains are compressed by the on-rushing solar wind until the "pressures" balance, and the solar wind is diverted around the magnetopause. Abrupt variations in the solar wind and IMF parameters obviously change the location of the balance (magnetopause) and the degree of compression of the geomagnetic field. This change in magnetospheric size and the resulting internal adjustments of the magnetosphere produce the phenomenon known as a geomagnetic storm.

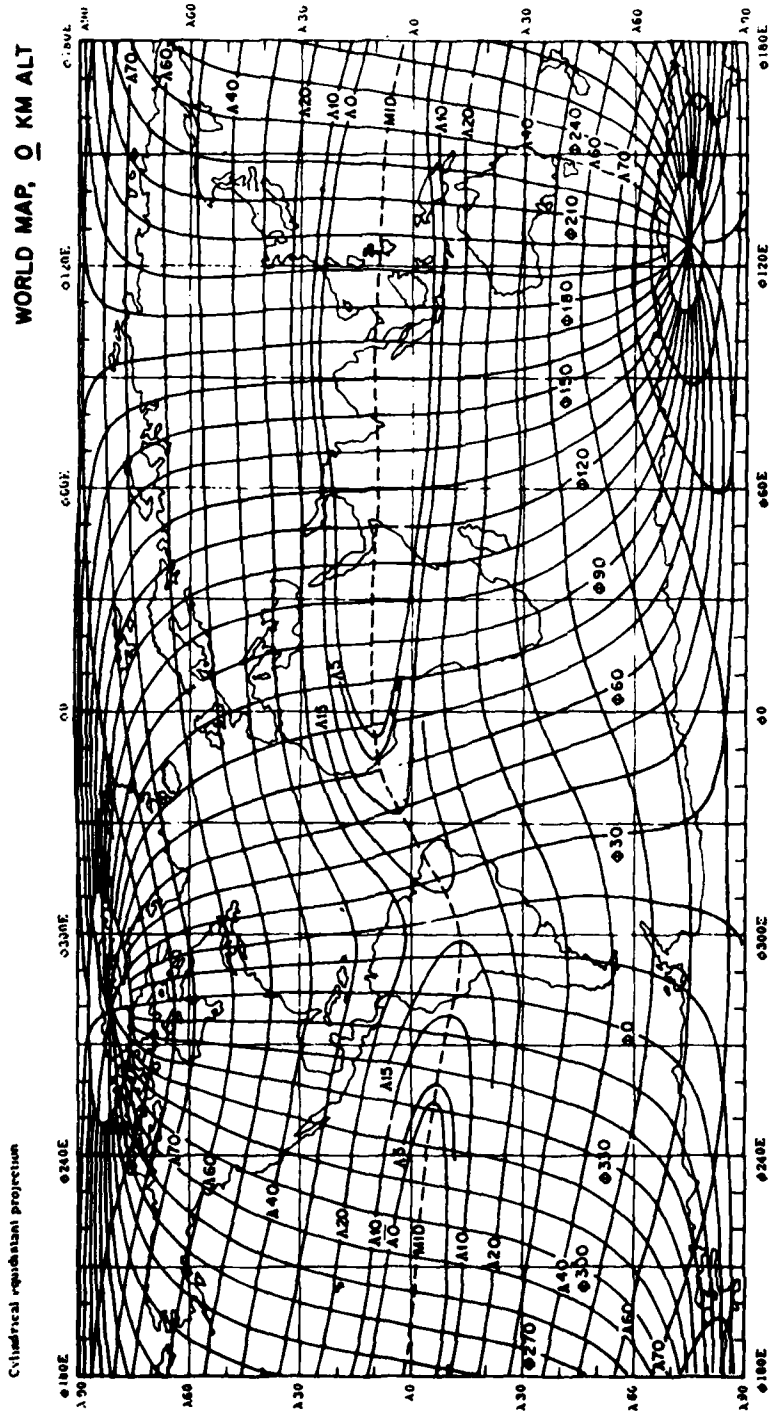
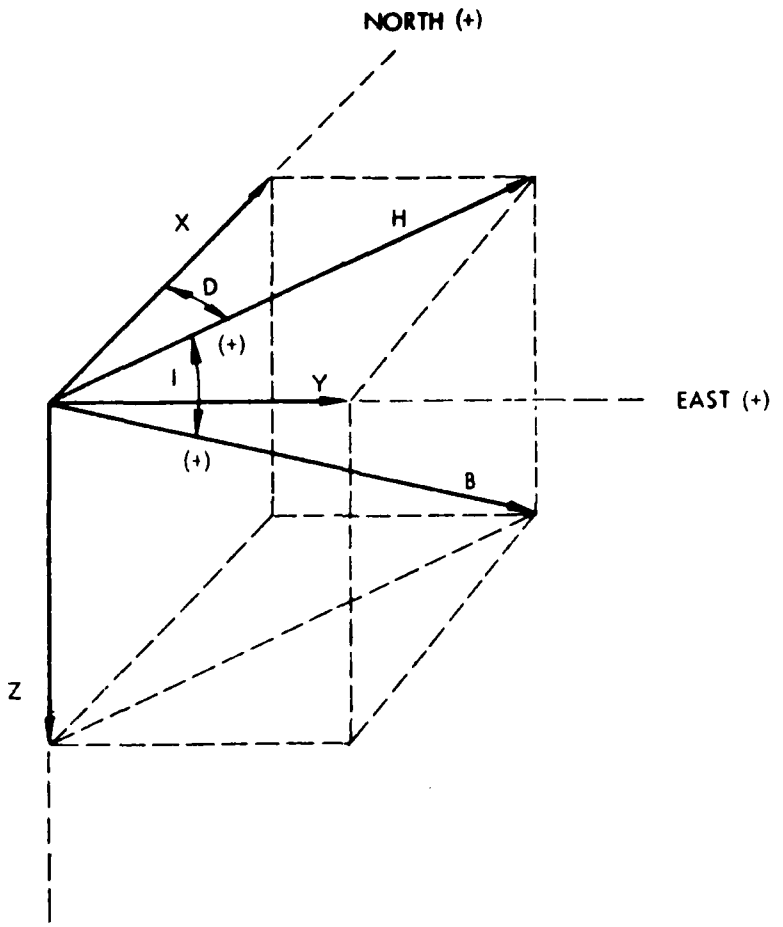



Figure 10.8 Geomagnetic Coordinates (Cladis, et. al. 1977).




D = DECLINATION
H = HORIZONTAL INTENSITY
B = TOTAL INTENSITY
I = INCLINATION
Z = VERTICAL COMPONENT
X = NORTH COMPONENT
Y = EAST COMPONENT

Figure 10.9 Elements of Geomagnetic Field (Prochaska, 1980).

### 10.6.1 Storm Phases

The classical geomagnetic storm observed by ground based instruments is composed of the average behavior recorded by magnetometers. Although there are large variations between individual geomagnetic storms, phases of the classical storm are used to describe the storms as a group.

A typical sudden-commencement magnetic storm shows four characteristic phases (see Figure 10.14): (1) sudden storm commencement (SSC), an abrupt increase in the horizontal magnetic field (H) seen nearly simultaneously worldwide; (2) an initial phase, lasting from a few minutes to a few hours during which H decays to pre-storm value (this phase is not generally characterized by large, random variations); (3) a main phase, lasting from about one to three days, in which H is below the prestorm value, first decreasing and then increasing more slowly toward the pre-storm value and

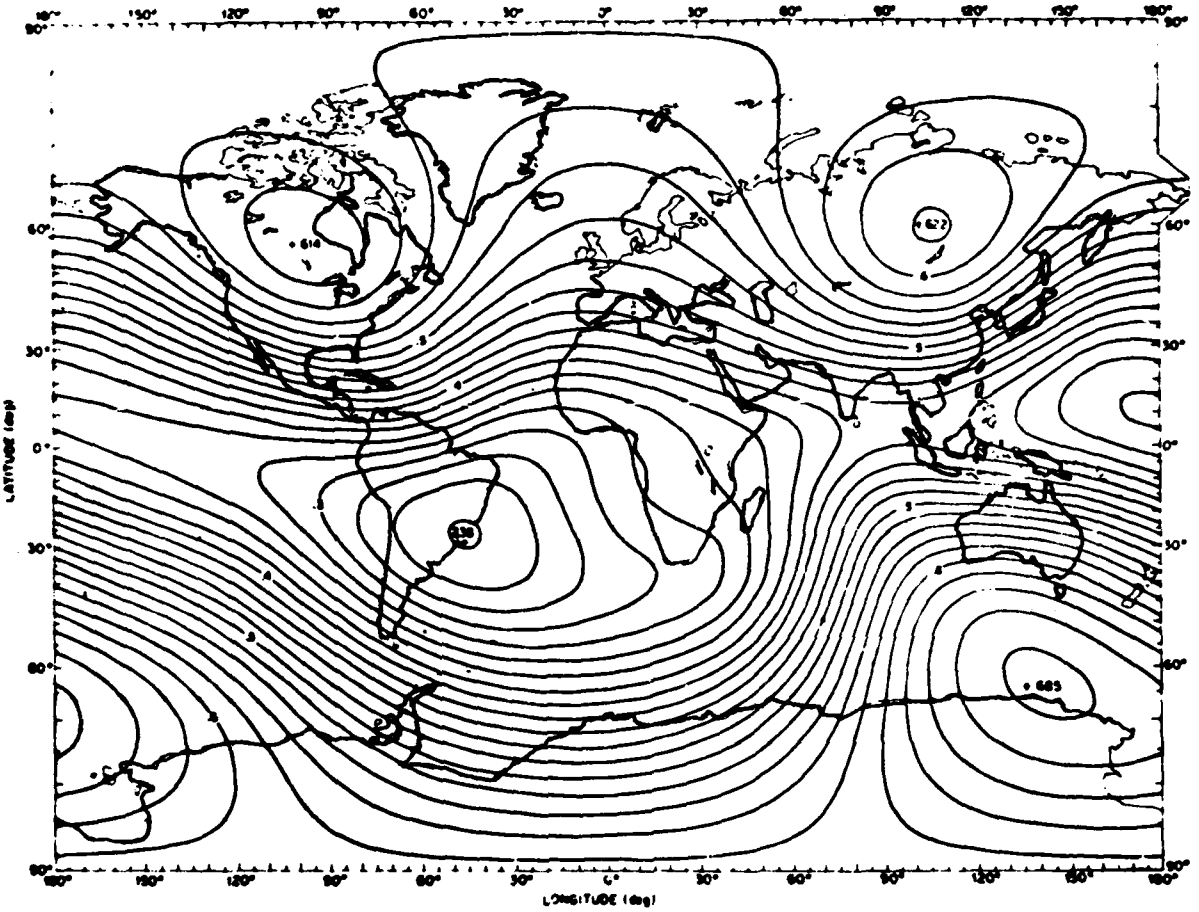


Figure 10.10 Contour Map of Geomagnetic Field Strength (note tilt to rotational equator) (Cladis, et. al. 1977).

during which there are many large, random variations of the field; and (4) a recovery phase after the end of the main phase (as indicated by a decrease of the random variations) in which H continues to rise toward, or perhaps slightly above the pre-storm value.

We can explain the storm phases using a shock wave ahead of a dense plasma cloud in the solar wind. The magnetosphere is compressed by the shock and adjusts to the solar wind in and following the shock wave. The sudden storm commencement is commonly identified with the compression of the magnetosphere by the shock; the initial phase with the magnetosphere's adjustments to the shock wave; and the main phase with arrival of the post-shock plasma cloud which establishes the ring current.

Before a storm, at subauroral and polar latitudes, the magnetic field is relatively quiet, with only slow amplitude variations. Even though the magnetic field is "quiet" in terms of the mean worldwide activity, there are daily disturbances (substorms and magnetic bays) in the auroral zones. Regular

diurnal patterns of visual and radio aurora are observed. Other observing sites experience normal quiet field variations, due mostly to $S_q$ and L.

During the storm onset, that is, during the first 5-10 minutes after the SSC impulse, no major storm phenomena (other than the SSC) are seen at low latitude. Here, the impulse almost invariably produces an increase in the horizontal field strength, with a rise time of 1-10 minutes. The shape of the SSC is more variable and may not be clearly visible in the auroral or polar regions because of the increased background level of agitation. Not all geomagnetic storms have a discernable SSC. Moreover, an SSC-like impulse (called an SI-sudden impulse) may occur without the following plasma cloud.

During the initial phase at low latitudes, the mean value of H often remains above normal for several hours immediately before the main phase of the storm. This interval is called the initial phase. In many (but not all) storms, the SSC impulse coincides with the start of the initial phase. The amplitude of the initial phase tends to decrease with increasing latitude, and a well defined initial phase is not normally visible at auroral and polar latitudes. Small variations in amplitude, with a periodicity of minutes rather than hours, occur immediately following the SSC. These variations, or micropulsations are strongest at auroral latitudes. Particles dumped from the trapped radiation zones or injected from the magnetotail attempt to stabilize on lower (stronger) field lines during this period.

At low latitudes, the onset of a large depression in H marks the beginning of the "main phase" of the storm. At high latitudes, large-amplitude magnetic bay disturbances occur. The magnetogram trace at high latitudes is complex, with overlapping positive and negative bays. Irregularities in the shape of the low latitude depression in H can often be correlated with details of the high latitude bays. Disturbances largely confined to the auroral zone during the prestorm and initial phase now spread to much lower latitudes. The auroral oval will also expand and move equatorward during the main phase of a storm. The main phase for intense storms tends to show a greater rate of change than the main phase for weaker storms; i.e., larger variation but shorter duration.

The main phase is thought to result from a combination of two actions. The plasma sheet is forced inward on the nighttime side of the earth, approaching within 5-6 $R_E$. The prevailing cross-tail current (flowing from sunrise to sunset) augments the quiet ring current. Simultaneously, the solar-wind plasma which followed the shock front flows into the ring current system. The ring current is now actually a sheet current extending from 2-3 $R_E$ above the surface to, perhaps, the inner surface of the magnetosphere. It is made up of a westward flow (clockwise from above the north pole) of keV protons. The current will generate a magnetic field which is opposite to the main earth field. The net result is a decrease in field strength measured at low latitudes (where most of the main field is parallel to the earth's surface and anti-parallel to the ring current field).

The current sheet is essentially in the plane of the geomagnetic equator. Nonetheless, the protons' random motions will carry them (or some of them) along field lines. Some will be lost when they collide with atmospheric particles at low altitudes in the high latitudes. This causes heating and expansion of the atmosphere, which causes more particles to be lost. Visible

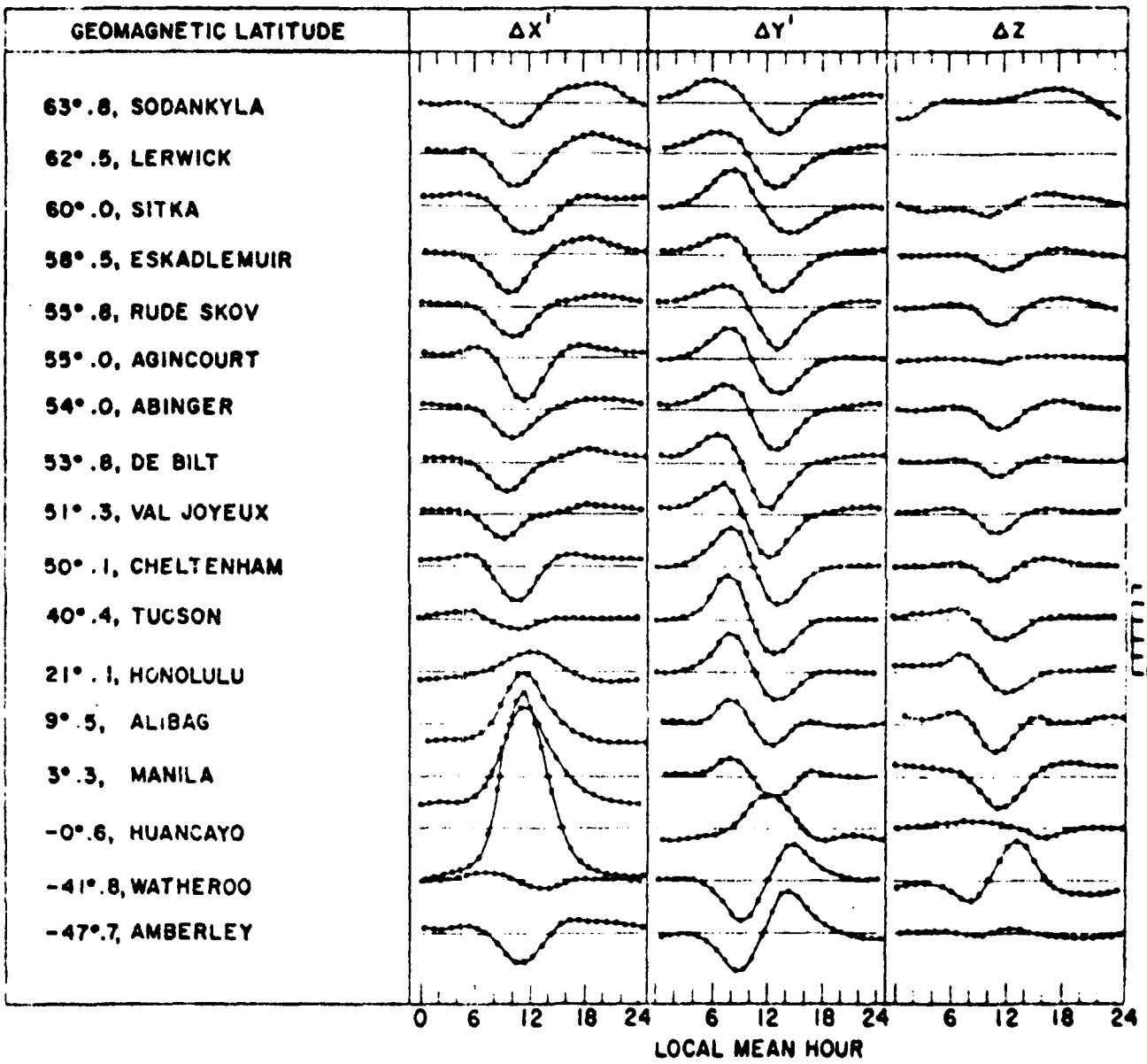



**Figure 10.11** Quiet Day Magnetometer Traces at Various Latitudes (from Valley, 1965).

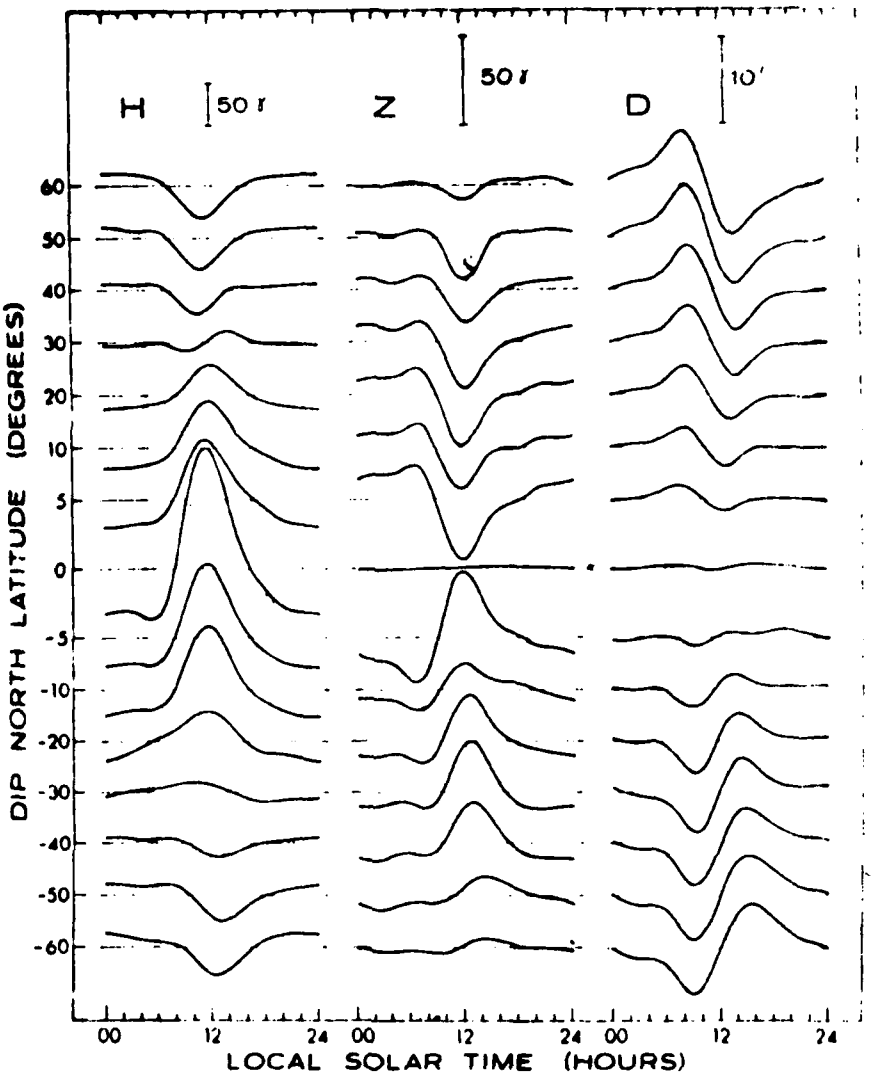



Figure 10.12 Quiet diurnal magnetometer variations for equinoctual months near solar maximum (from Akasofu and Chapman, 1972).

and radar aurora also occur in the auroral ovals. Density fluctuations in the plasma stream feeding the ring current and variable loss rates account for the variations observed on magnetometer traces. Of course, additional shocks may be embedded in the plasma stream. When the plasma stream is depleted, the remaining ring current will be gradually absorbed by the aforementioned processes, and the storm will abate.

At low latitudes, the H component gradually returns to the prestorm value during the "recovery phase". At high latitudes, the amplitude of the bay irregularities decreases, and the bays are again isolated as in the prestorm phase. The H component is well below its prestorm value at the beginning of the recovery phase, and may remain below normal for several days.

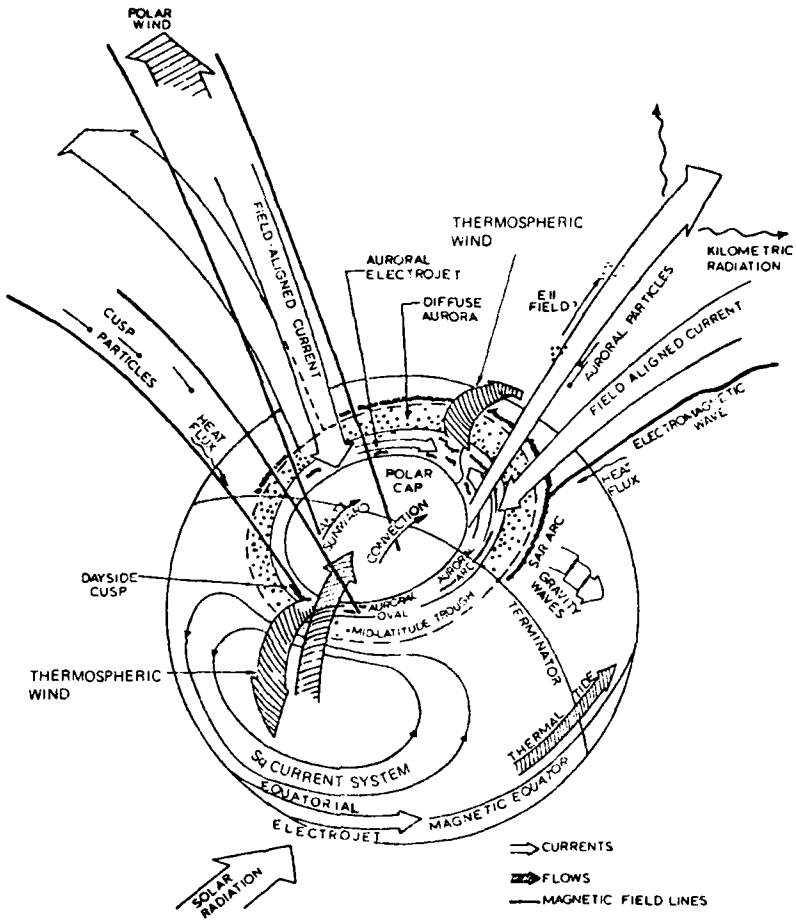



Figure 10.13 Magnetospheric Current Systems (from National Research Council, 1981).

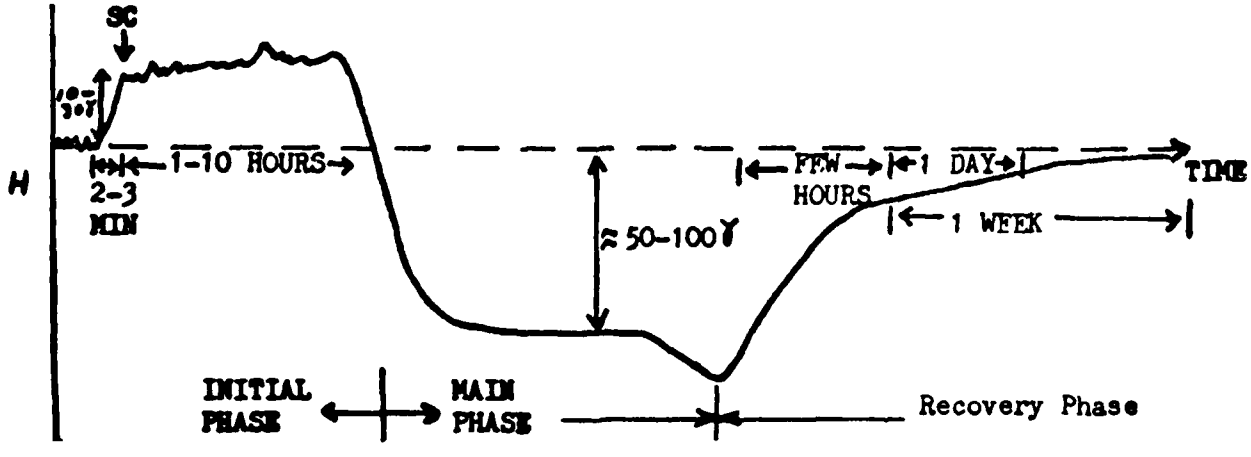



Figure 10.14 Phases of Classical Geomagnetic Storm.

Geomagnetic storm features vary with the location considered. Figure 10.15 shows magnetogram traces illustrating the development of a sudden commencement storm in (1) the polar region (Godhavn), (2) the auroral zone (College), and (3) at low latitudes (Honolulu and Huancayo). The largest disturbance variations occur in the auroral zones; these are about 10° wide centered at about $\pm$ 67° geomagnetic latitude (see Figure 10.16). During a large magnetic storm, the maximum range of the field components at an auroral zone station will be about 2500 gammas or more. The amplitude of the disturbance decreases with increasing latitude to about one-half the auroral value near the geomagnetic poles. At latitudes lower than the auroral zone, a sharper decrease occurs. The disturbance level drops to about one fifth the auroral value at 50° geomagnetic latitude and reaches a broad minimum of about one-eighth the auroral value below 30°. Near the equatorial electrojet, there is a secondary maximum of about one-fourth the auroral zone value.

The largest variations of the field are highly irregular and do not repeat in any predictable fashion from one day to the next. Generally, the largest variations have durations of about 15 minutes to 3 hours, with amplitudes decreasing for both longer and shorter durations. At auroral zone stations, these random variations tend to be shorter but with higher rates of change than at lower latitudes. There does not appear to be a sharp dividing line between normal (or quiet) conditions and abnormal or storm conditions. Broadly speaking, magnetic storms are strongest in the auroral zones and near the geomagnetic equator.

### 10.6.2 Types of Magnetic Storms

Magnetic storms may be classified according to the characteristics of individual storms or according to their relationship to other storms. In the first case, the usual division is into sudden commencement (SC) and gradual commencement (GC) storms; in the second, into isolated and 27-day recurrent storms.

Intense flares usually precede the strongest storms by about two days, and most of these storms are sudden commencement ones. Optical flares of importance two or greater show a close statistical association with strong geomagnetic storms, especially if the flare was accompanied by major radio bursts. Since type II bursts are produced by a shock wave propagating through the solar corona, type II bursts are generally associated with SC storms.

Recurrent storms are thought to be due to current sheet crossings (and other long duration phenomena) which rotate with the sun. Strong recurrent plasma streams may also produce an SC storm. The division of storms into SC and GC categories and into isolated and recurrent ones is considered two independent processes.

The designation of a storm as SC depends on the identification of a storm sudden commencement (SSC) impulse occurring near the beginning of a storm. There is no precise definition of an SSC impulse. When a sudden change in the amplitude of a standard magnetogram record, with rise time of less than about 10 minutes, is recorded simultaneously (within one minute or so) by several observatories, this event is called a "magnetic impulse". Prevailing practice seems to be to list a storm as SC if at least two widely separated stations record such an impulse 24 hours or less before the onset of other storm

characteristics. There are several complicating factors to this method of classification. Magnetic sudden impulses (SI), similiar in all other respects to SSC's, often occur in the absence of storms. Many storms are preceded by several, rather than one, impulse, and additional impulses frequently occur during the storm. Impulses preceding strong storms are generally larger than impulses preceding weak storms, although there are many exceptions. It follows that impulses preceding strong storms are less likely to be obscured by random fluctuations than are the impulses that precede weaker storms. The strongest storms are usually SC, but it is not known whether this occurs primarily from the method of storm classification or from an intrinsic difference between SC and GC storms.

A somewhat similar problem arises in attempting to separate isolated from recurrent storms. A storm is considered to be isolated unless it clearly forms part of a recurrent sequence (i.e., recurring approximately every 27 days for several rotations). The selection of recurrent sequences depends entirely on terrestrial observations, and standard selection criteria have not been established. There is, moreover, a marked annual variation of terrestrial storm activity which makes it more difficult to identify storms during the solstitial months. These problems are not too serious during years of low solar activity when recurrent storm sequences are the dominant phenomena. During the more active years, however, the identification of recurrent geomagnetic storms becomes very subjective.

Identification of a recurrent geomagnetic storm is attempted using a plot of geomagnetic activity versus day of the solar rotation. A recurrent storm is one which appears at the same part of at least three successive solar rotations and is of similar intensity. The recurrent storm may slowly walk in time; that is, it may have a recurrence period slightly more or less than 27 days. This may be due to differential rotation of the source region or modification of the plasma stream by subsequent solar activity. A true recurrent storm will <u>not</u> skip a rotation or fluctuate wildly in intensity from rotation to rotation. Recurrent storms are generally not the largest geomagnetic storms, but they may be extensively modified by large solar flare storms.

## 10.7 Geomagnetic Substorms

Two theories currently exist on the production of geomagnetic disturbances. The older suggests energy storage by the magnetosphere. In this theory, the magnetosphere continuously removes energy from the solar wind and IMF and stores it in the stretched, open magnetic field of the geomagnetic tail. This stored energy is sporadically released in explosive bursts known as substorms. The energy is dissipated in atmospheric heating, emission of electromagnetic waves (e.g., light) as aurora, and in production of electrojet currents. Under this theory, substorms are produced in a slightly different way from geomagnetic storms.

The second theory holds that the magnetosphere is similar to a dynamo. Such a dynamo is capable of converting as much as $10^{12}$ watts from the solar wind into disturbance energy at any time; the actual amount converted at any instant is dependent on the amount of energy delivered to the magnetosphere. Measured by the Akasofu parameter (Epsilon) (Akasofu, 1978), the energy

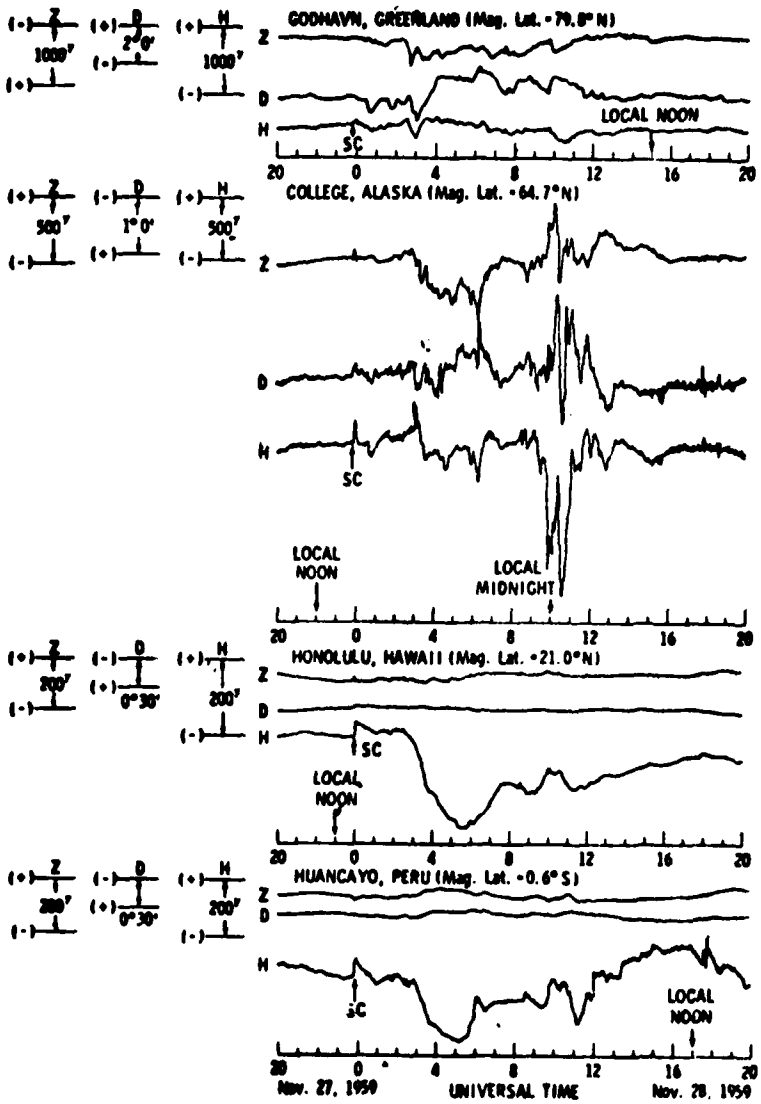



Figure 10.15 Development of an SSC Storm at Several Latitudes (Hess and mead, 1968).

available establishes limits on the type of effects to be expected. The ranges on Epsilon (in ergs/sec) and anticipated effects are:

$10^{17}$-$10^{18}$ - polar cap effects,
$10^{18}$ - substorm onset,
$10^{18}$-$10^{19}$ - substorms, aurora, minor geomagnetic storm, and
$10^{19}$ - major geomagnetic storms.

Regardless of the theory, the north-south component of the IMF is related to the frequency and intensity of occurrence of geomagnetic disturbances. Since this IMF component controls the injection efficiency of energy into the magnetosphere, it is not surprising that it would affect the frequency of energy release. A southward component increases the energy storage (or

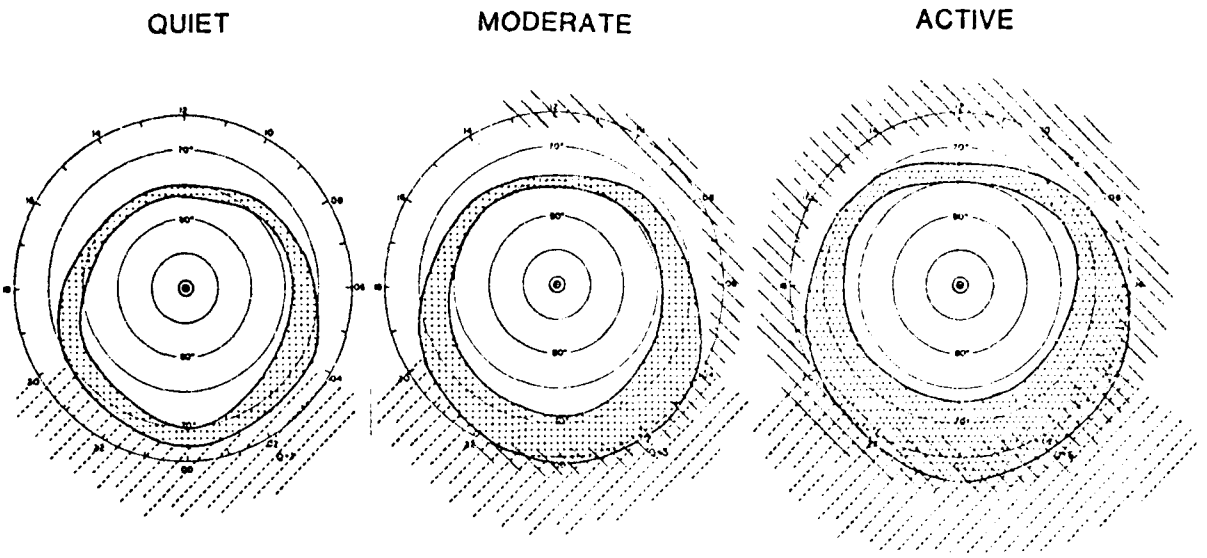



Figure 10.16 The average auroral oval position (dots) for varying levels of geomagnetic activity. Dashes identify the subauroral trough and lines the region of particle precipitation.

delivery) and the substorm/storm frequency; a northward component decreases both. Unlike the effect of a solar flare on the ionosphere (which is maximized by the sharpness of the event--how quickly does it rise and how large does it get?), a geomagnetic disturbance seems to depend purely on the amount of energy available. It will remain at a given level of disturbance as long as energy input continues. Substorms seem to result from small, brief energy injections.

A geomagnetic storm is a period of frequent, intense substorms. Moreover, the onset of the main phase of a geomagnetic storm coincides roughly with a southward turning of the IMF. A northward change of the IMF seems to coincide with the sudden end of the substorms. Indeed, the erratic field variations which characterize the geomagnetic main phase observed at high latitudes are commonly identified with substorms.

A magnetospheric substorm is most dramatically manifested in the auroral substorm. The auroral substorm has two characteristic phases: the expansive phase and the recovery phase (see Figure 10.17). The first indication of a substorm is a sudden brightening of one of the quiet auroral arcs lying in the midnight sector of the auroral oval (or a sudden formation of an arc at that location). This is followed by a rapid poleward motion of the brightened arc, which results in an "auroral bulge" in the midnight sector. This bulge expands in all directions. In the evening sector (sunset to midnight quadrant) of the expanding bulge, a large scale fold appears in the arc and travels rapidly (up to 5 km/sec) westward along the oval. This is referred to as a westward traveling surge. In the morning side of the bulge, arcs appear to disintegrate into "patches" which drift eastward and disappear.

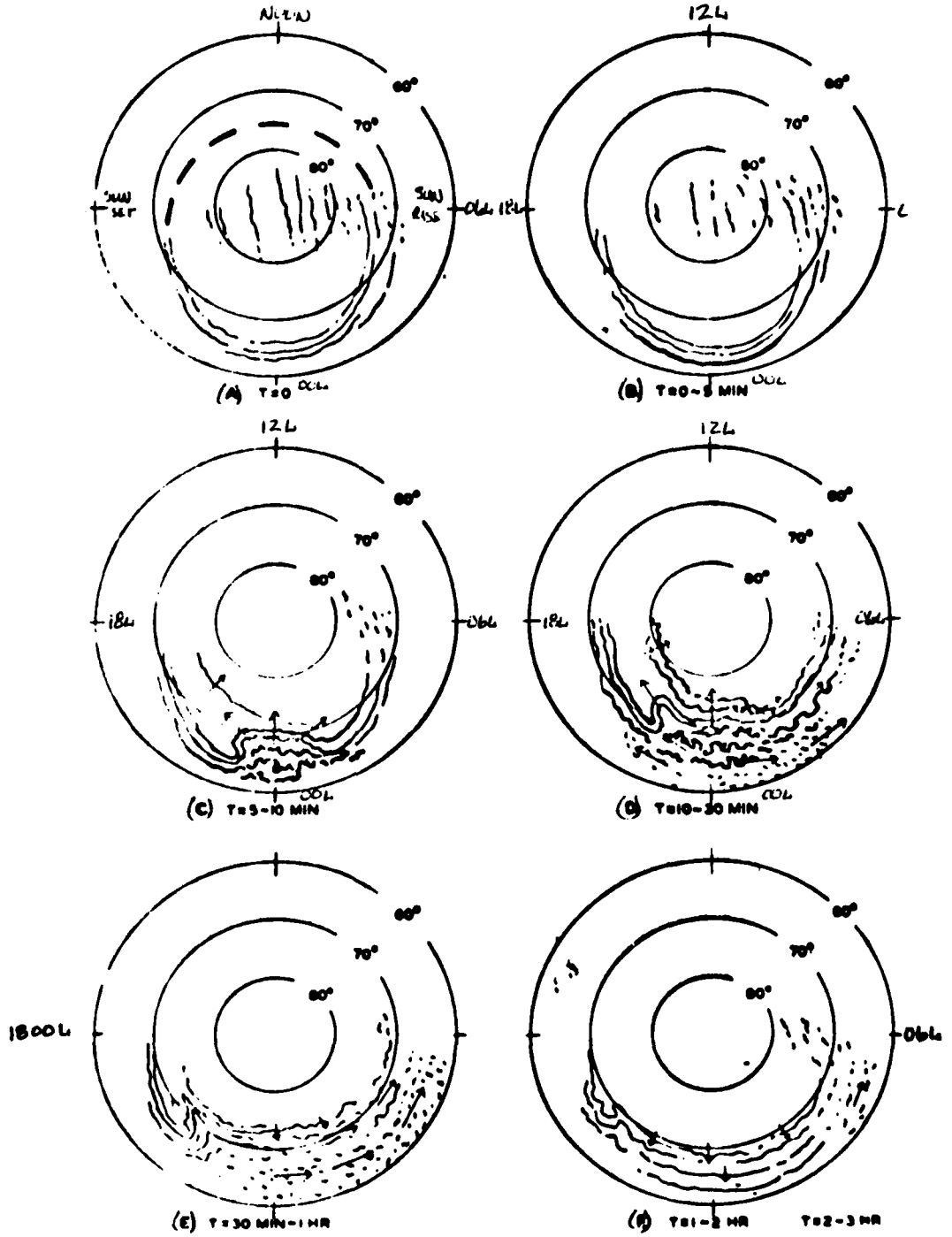



Figure 10.17 Geomagnetic Substorm Development (from Akasofu, 1968).

When the expanding bulge attains its highest latitude, the recovery phase begins. The bulge begins to contract. The auroral forms and locations slowly return to normal. From onset of the expansive phase to the end of the recovery phase generally takes two to three hours. Typically found in the midnight to dawn sector, substorms seem to occur at the rate of 1-3 per day during quiet times.

## 10.8 Geomagnetic Activity Indices

The degree of magnetic disturbance during each Greenwich day is indicated by a variety of indices according to international conventions. Many magnetic observatories throughout the world take part in these observations. They assign local indices to magnetic activity and pass their indices to the Permanent Service of Geomagnetic Indices at Gottingen, Germany. There, the local indices are used as the basis for calculating world indices.

Several indices are used to describe magnetic disturbances and their relationship to solar and geophysical phenomena. The differences in the character and intensity of the variations with latitude influence the choice and derivation of activity indices. In auroral zones, measures derived over periods of an hour or less are important. In middle latitudes, the indices for intervals of 3 hours to 1 day are most used. Near the equator, indices based on an interval of a day or more appear more suitable.

The K index (adopted 1939) is a measure of the irregular variations of standard magnetograms and is used as an indicator of the general level of magnetic activity caused by the solar wind. Each observatory computes a value of K for each 3-hour interval from the largest 3-hour range (R) in X, Y, Z, or H, D, and Z. R is the range between the highest and lowest deviations of the sensor from the quiet day curve. Since observatories in or near the auroral zone record much larger variations than others, the correspondence between K and R is adjusted for each observatory to permit comparison of K values from different stations. A two-letter subscript (e.g., $K_{SI}$ for Sitka) is frequently used to identify the stations. K values are integers ranging from 0 through 9 and are quasi-logarithmic. The K index from Fredericksburg, Virginia, ($K_{FR}$) has been taken as a standard measure of geomagnetic activity for the continental United States.

The $K_p$ index ("p" for "planetary") was intended to measure the worldwide average level of activity and is in widespread use. Its correlation with solar-wind parameters and specific types of magnetic activity varies widely. It is often a good measure of certain auroral-zone activity but a poorer measure of polar-cap disturbance. It is based on the K indices from twelve selected stations between geomagnetic latitudes 48° and 63°. Values of K for a given station are first used to compute the $K_s$ index ("s" for "standardized"), from tables which reflect the characteristic seasonal behavior at the station and thereby remove local variations. The $K_s$ index ranges continuously from 0.0 to 9.0 but is quoted in thirds of an integer, using the three symbols, -, 0, and +. The interval from 3.5 to 4.5 includes the $K_s$ values 4-, 4o, and 4+, and other intervals are correspondingly defined. So $K_s$ can assume 28 values: 0o, 0+, 1-, 1o, ...8+, 9-, and 9o. The $K_p$ value for each 3-hour interval is derived using the $K_s$ values from the 12 stations.

Because the 3-hourly K, $K_s$ and $K_p$ indices are defined with a quasi-logarithmic scale, they are not suitable for simple averaging to obtain a daily index (still, this is not uncommonly done despite the illogic). To convert to a roughly linear scale (that is, reconversion to an equivalent range), the $a_k$index is defined from K, and the $a_p$index from $K_p$ as indicated in Table 10.1. The average value of $a_k$ or $a_p$ over the eight 3-hour intervals in a day is then defined to be the daily $A_k$index or $A_p$index, respectively (single-station or planetary).

The $a_k$ and $A_k$ indices are often quoted in units of gammas (range of field strength) by multiplying by a calibration factor f for the particluar station. Here $f = R_9/250$, where $R_9$ is the lower limit of R for K=9. Thus, at Fredericksburg, f=(500 gammas)/250 = 2 gammas. As an example, if R = 27, then K = 3, and $a_k$ = 15, or $a_k$ = 15 X 2 gammas = 30 gammas. Begun in 1951, the indices $a_k$, $A_k$, $a_p$, and $A_p$ are currently in wide use.

Although several other indices are calculated, the last ones we will consider are the real-time "$a_p$" and "$A_p$" calculated at AFGWC. Most geomagnetic data is reported once per day showing the K value for each 3 hourly period. AWS stations report in real-time (every 90 minutes). The data are reported to AFGWC in chart divisons as read from the magnetometer chart recorder. The computer automatically decodes these reports, converts chart divisions to gammas, converts gammas to K indices and $a_k$, and calculates $a_p$ and $A_p$ using all non-polar stations. Thule is not used in the calculation of $a_p$ and $A_p$, because it is negatively correlated with auroral activity (i.e. during quiet times, the auroral oval shrinks towards the polar region, leading to higher gamma values at Thule). The geomagnetic indices calculated every 90 minutes are transmitted to NORAD and Sunnyvale AFS, where they are used to update models of neutral atmospheric density.

## 10.9 Geomagnetic Activity Forecasting

Extensive study of long period records of geomagnetic data has shown that geomagnetic disturbances have a tendency to recur at 27-day intervals. These "recurrent storms" have been attributed to passage through the current sheet or the occurrence of relatively high speed solar wind streams. Numerous geomagnetic storms do not show a recurrent pattern, but rather are associated with major solar flares or other solar events. We will first look at prediction techniques for recurrent storms and then consider solar event associated geomagnetic disturbances.

The forecasting of recurrent geomagnetic storms resembles persistence forecasting. Current sheet crossings tend to persist for several solar rotations, so a recurrence forecast is possible. If a non-flare associated storm occurs, it may be a good idea to forecast its recurrence 27 days later.

$a_k$ and $a_p$ for Given Values of K and $K_p$

| K | $a_K$ | K | $a_K$ | $K_p$ | $a_p$ | $K_p$ | $a_p$ | $K_p$ | $a_p$ | $K_p$ | $a_p$ | $K_p$ | $a_p$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 6 | 80 | 0o | 0 | 2o | 7 | 4o | 27 | 6o | 80 | 8o | 207 |
| 1 | 3 | 7 | 140 | 0+ | 2 | 2+ | 9 | 4+ | 32 | 6+ | 94 | 8+ | 236 |
| 2 | 7 | 8 | 240 | 1- | 3 | 3- | 12 | 5- | 39 | 7- | 111 | 9- | 300 |
| 3 | 15 | 9 | 400 | 1o | 4 | 3o | 15 | 5o | 48 | 7o | 132 | 9o | 400 |
| 4 | 27 | | | 1+ | 5 | 3+ | 18 | 5+ | 56 | 7+ | 154 | | |
| 5 | 48 | | | 2- | 6 | 4- | 22 | 6- | 67 | 8- | 179 | | |

Table 10.1 Geomagnetic Indices Conversion.

The forecasting of the first occurrence of a "recurrent" storm is much more difficult, because the relationship between high speed wind streams and geomagnetic disturbances is not perfect. It requires accurately forecasting a particular current sheet configuration. In other words, "apparently" similar solar wind conditions do not produce the same geomagnetic results. In most cases, this is probably a consequence of prevailing current sheet geometry. The current sheet has its maximum inclination to the solar equator near sunspot maximum. Current sheet crossings are, therefore, rarer. The likelihood of being on the "wrong side" of the sheet is similarly increased compared to solar minimum. At minimum, we are closer to the sheet, and it has fewer transient distortions. In order to use current sheet crossings as an aid to geomagnetic forecasting we must (1) be able to locate the solar position of the current sheet, (2) understand the time relationship between solar position and earth passage, (3) consider the flare and active region effects on the sheet geometry, and (4) determine the heliographic latitude of the earth.

Solar wind data, when available with sufficient regularity, can provide important short range forecasting information. A large directional change in the solar wind (often as much as $45^{o}$ east or west) will often precede the crossing by 12 to 24 hours, with a shift of at least equal magnitude and opposite direction occurring some 3 to 12 hours prior to passage. The solar wind speed is normally at a minimum just prior to passage (since the sheet is imbedded in low speed wind), and it jumps to a maximum shortly after the crossing. Likewise, a sharp jump in solar wind density and IMF direction reversal is typically associated with the crossing. Constant fluctuation of these parameters may mean that the earth is moving nearly parallel to or in "contact with" the current sheet.

Flares and active regions can affect the timing and severity of recurrent geomagnetic storms. A large flare ahead (west) of a warpage (previously termed SSB) often delays and somewhat obscures the SSB-associated disturbance. A large flare east of the warpage may cause passage to occur earlier than otherwise expected. In addition, the flare east of SSB may

cause the geomagnetic disturbance to persist longer and with a generally lower magnitude than would normally be expected for either a flare or sector disturbance. If you think of the flare blast wave and plasma cloud distorting or reshaping the current sheet, these observations will make more sense. The current sheet is not isotropic; nor is it homogeneous. Waves may exist radially away from the sun and circumferentially about the sun (think of a ballerina's fluted skirt bouncing and waving as she dances). The waves may move and change with time. If a very large solar region(s) is, or recently (in the last several rotations) was located just east of a SSB, the warpage and crossing may occur as much as 2 days later (i.e., 7 days after CM passage of the feature) than would normally be expected. (Contrast with flares which imply open field lines and a high speed stream. A strong sunspot group will be associated with strong, closed magnetic field lines and low speed/low density wind streams.) The geomagnetic storm associated with the crossing can be either a sudden commencement storm or a gradual commencement storm. If a sector boundary disturbance is not flare enhanced, the resulting geomagnetic storm will seldom exceed an $A_p$ of 35. Very narrow "sectors" are common near solar maximum and near the solstices and often have little or no effect on the geomagnetic field.

In the past few years, low latitude coronal holes have been identified as the origin of some high-speed solar wind streams and their associated recurrent geomagnetic disturbances. Studies done during the last (early 1970's) solar minimum show an excellent correlation between coronal holes and geomagnetic storms. Coronal holes can be observed in helium (He I) 10,830A spectroheliograms and in EUV and x-ray photographs taken from outside the earth's atmosphere. Currently, there are studies underway using east-west solar radio scans and Iron (Fe) 5303 A (green line) solar limb scans to detect coronal holes. The central meridian passage of a coronal hole occurs 2-3 days before the wind from the hole would reach the earth. Nearby flares may significantly alter this timing as may current sheet geometry. During solar maximum, the interplanetary magnetic field is much more complicated. If the hole is associated with the wrong current sheet geometry a disturbance is unlikely. Finally, small equatorial holes associated with solar maximum seem unrelated to geomagnetic activity.

Solar flares release many protons and electrons of lower energy than those which produce PCAs. These slow protons (400 to 700 kilometers per second or slightly higher) are associated with a shock wave in interplanetary space. This shock wave is the leading edge of a more dense (maybe 10-20 $cm^{-3}$) plasma region. The magnetic field strength in the shock region is usually higher than that of the ambient field. This more dense (flare enhanced) solar wind can cause a geomagnetic storm. The forecasting of solar flare associated geomagnetic disturbances involves three main considerations: flare characteristics, flare location, and storm timing.

Flare characteristics favorable for producing geomagnetic storm particles include those indicative of a PCA event. The characteristics are indicators that the flare may have expelled particles at all energy ranges. The Comprehensive Flare Index (CFI) (Dodson and Hedeman, 1971) is one way of summarizing what's important. This work suggests that geomagnetic storms are produced only by "significant" flares defined as having one or more of the following characteristics:

(1) Sudden Ionospheric Disturbance (SID) of importance 3 or greater;

(2) Hydrogen-alpha flare size of 3 or greater;

(3) Associated 10.7 cm radio burst of at least 500 SFU;

(4) A type II sweep burst; or

(5) A type IV sweep burst of at least 10 minutes duration. Moreover, the CFI must be greater than 10. For CFI=6 to 10, a small storm, or no storm will result. For a CFI of less than 6, no storm will result. The CFI is calculated as follows:

CFI = SID importance (0-3) + H-alpha flare size (0-3) + characteristic of the log of the 10 cm flux (=3 for 1000 to 9999 SFU burst, for example) + characteristic of log of the 200 MHz burst flux + radio burst dynamic spectrum (type II = 1, type IV = 2, type IV greater than 10 minutes = 3).

While the index is based on a fairly limited data set and is somewhat subjective in nature, it does provide a suggestion of what parameters to look at when analyzing a given flare. One might, for example, replace the SID importance with some parameter related to the size and duration of the 1-8A x-ray burst. This is, after all, what the CFI is really attempting to measure--the energy output spectrum of the flare at x-ray, optical, and radio frequencies. A grocery list of magnetic disturbance indicators looks very similar to those found with many major flares.

In general, optical parameters associated with storm production include:

(a) The flare occurred in a magnetically complex sunspot group;
(b) Parallel ribbon flare (loop prominence system);
(c) 2B or greater flare;
(d) Over 30% coverage of the main spot umbra by the flaring material;
(e) White light flare.

Radio indicators include:

(a) The region had a metric wavelength noise storm (type I) in progress prior to the flare;
(b) Castelli U burst;
(c) Power spectrum increasing toward lower frequencies;
(d) Type II burst followed by a type IV noise burst;
(e) Type IV burst lasts longer than 15 minutes.

Note that these criteria include other than the PCA producing flare indicators. Many flares release sufficient lower energy particles to produce a geomagnetic storm without producing the high energy particles necessary to cause a PCA. Typically, these flares lack the large discrete frequency radio bursts. Likewise, many small, radio-rich (not necessarily large) flares seem to effectively combine their efforts to produce a dense plasma stream (often containing several shocks). If a region produces several such events prior to passing 40°W, it may produce a disturbance. Of all types of radio, low discrete frequencies (i.e., 245-410mhz) and type IV sweep bursts seem to be the most critical for plasma injection into the interplanetary medium.

The direction of the photospheric magnetic field at the flare site is also suggestive of the potential geomagnetic effects of the flare. A southward component seems to result in a higher level of geomagnetic activity than is the case for the northerly component (Lundstedt, et. al., 1981). A southward component would have a plus (outward) field north of the flare and a negative (inward) field south of the flare.

The location of the flare is also an important consideration in magnetic storm forecasting. The most important location consideration is the flare central meridian distance. Unlike the high energy, low density PCA-producing particles which follow the spiral magnetic field lines through the interplanetary medium, the dense, low energy particle cloud carries the solar field lines out into interplanetary space. (It dominates the prevailing IMF structure and can reshape it. This can set the stage for a later PCA event from what would seem to be an unfavorable longitude.) As with the ambient solar wind, this plasma flows nearly radially (equivalent to a typical solar wind particle). Therefore, a flare which occurs within $45^{o}$ of central meridian has a higher probability of producing a geomagnetic disturbance than does one closer to either limb.

The heliographic latitude of the flare is also important. The closer this matches the $B_o$ angle of the sun, the larger should be the geomagnetic disturbance produced by the flare. As seen from the earth, the solar equator is north of the ecliptic plane from December to June and south for the remainder of the year. Statistically, it is known that flares in the northern hemisphere of the sun have produced more and larger geomagnetic disturbances than flares in the southern hemisphere, but no physical mechanism for this effect has been determined. Geometrically, the storms should be evenly distributed in latitude, but separately peaked in time. Northern hemisphere flares should be more effective storm producers in the fall and vice versa. These statistics are somewhat contaminated by the ease (or lack of it) of establishing a southward solar magnetic field at the flare site in each hemisphere in successive magnetic cycles.

The onset of a flare produced geomagnetic storm varies between one and 3 days after the flare. The most probable delay time between flare and onset is 36 hours (equivalent to a shock moving at over 1000 kilometers per second). The onset of the storm may be delayed for flares further from central meridian. Flares which occur east of central meridian tend to produce longer lasting and more intense geomagnetic storms than do flares west of it. Flares east of central meridian eject a plasma cloud which acquires a slight tangential velocity component toward the earth (due to solar rotation--producing a sort of shock), while those west of CM acquire a component away from the earth (lessens the net component toward earth). Thus, the geomagnetic storm from a flare east of central meridian should be more intense than one from a flare west of it.

Flares which produce a number of moderate energy events or a single high energy particle event show a tendency to produce a plasma enhancement which persists for one or more solar rotations. These "recurrent" storms differ from normal recurrent disturbances or high speed streams in three main aspects:

(a) They generally persist for only one rotation after the flare;

(b) Particle energies, while low, are generally higher than those associated with true recurrent particle streams; (a 5 MeV proton event may precede or accompany the flare-produced recurrent geomagnetic disturbance); and

(c) In several cases, they have reached the earth several days earlier than a true recurrent stream.

The prominence activity known as an EPL, an erupting prominence on the limb, ejects low speed (but above escape velocity) plasma into the solar wind. The slower portion of the plasma cloud receives a significant tangential velocity from the rotating sun and IMF. For an EPL on the east limb, the chance of the plasma reaching the earth is greater than for one on the west limb. Such a disturbance will typically begin when the location of the EPL reaches 10° to 30° W.

On the disk, an erupting prominence (filament may show as a disappearing filament (DSF)) may inject storm plasma into the IMF. A DSF in which the material has a marked blue shift (toward earth) has a better chance of being an erupting filament than one with a red shift. The timing of a DSF geomagnetic disturbance onset is much the same as the timing of a flare associated disturbance. Just as with the solar flare, a southward photospheric magnetic field component (+ field north of the filament) at the filament site seems to be conducive to greater geomagnetic activity (Joselyn and McIntosh, 1981).

Active surge regions (ASR) occasionally produce brief sprays which eject plasma. Like the EPL, an east limb ASR may then produce a geomagnetic disturbance as the location where it occurred passes approximately 30° west. These disturbances are not usually large or long-lived unless several flares or low frequency radio bursts have also occurred in this region while on the disk.

Decay time of a magnetic disturbance is also difficult, but not impossible to forecast. Assuming a storm is not augmented by additional flare activity, statistical data can be used to good effect. Tables similar to persistence probability have been compiled at AFGWC and provide reasonable guidance on the duration of minor ($A_p$ greater than 30) and major ($A_p$ greater than 50) disturbances. Qualitatively, it is possible to make several generalizations:

(1) Larger storms decay more rapidly that smaller storms, probably because the associated ring current penetrates more deeply into the atmosphere and is thus more quickly absorbed when the influx of storm energy stops.

(2) Solar minimum storms decay more rapidly. This is probably due to the entire atmosphere being initially cooler (lower level of solar radiation) and so more able to absorb larger amounts of heat quickly.

(3) More and larger disturbances occur near the equinox. Flare storms are larger than recurrent storms, and there is more likelihood of getting the brunt of the storm (assuming it comes from a flare on the earth's side of the current sheet) when the earth is most distant from the current sheet.

(4) The maximum level of geomagnetic activity occurs about two years after sunspot maximum. This is the time when the solar polar magnetic fields

are reversing. Hence, the current sheet is weakest and most confused. This may result in numerous crossing and incomplete blocking of plasma streams from opposite hemispheres.

## 10.10 Aurora

For centuries, mankind has been fascinated by brilliantly colored lights which seemed to dance across the near-polar sky. These brilliant displays sometimes continued for hours, during which the sky seemed ablaze with color, form, and motion. Curiosity about this phemonenon was especially strong in Scandanavia, where early auroral observations were documented. More recently, scientists and amateurs have done much of their auroral observing in Canada and Alaska, with the University of Alaska at College, Alaska (near Fairbanks) being the center for much of this work.

Auroral emissions are produced by precipitating low energy particles. Protons and electrons with kinetic energies between 0.3 and 20 keV (average energy 6 keV) produce the optical aurora. These particles spiral down magnetic field lines into the earth's atmosphere where they collide with neutral (uncharged) atmospheric particles and give up energy to ionize or excite them. Collisions are most likely at about 100 km. The primary atmospheric particles at this height are atomic oxygen and molecular nitrogen. These are the particles most commonly ionized or excited by the collisions. These particles store the energy briefly and then emit it as electromagnetic waves (light). The most important visible emissions are the green (5577A) and red (6300A and 6364A) emission lines of atomic oxygen and various blue and red molecular nitrogen bands. Other significant auroral emissions occur across the electromagnetic spectrum from radio waves to x-rays. The optical emissions range in intensity (as observed by human eye from the ground) from below visual threshold to roughly equivalent to full moonlight.

There are two primary source regions for auroral particles: the plasma sheet and the magnetosheath/polar cusp. We previously examined the plasma sheet and found it a region of dense plasma extending from about 10 $R_E$ on the anti-solar side of the earth back into the tail. The magnetic field lines through it connect to the earth and channel the particles earthward along horn shaped appendages into the auroral oval. These are the locations of precipitating auroral particles on their way from the plasma sheet to the atmosphere. The magnetospheric field lines which run along the magnetopause enter the magnetosphere along the cusp and reach lower altitudes on the sunward side. These magnetic field lines connect into the earth in an oval shape, broader and further equatorward on the anti-solar side (see Figure 10.18).

### 10.10.1 Auroral Location

The auroral oval is the intersection of the magnetic field lines on which the precipitating particles are trapped with the earth's atmosphere at a level of atmospheric density high enough to make the collision probability high. The auroral oval roughly coincides with the curve of intersection of the trapping region of the magnetosphere with the ionosphere. The geomagnetic field lines which intersect the earth poleward of the auroral oval are those which are swept back by the solar wind to form the magnetotail. Those

equatorward of the oval are closed, generally undisturbed field lines which encompass the plasmasphere. The auroral oval is the instantaneous location of the aurora.

The size of the auroral oval is not fixed. During very quiet periods the oval shrinks, contracting toward the poles. Its midnight dipole latitude can then be near 70° or higher. During a great magnetospheric storm, both the inner and outer limits of the auroral oval shift equatorward. The equatorward boundary can reach 50° (dipole) latitude, or even lower, during very intense storms. At such times, the polar cap also expands, and the geographic region which normally experiences the most frequent appearance of the aurora (65-67° geomagnetic in the midnight sector) may be poleward of the auroral oval. When a typical auroral station is located poleward of the enlarged oval, it becomes, temporarily, a polar cap station. The mean position of the maximum of the aurora is about 23° from the magnetic pole on the nightside and 15° from the pole on the dayside. These distances are approximately equal in both hemispheres.

The size of the auroral oval can be correlated to various geomagnetic indices. Two indices commonly correlated to the size of the auroral oval are $K_p$ and Q. Kp was correlated to show that the geomagnetic latitude of auroral occurrence moves equatorward as the $K_p$ value increases. A decrease in $K_p$ of 1 results in a shrinkage of the oval of about 2° poleward. As Q increases, the auroral oval shifts equatorward.

Aurora is observed with greatest frequency in a circular band about each pole. The bands are not completely symmetric from hemisphere to hemisphere, but are centered 20°-25° from each magnetic pole. This band of maximum occurrence is known as the auroral zone. It is, very loosely, the locus swept out by the midnight sector of the auroral oval. The auroral zone is of particular signficance in ionospheric analysis and is discussed in this context elsewhere.

### 10.10.2 Auroral Appearance

Aurora visible to the unaided eye has a curtain-like structure and a fairly definite lower border at approximately 100 kilometers. It is uncertain if a single curtain ever extends completely around the oval, but east-west lengths of at least several thousand kilometers have been observed. The north-south dimension of a single curtain is very small, only a few hundred or even tens of meters.

The complexity of the auroral form serves as an indicator of the activity of the aurora. The simplest type of auroral curtain appears as an arch or arc of nearly uniform brightness. From above, the arc may appear as a thin line. More active aurorae develop apparent vertical striations (rays) and are called rayed arcs. Seen from above, a rayed arc reveals the rays to be very fine "pleats" in the auroral curtain. As the aurora becomes more active, it develops waves, folds, or curls of various scales. A very active rayed band with a well-developed fold (maybe 100 kilometers long) is often called a drapery. The pleats and folds are aligned along magnetic field lines.

The auroras are divided into discrete and diffuse. A discrete aurora appears as a single, bright strand of emission separated from other discrete

auroras by a dark space a few tens of kilometers in width. When seen from the ground, it has a curtain-like structure. A diffuse aurora appears as a broad band of emission with a width of at least several tens of kilometers. It may not be easily visible from the ground, but can cover half of the sky. It generally fills the auroral oval and can be used to define the instantaneous position of the oval if all-sky or satellite photos are available. It probably results from lower densities of lower energy auroral particles than are required to produce discrete aurora.

Geomagnetic storms and substorms both produce aurora. Auroral substorms occur during all geomagnetic conditions but are most common during geomagnetic disturbances. Geomagnetic storms are always accompanied by auroral substorms, especially during the main phase. Auroral activity and geomagnetic substorms are both correlated to the auroral electrojet. In the diffuse auroral region, an eastward auroral electrojet commonly occurs in the evening sector. During a substorm, the morning sector contains westward auroral electrojet currents.

AFGWC uses auroral observations from three instruments: satellite imagery, precipitating particle data, and auroral radar returns. Auroral position is inferred using magnetometer observations, solar wind data, and ionospheric analysis (ionosondes, polarimeters, and riometers).

Auroral imagery is provided by spacecraft. These spacecraft are in low, near-polar orbits, and each vehicle completes its orbit in approximately 90 minutes. The spacecraft are sun-synchronous, so they always cross the equator on the sunward side of their orbits at the same local time, and the anti-sunward side at another specific local time. A noon-midnight satellite, for example, crosses the equator at noon on one side and midnight on the other on every orbit. The imagery is the product of a visual (actually visible-infrared) sensor. The sensor continually scans a thin region across the path of the spacecraft and builds a "picture". For auroral observations, the sensor is set for high gain (sensitivity). This sensor gives its best observations over the winter pole near midnight, with a new moon (i.e. the darkest possible conditions).

Precipitating particle measurements are made by a device known as the J sensor. Particles with the proper energy to produce aurora are measured in counts per unit time. When the satellite moves from the magnetic field lines of the trapped radiation belts to those of the auroral region, this sensor records a sudden upward jump. Figure 10.19 shows the result of a typical pass accross the south polar cap. A computer program at AFGWC locates this jump and computes where the sampled particles of the jump will produce the equatorward edge of the diffuse aurora. This location of the auroral oval edge is used to compute an equivalent Q index, $Q_e$.

The High Latitude Monitoring System (HLMS) includes an auroral radar located at Anchorage, Alaska which is pointed toward the geomagnetic north pole (near Thule, Greenland). The output of this system is sent to AFGWC where it is displayed as a plot. The existence of large values (strong returns) indicates radar aurora, a phenomenon observed by radar and associated with optical aurora. Auroral radar returns show an increase in intensity as well as an equatorward shift of the most intense region during the expansive phase of a substorm. The recovery phase is marked by a decline in intensity of auroral returns.

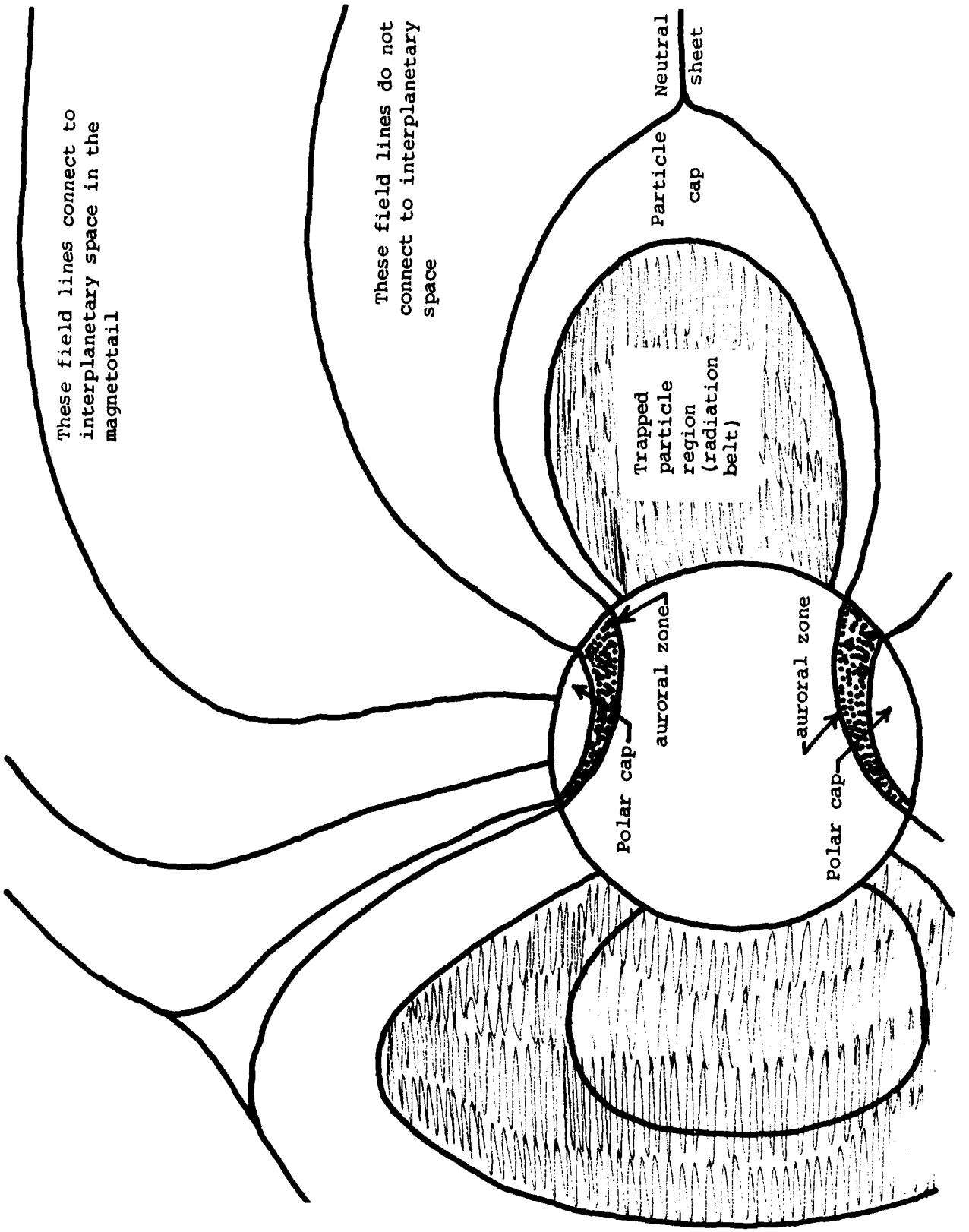



Figure 10.18 Magnetospheric Connection into the Auroral Zones.

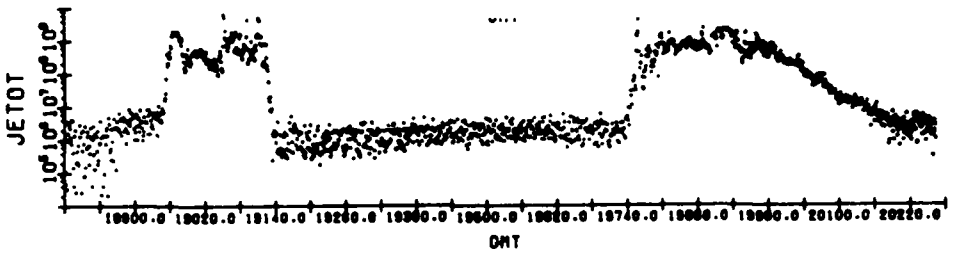



Figure 10.19 Electron Energy Flux (KeV/cm$^2$ sr sec) Versus Magnetic Latitude for a South Polar Pass (After Hardy, et. al., 1981).

The equivalent Q index, $Q_e$, is computed from auroral observations. Larger values of $Q_e$ equate to a diffuse auroral boundary nearer the equator. Since Akasofu's substorm model has an expansive phase, when the auroral locations move equatorward, and a recovery phase, when its locations move poleward, we can relate $Q_e$ changes to substorm phases. The expansive phase is marked by increasing $Q_e$ values and the recovery phase by declining ones.

## 10.11 Summary

The geomagnetic field is, to a good approximation, dipolar. It's interaction with the solar wind produces the magnetosphere, an approximately 10 $R_E$ radius cavity about the earth. Bounded by the magnetopause, the magnetosphere defines the region in which the geomagnetic field controls particle motions. The trapped radiation belts and the auroral zones are partial consequences of this structuring. Variations in the solar wind translate into magnetospheric variations. Depending on their magnitude, these variations produce geomagnetic substorms and may result in increased auroral activity. The magnetosphere is closely tied to and exerts a major influence on the earth's upper atmosphere and the ionosphere.

# CHAPTER 11

# THE IONOSPHERE: FORMATION AND VARIATION

Beginning about 75 km above the earth's surface ionized particles become a significant fraction of the atmosphere. Indeed, it is here, in the region known as the ionosphere, that ions and electrons are first present in sufficient quantities to affect the propagation of radio waves. The ionosphere is born of the interaction between solar radiation and the earth's atmosphere and magnetic field. An understanding of the ionosphere first requires a basic knowledge of the earth's atmospheric structure. Its variability is an important factor in the ever-changing ionosphere.

## 11.1 The Neutral Atmosphere

The atmosphere is in continual motion in response to differential solar heating. Warmer air rises, displaced by descending cool air. This overturning occurs in both vertical and horizontal planes, and is evidence that sunlight is not equally intense at all locations. Atmospheric mixing has numerous consequences in meteorology. Since a few of these are significant even above 75 km, they must be considered. Yet mixing occurs in only a limited segment of the atmosphere. Mixing, differential heating, and the earth's rotation produce a limited stratification of the atmosphere, and it is these strata to which we first turn our attention.

### 11.1.1 Temperature Regimes

Weather, as the term is generally applied, refers to phenomena of the lowest layer of the earth's atmosphere, the troposphere. This layer is characterized by a fairly steady decline in temperature (approximately 6.5 $^{o}$K/km) with altitude. Variable concentrations of water in the atmosphere may cause the actual variation of temperature with height to be quite different from this average.

It was believed, before the beginning of the twentieth century, that the temperature continued to decline up to about 50 km where the atmosphere merged into interplanetary space. However, balloon-borne thermometers revealed a nearly isothermal region beginning near 11 km at mid-latitudes. The temperature of this layer is near 220$^{o}$K. The level at which the temperature profile becomes isothermal is called the tropopause.

The stratosphere lies above the tropopause. In this region, the temperature is initially constant (for the first few kilometers) and then increases with height to the top of the layer. The top of the stratosphere is named the stratopause and occurs at about 45 km.

Above the stratosphere, the mesosphere is characterized by decreasing temperature with height. The top of this layer, known as the mesopause, occurs between 80 and 85 km altitude. The mesopause is the coldest level of the entire atmosphere, with a temperature near 180$^{o}$K.

The layer above the mesopause is the thermosphere. Thermospheric temperature increases vertically (see Figure 11.1). When considering the

thermospheric temperature, it is important to remember that we are speaking of km temperature, not sensible temperature. The thermospheric base near 90 kilometers marks the onset of a temperature inversion and divides the atmosphere into chemical regimes. The top of the thermosphere, the thermopause, marks a return to an isothermal temperature field (Figure 11.5).

### 11.1.2 Chemical Composition Regimes

The temperature gradient establishes four distinct portions of the atmosphere--the homosphere, heterosphere, exosphere, and protonosphere. They encompass the temperature structure outlined above.

The homosphere makes up the lower 100 km of the atmosphere. It includes the troposphere, stratosphere, and mesosphere. Vertical mixing of the atmosphere occurs in the lower atmospheric layers and keeps the relative concentrations of gases nearly constant. The troposphere is composed of (approximately) 78% molecular nitrogen, 21% molecular oxygen, and 1% argon, with variable concentrations of such gases as carbon dioxide and water vapor. The decreasing temperature with height in the troposphere and mesosphere allows convective mixing of gases. The relatively high density of atmospheric particles in the stratosphere and troposphere means frequent collisions between particles, resulting in further mixing of gases. This region is sometimes called the turbosphere because of the turbulent mixing which keeps the densities of molecular nitrogen and oxygen relatively constant. This sameness in chemical composition results in the region's name.

Near the stratopause (about 50 km), there is a small but significant amount of ozone. Above the stratopause, the concentration of molecular oxygen is insufficient, and below it, there is insufficient atomic oxygen to produce a significant amount of ozone. Thus, there is only a small region near the stratopause where there is an appreciable concentration of ozone. This layer is sometimes called the ozonosphere. Ozone absorbs virtually 100% of the incoming solar extreme ultra-violet (EUV) with wavelength less than 2900 A. This radiation would be lethal to many forms of life present on earth (including unprotected man). The ozone layer shields the earth from this radiation and heats the stratopause. Clorofluorocarbons (formerly used as aerosol spray can propellants) cause rapid loss of ozone. When the importance of this gas in destroying the ozone layer became known, such gases were banned from U.S.-made aerosol sprays. High levels of solar activity and high altitude aircraft exhausts also temporarily deplete the ozone layer.

Above the mesopause (near the top of the homosphere's highest level), the temperature increases steadily with height (toward a value dependent on the level of solar activity; see Figure 11.2). This level is also known as the turbopause, because convective mixing ceases above this level. The absolutely stable lapse rate of this region (above the homosphere) eliminates most vertical motion, or convection. Moreover, since total atmospheric density is decreasing with height, fewer collisions occur than in the lower layers. Both of these processes minimize atmospheric mixing and allow diffusion to become important. If you take a container of water and oil and mix it you get a "solution" of oil and water. Let the container sit a few minutes. The oil and water will separate; with the heavier water molecules settling to the bottom, and the lighter oil molecules rising to the top. A similar situation

occurs in the atmosphere above the turbopause. Mixing stops, and the various atmospheric constituents separate out to reach hydrostatic equilibrium separately. The particles diffuse through the region, with heavier particles dominating low in the atmosphere, and successively lighter constituents becoming important higher in the atmosphere (diffusive equilibrium). This layered region extending upwards to 500 km is called the heterosphere (hetero is Greek for different).

The major constituents of the heterosphere are molecular nitrogen ($N_2$), relative weight per particle of 28; molecular oxygen ($O_2$), 32; atomic oxygen (O), 16; argon (Ar), 39; helium (He), 4; and atomic hydrogen (H), 1. Ordered by decreasing weight, they are Ar, $O_2$, $N_2$, O, He, and H. In diffusive equilibrium, we would expect the heavier gases (Ar, $O_2$, and $N_2$, the dominant gases of the troposphere) to be relatively unimportant at higher levels, where the lighter gases, especially atomic hydrogen, tend to gather.

This actually occurs in our atmosphere as shown in Figure 11.3. Atomic oxygen is the dominant constituent from just below 200 km to approximately 550 km altitude. Helium is dominant above that level, and hydrogen dominates even higher.

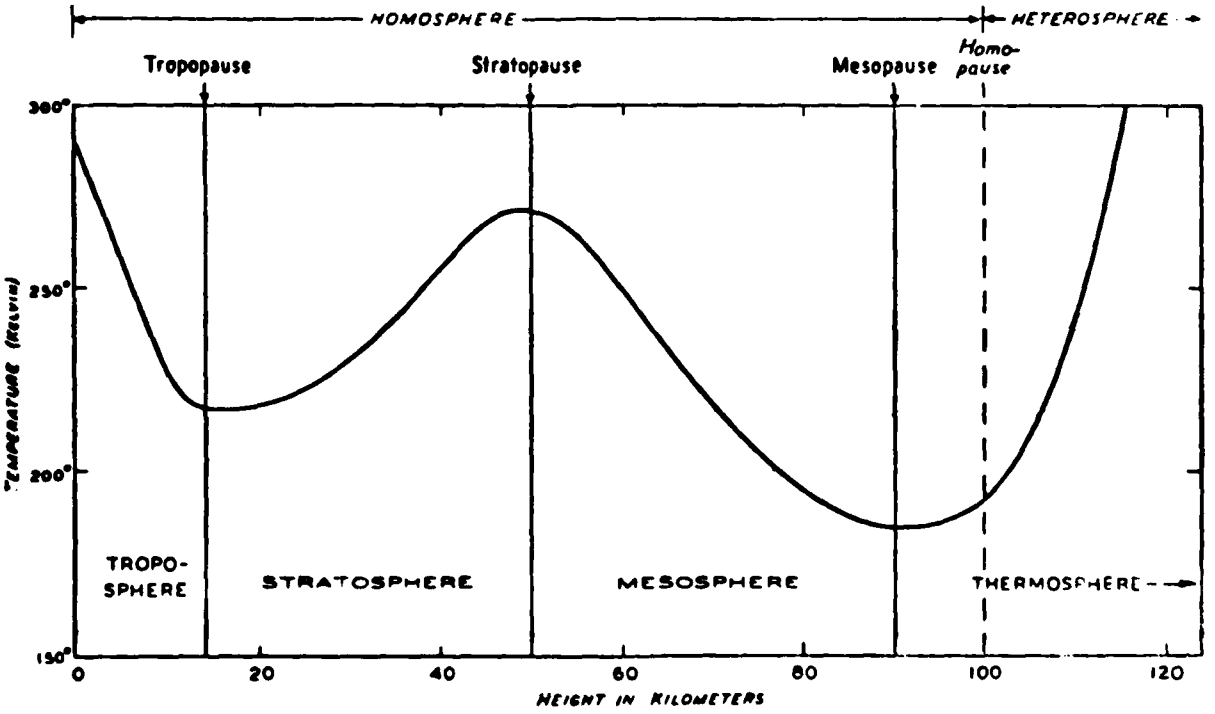



Figure 11.1 Atmospheric Temperature Regimes (from Jacchia, 1975).

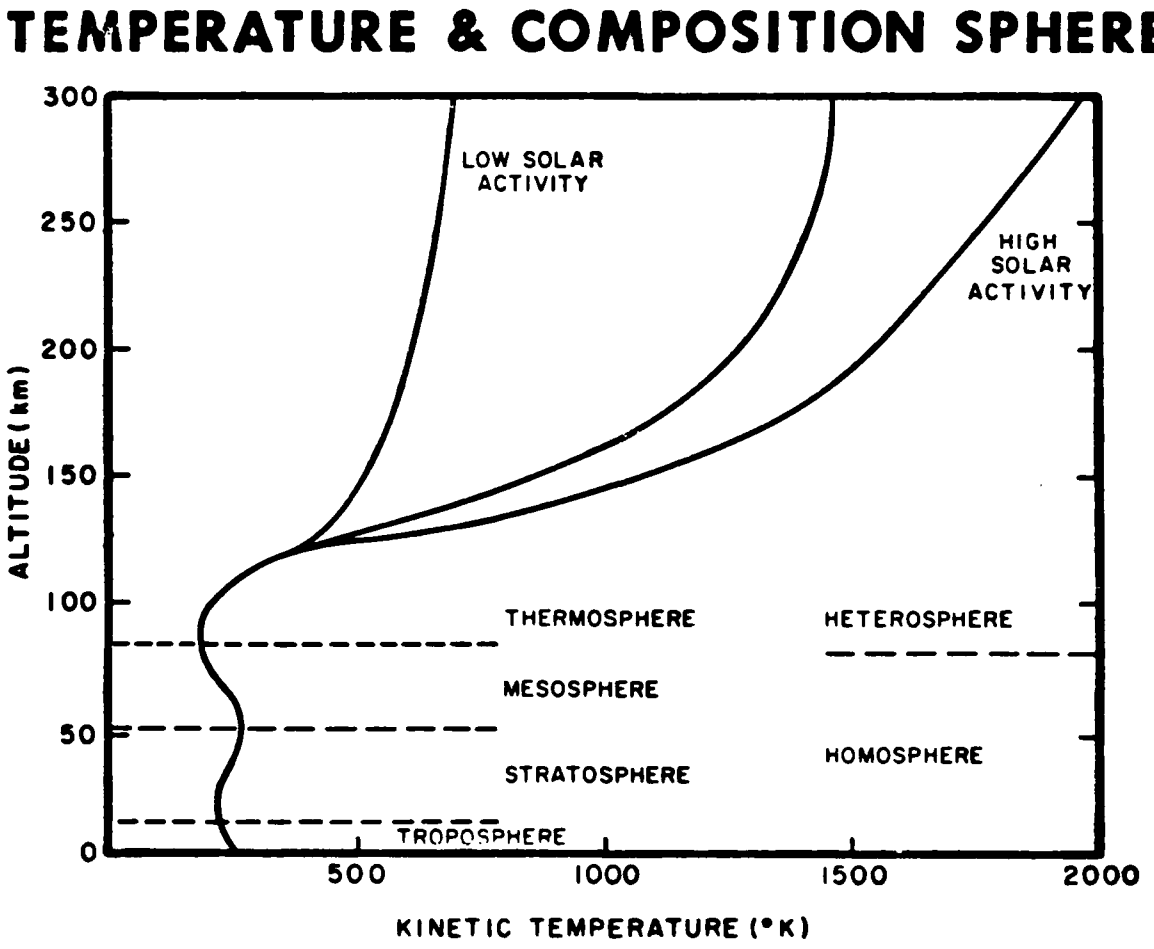



Figure 11.2 Effects of Solar Activity on Atmospheric Temperature Gradient (MITRE, 1972).

The thermosphere, defining the center of the heterosphere, is primarily heated by atomic oxygen, which absorbs EUV radiation of 1000 to 2000 A. Small contributions to the heating come from precipitating charged particles (mostly in the auroral zones) and certain chemical reactions. The energy is converted to thermospheric heat.

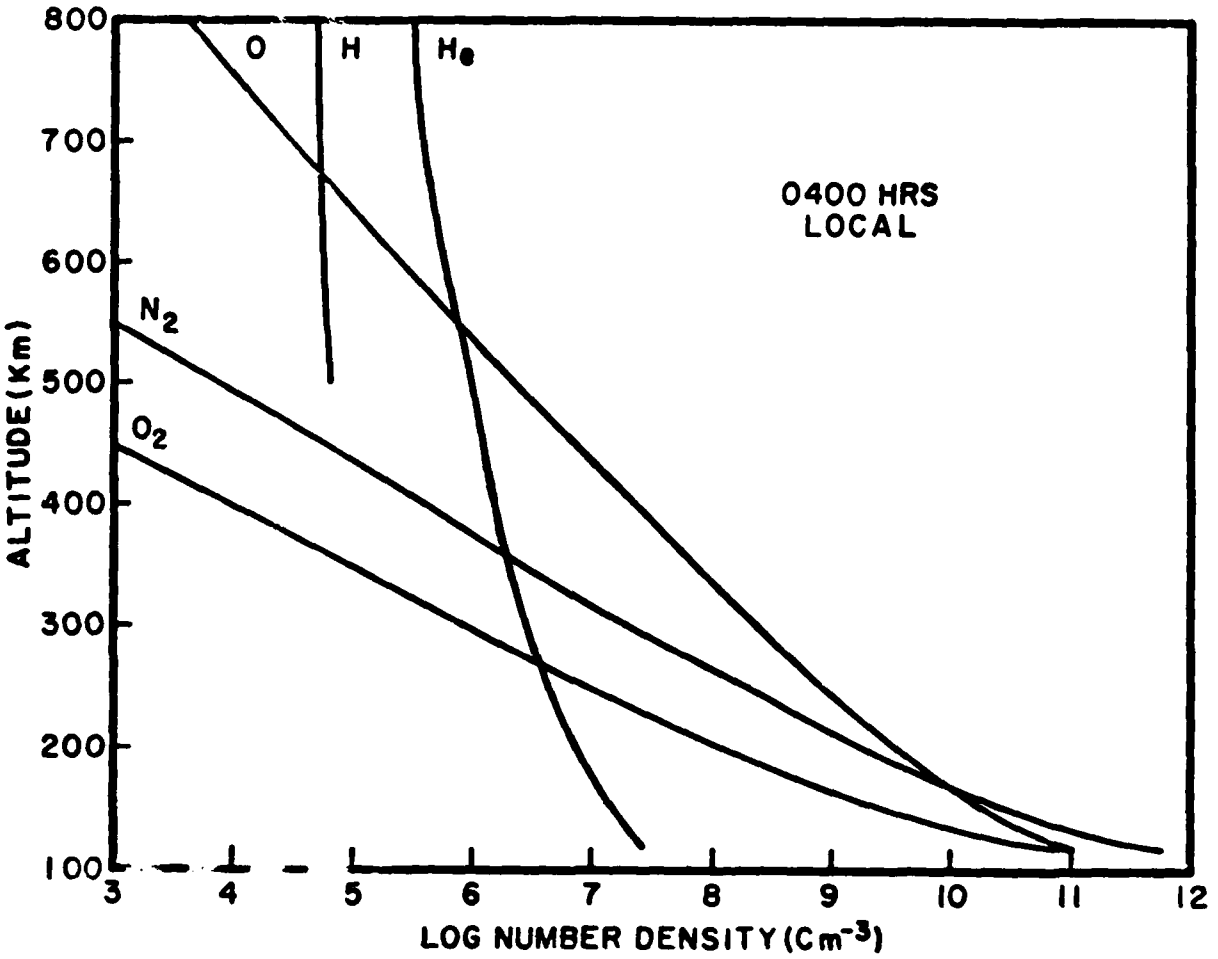



Figure 11.3 Elemental Density Variations with Height (MITRE, 1972).

Above the heterosphere, the exosphere extends upwards to nearly 1000 km. The concentration of a gas in hydrostatic equilibrium is a function of temperature. Figure 11.2 shows temperature change through the thermosphere for low and high solar activity. The highest temperature in the atmosphere is in the isothermal region at the top of the thermosphere and is called the

exospheric temperature. We can assume it to be the temperature of the top of the atmosphere. Hydrostatic equilibrium gives a higher density of any gas at a given level when the temperature is higher. For a gas in which the temperature varies as it does in the thermosphere, the variation may be related to the exospheric temperature. For a higher value of exospheric temperature (higher solar activity), all the gases (except hydrogen) shift to higher densities at any heterospheric level (contrast Figure 11.3 to Figure 11.4). For higher levels of exospheric temperature/solar activity the level where a constituent (like atomic oxygen) becomes dominant occurs higher (from 275 to near 1000 km) than for a lower exospheric temperature/solar activity level (below 200 to 550 km).

The exosphere is a region of continual loss of atmospheric particles due to long mean free paths and high kinetic energies. The very low particle density at high altitude means the average distance between particles is large. This, in turn, leads to the long mean free paths. At 800 kilometers, for example, the mean free path of atmospheric particles is approximately 160 km. A neutral particle in the earth's atmosphere is restrained from escaping by only two things: gravity and collisions. In the upper thermosphere, the large kinetic temperature (high speed) means that a large portion of the particles are traveling with greater than escape velocity (fast enough to overcome gravity). So long as they remain electrically neutral these particles can easily escape unless they are directed toward the earth's surface, collide with another particle and give up energy (slow to below escape velocity), or are deflected toward the earth. The large mean free path of the upper thermosphere makes collision probability very low, so many of these particles escape the earth. Those traveling too slow to escape may later absorb incoming solar EUV radiation and reach a kinetic temperature high enough to escape.

Lighter particles, such as hydrogen and helium, are traveling at a higher speed than heavier particles ($N_2$, $O_2$, Ar) at the same kinetic temperature, and so escape more readily from lower altitudes than do heavier particles. Because of this, there is no fixed lower boundary of the exosphere. It is generally thought to be somewhere between 500 and 1000 km. Hydrogen and helium dominate at this level, and they are steadily lost in the exosphere. Variations of exospheric temperature occur due to such factors as solar activity, time of day, and latitude.

### 11.1.3 Density Variations

Variations in total density as a function of temperature at any height may be constructed using a series of charts like Figure 11.4 for each atmospheric component, and adding all the densities for a given temperature. We can display the results as shown in Figure 11.6. We then consider temperature and density variations interchangeably; to speak of one is to speak of the other. (Note on Figure 11.6 that the density change with exospheric temperature is especially rapid near 500 kilometers). Atmospheric density at a given altitude changes in response to many factors including local time, latitude, altitude, and level of solar and geomagnetic activity.

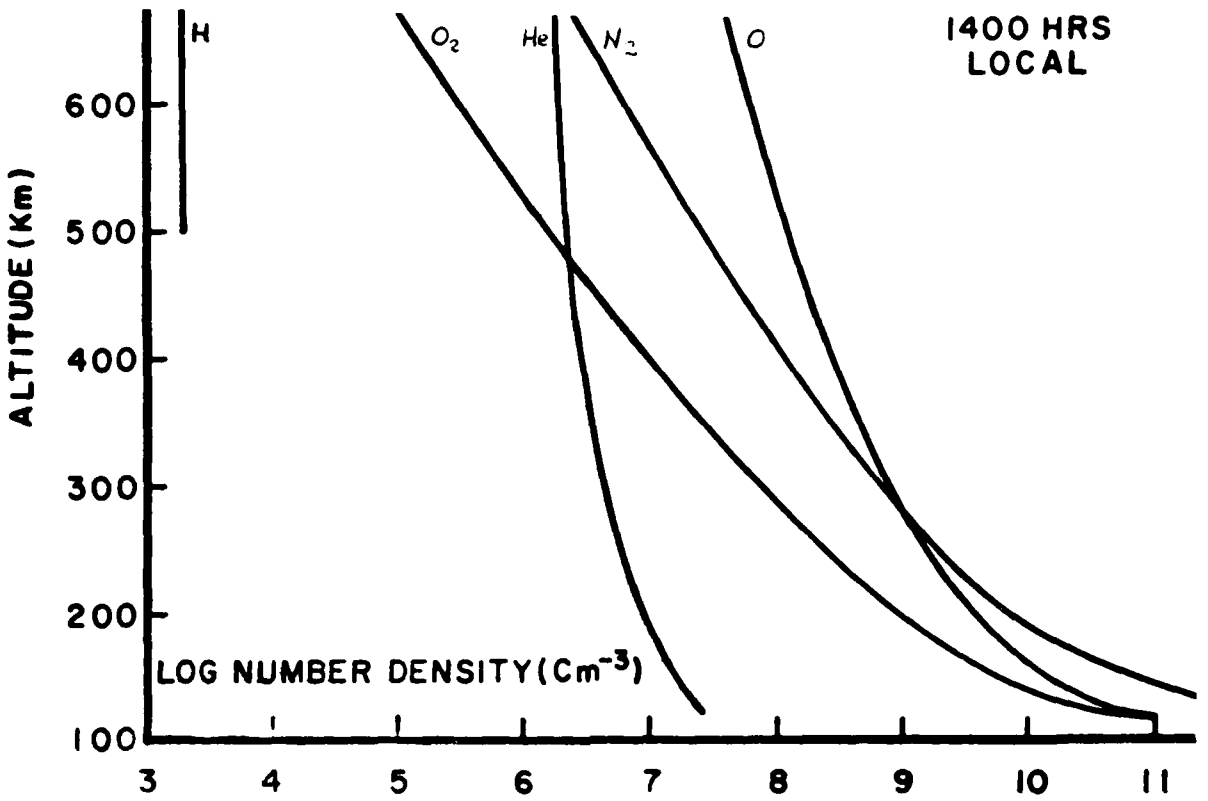



Figure 11.4 Density Variations with Height for High Solar Activity (MITRE, 1972).

The amount of solar radiation received at a point on the earth's surface (insolation) is a function of local time. Similarly, exospheric temperature at a point in the atmosphere is a function of local time, since the exosphere is heated by absorbing solar EUV. Figure 11.7 shows derived density, based on actual satellite orbits at various altitudes. Note that the peak density occurs about 1400 to 1500 local time. Minimum density occurs shortly before sunrise.

Insolation decreases with distance from the subsolar point (the point on the earth's surface directly between the center of the earth and the sun). Figure 11.8 shows this latitudinal effect (combined with the diurnal variation) on computed exospheric temperature for different seasons. The temperature (and thus the density) peak does not simply shift its location with the subsolar point, but changes its intensity from equinox to the June solstice. This also involves the semiannual variation.

Averaging the density over local time and latitude (to remove sun angle effects) reveals that the average worldwide density varies in a cyclic fashion with a semiannual period (see Figure 11.9). The highest average density

occurs in October, with a secondary peak in April. The lowest average density occurs in July, with a secondary minimum in January. This effect is actually the result of two features: the variable earth-sun separation and the higher solar altitude (and longer day) in the summer hemisphere.

The sun's EUV output varies in a pattern similiar to sunspot number (SSN). This variability translates into a variation of energy available to the thermosphere. This results in variation of exospheric temperature, which, in turn, produces a solar cycle variation of atmospheric density. A measurement of EUV flux could give a good estimate of the density, but SESS receives no real-time EUV flux measurements. Such measurements would have to be made by spacecraft.

Little EUV radiation reaches the ground. Ground based solar EUV measurements are, therefore, impossible. Direct EUV flux observations have been made only rarely, but we can infer the value based on 2800 MHz flux measurements. EUV and 2800 MHz fluxes show a fairly good correlation. The 2800 MHz flux is better known as the 10.7 cm flux (or F10). Although the correlation is not perfect (and varies from one sunspot cycle to the next), the patterns are similar enough to be useful.

Daily values of 10.7 cm radio flux are observed near Ottawa, Canada at 1400Z, 1700Z, and 2000Z. These times equate to nearly 0900, 1200, and 1500 local time. There is some atmospheric attenuation of the received flux, dependent on sun angle, with the minimum at local noon. The 17Z (noon) flux value is the world standard and is archived in a manner similar to SSN. These values are used as "measurements" of the solar EUV radiation which heats the atmosphere. Two 10.7 cm flux values are used in most atmospheric density models, the daily value and a longer term mean. The mean (90 day) value is representative of the long term carryover effect of previous days' EUV fluxes. The effects of EUV/10.7 cm flux on atmospheric density are shown in Figure 11.10. Note that 10.7 cm flux has no significant impact on the atmosphere. It is, however, closely correlated with EUV (which strongly impacts the atmosphere).

When a geomagnetic storm occurs, large numbers of charged particles are dumped from the magnetosphere into the high latitude atmosphere. These particles ionize and heat the high latitude atmosphere by collisions, with the heating first observed several hours (2 to 10) after the geomagnetic disturbance begins. The change this geomagnetic storm heating makes on the normal diurnal/latitudinal temperature pattern can be inferred from Figure 11.11. Most heating occurs near 100 km, but changes in density and chemical composition extend from at least 300 km to over 1000 km and may persist for 8-12 hours following the end of the magnetic disturbance.

Density variations may have significant impact on radio wave and spacecraft operations depending on the altitude at which they occur. Increasing particle density between 60-100 km or higher can significantly alter the structure of the earth's ionosphere by providing additional material for ionization.

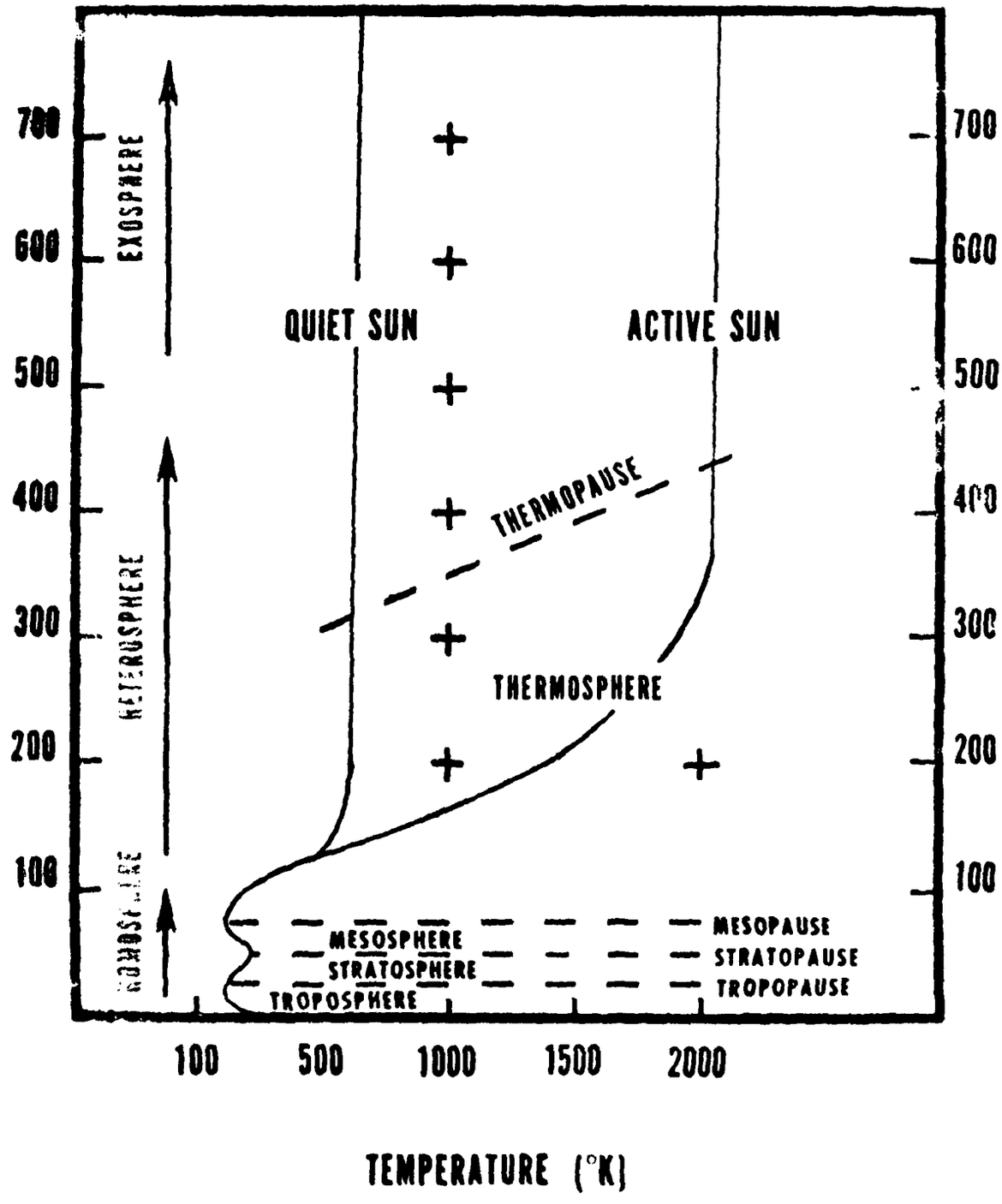



Figure 11.5 Effect of Solar Activity on Exospheric Temperature (MITRE, 1972).

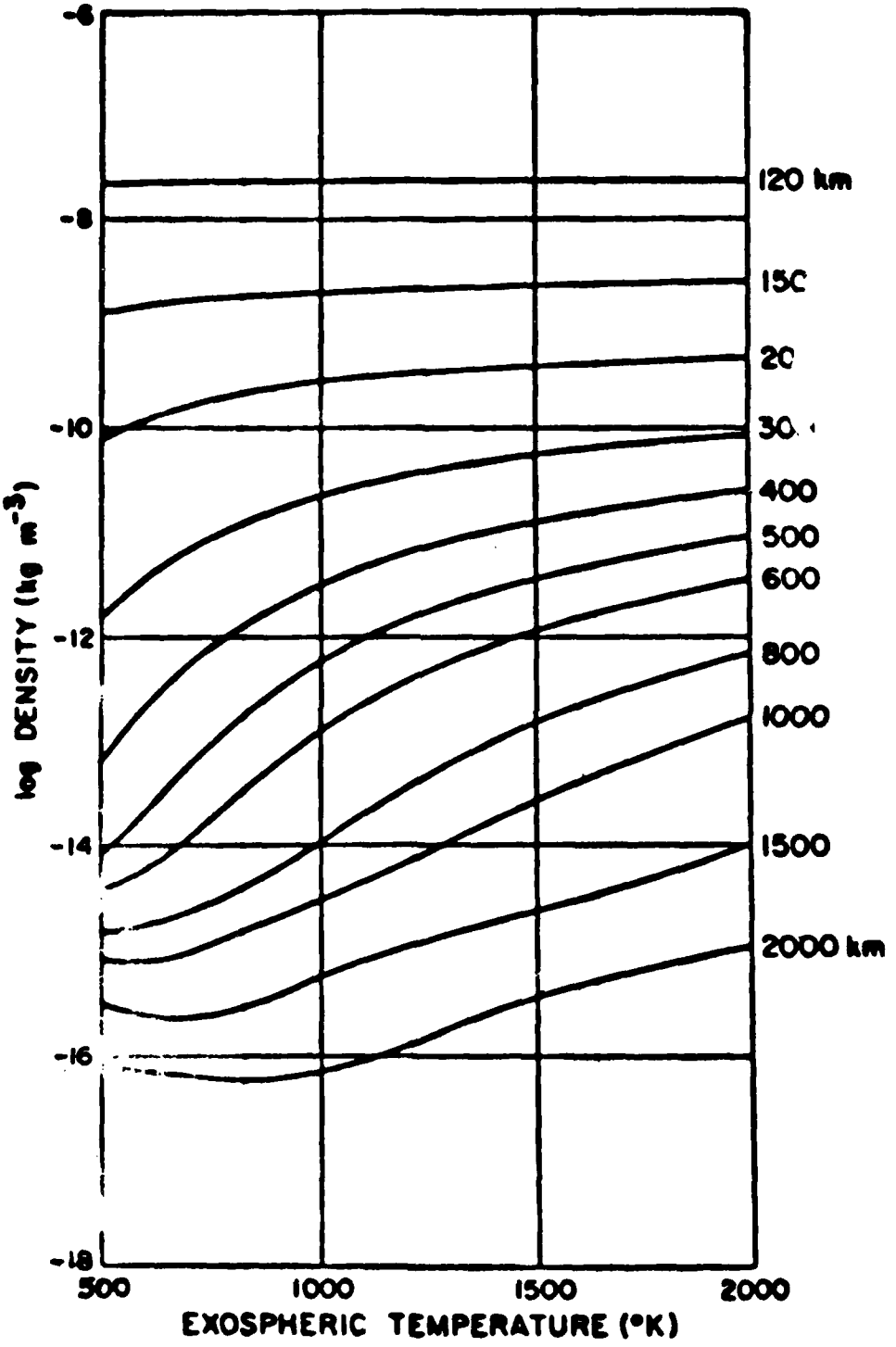



Figure 11.6 Atmospheric Density Variations with Altitude.

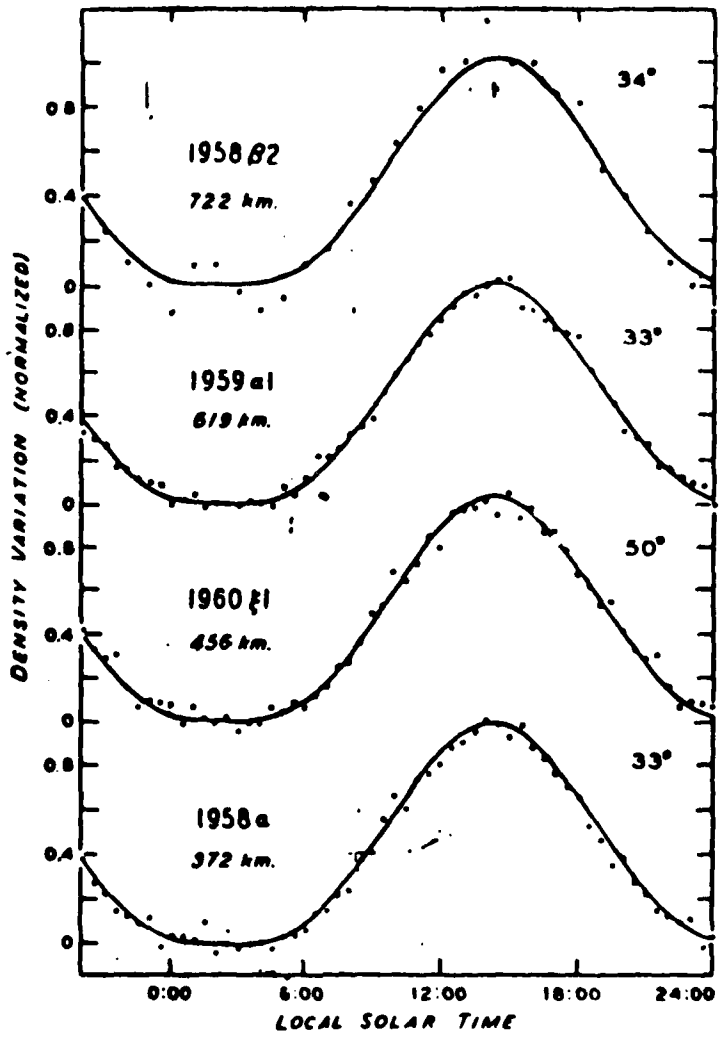
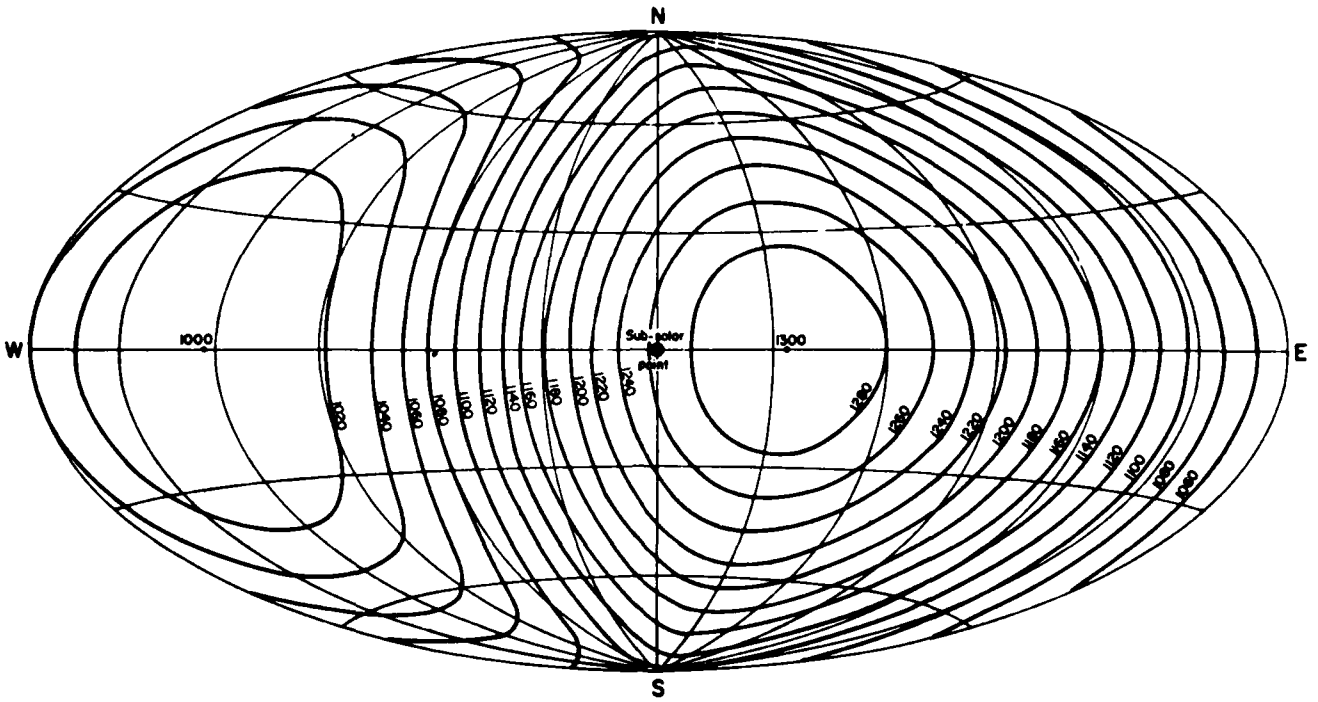



Figure 11.7 Density Variations Inferred from Spacecraft Drag Variations.

EXOSPHERIC TEMPERATURE DISTRIBUTION AT THE EQUINOXES
FOR $T_\infty$ = 1000° K




EXOSPHERIC TEMPERATURE DISTRIBUTION AT SUMMER SOLSTICE
FOR $T_\infty$ = 1000° K

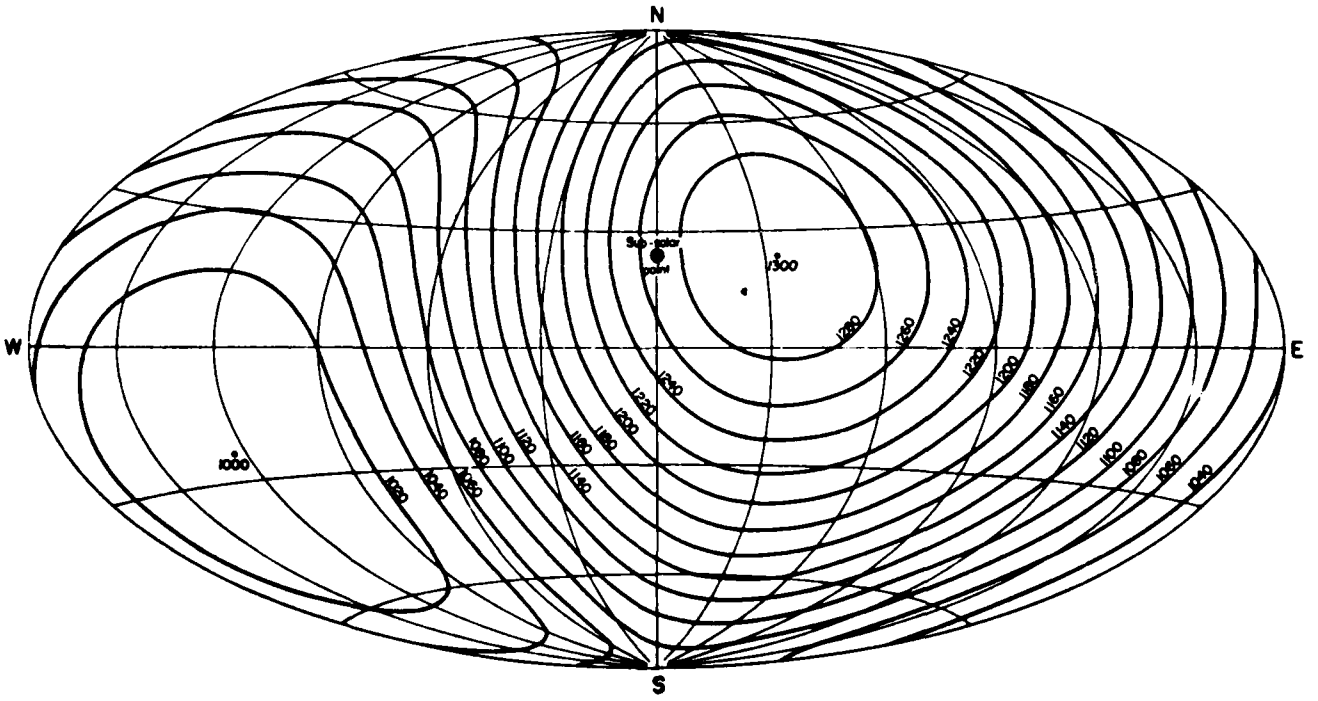



Figure 11.8 Variation in Exospheric Temperature by Latitude, Local Solar Time, and Season (Jacchia, 1965).

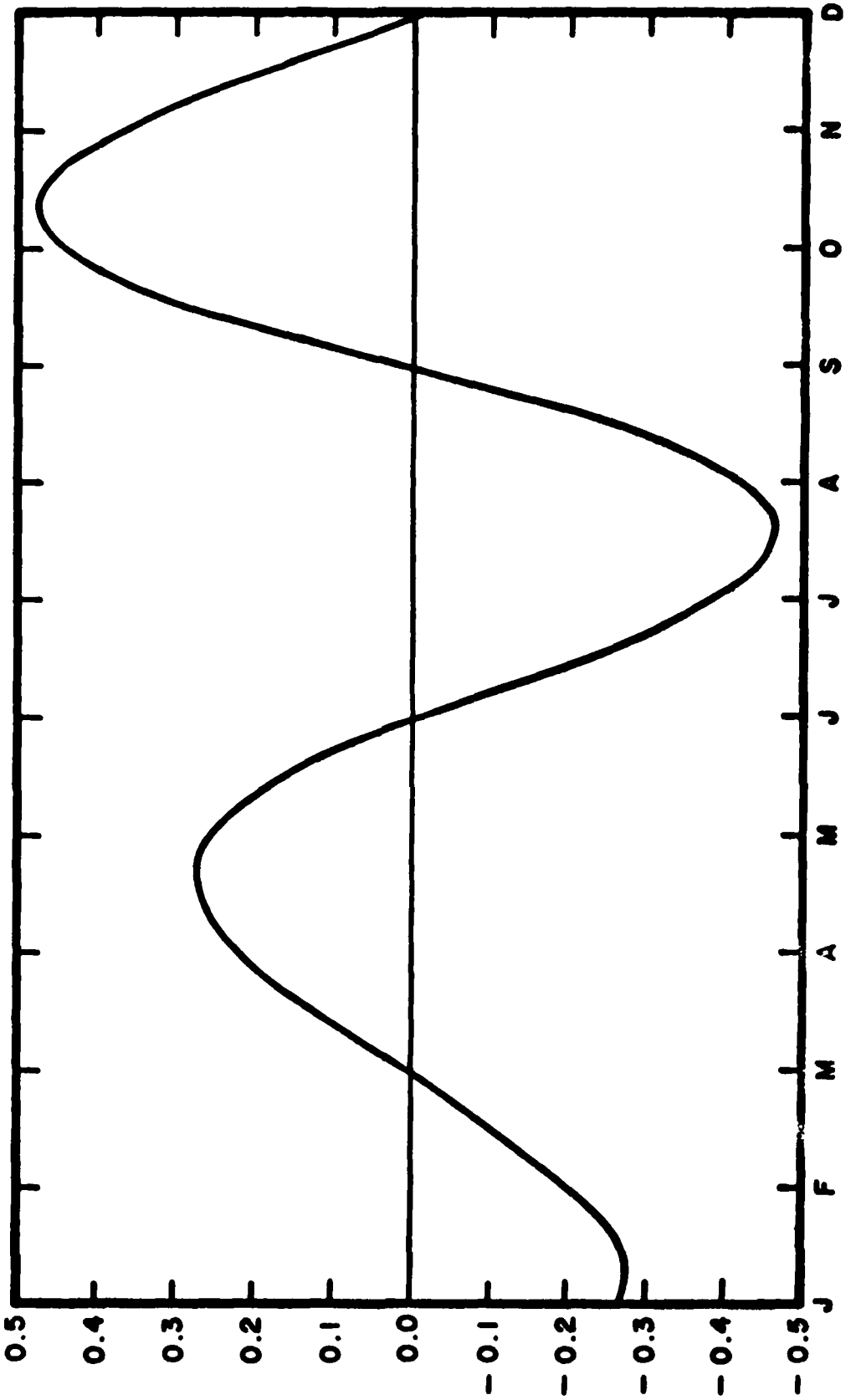



Figure 11.9 Semiannual Density Variations.

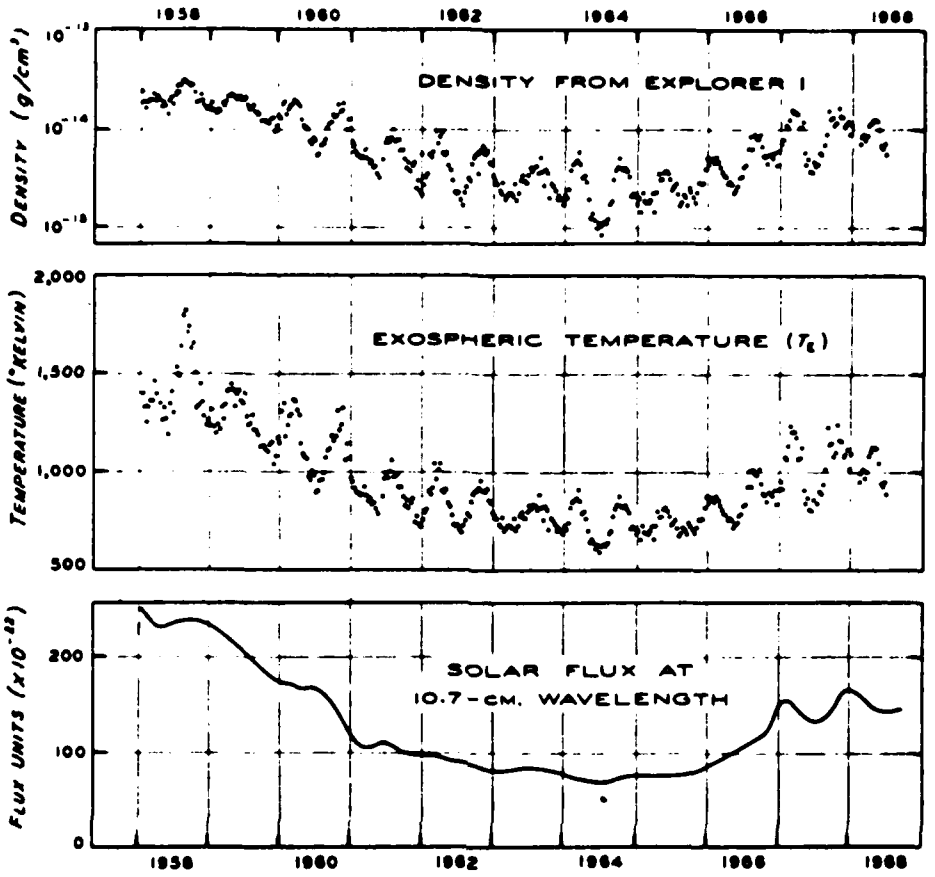



Figure 11.10 Solar Activity and Atmospheric Density (from Jacchia, 1975).

## 11.2 Ionospheric Formation

Electromagnetic radiation and energetic particles ionize and heat the earth's atmosphere. Although ionization occurs throughout the atmosphere, it is most effective (in creating significant concentrations of free electrons) at higher altitudes. Above about 75 km, a significant concentration of free electrons (and ions) persists for an extended time. Conversely, recombination (capture of an electron by an atom or ion) maintains electrical neutrality (or nearly so) at lower altitudes. Recombination also occurs at the higher altitudes, but it proceeds more slowly there due to the lower ambient density. The electron density at any altitude is thus determined by the relative rates of recombination and ionization at that altitude.

Ionization may result from photoionization (due to EM radiation) or collisions between atmospheric particles and energetic, extra-terrestrial particles. Solar x-rays and EUV are primarily responsible for photoionization, which is, in turn, the primary source of ionospheric ionization. Energetic particles (also referred to as precipitating particles,

cosmic rays, and corpuscular radiation) are significant for brief periods of time and in particular regions of the ionosphere. Energetic particles range in energy from a few eV to a few BeV, with the extremely high energy component (BeV protons and alpha particles) generally of galactic origin. The galactic cosmic ray component is relatively stable by comparison with solar cosmic rays and solar EM radiation received at a given location.

### 11.2.1 Chapman Layer Theory

The variation in electron density with altitude, known as an electron density profile (EDP), was first explained by Sidney Chapman in his pioneering work in 1931. Chapman suggested that only a beam of radiation and an atmosphere consisting of atoms capable of being ionized by the radiation were required. In the discussion which follows, only EM radiation is considered; a similar analysis results if energetic particles are considered. Recombination is ignored in the chapman model.

The model initially assumes a single component atmosphere in hydrostatic equilibrium (so density decreases with height) and a monochromatic (single wavelength) beam of radiation vertically incident on the top of the

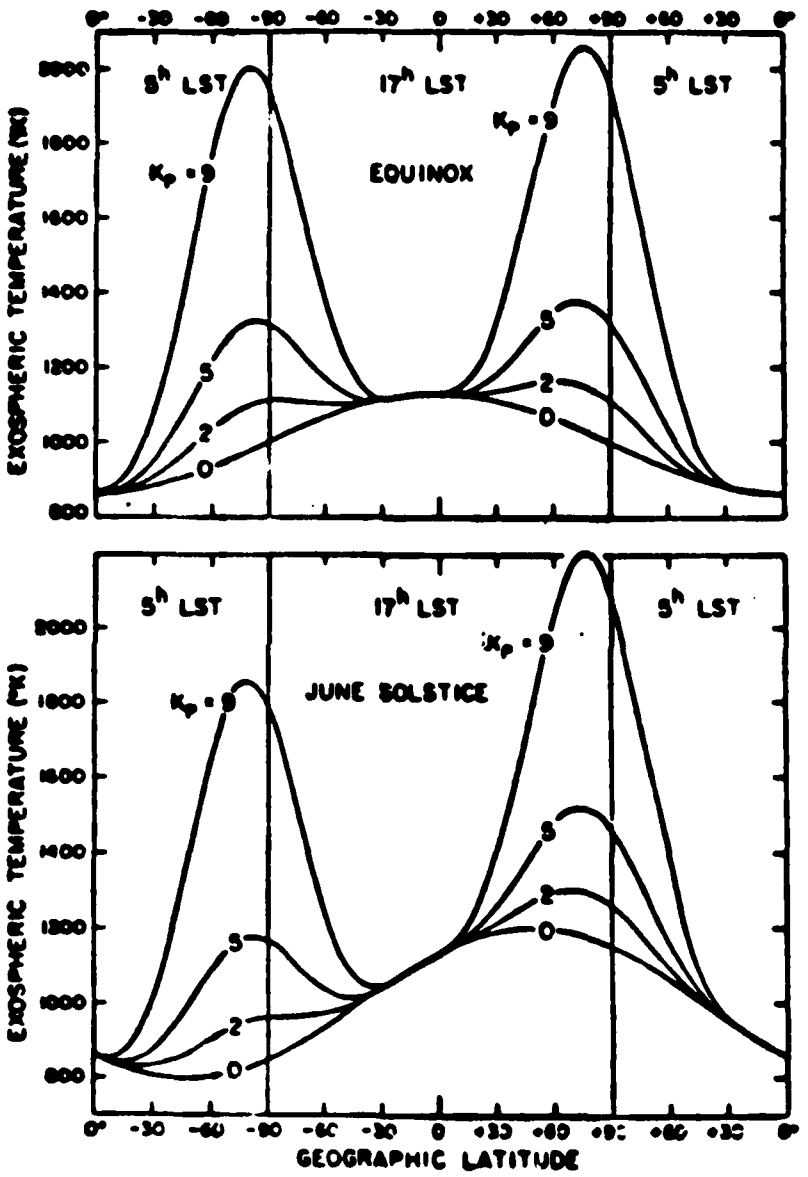



Figure 11.11 Effects of Geomagnetic Activity on Atmospheric Density.

atmosphere. The resulting EDP (the solid line in Figure 11.12) shows only one density maximum. Reference to the dotted lines in Figure 11.12 suggests the reason for such a structure. At the top of the atmosphere, the beam intensity (photon density) is great, but the atmospheric density is low; so only a small number of free electrons can be produced. As the radiation penetrates more deeply into the atmosphere it encounters increasing density and produces an increasing number of free electrons. The photon density reaching each lower altitude is reduced by the absorption which took place at higher altitudes (and extracted a portion of the beam's energy). The resulting electron density attains a maximum and decreases smoothly to near zero as the photon density falls to zero. This EDP is termed a layer (or, sometimes a Chapman layer) because of its smooth, single-peaked structure.

The EDP will change in response to changes in radiation intensity or atmospheric density. Increasing the intensity of radiation by shining the radiation more nearly normal to the top of the atmosphere or by shining more total radiation on the atmosphere will cause two changes. The total electron content (TEC) of the layer (area between the curve and the ordinate) will be increased, and the height of maximum electron density ($h_{max}$) will shift downward. The exact effect of changing the neutral density depends on the altitude(s) at which the change is made. In general, increasing the neutral density at altitudes at or above $h_{max}$ will increase $h_{max}$. It will not significantly alter the TEC, since a given quantity of radiation can ionize only a certain number of atoms of a particular type, regardless of the altitude at which they are located. Decreasing the neutral density above $h_{max}$ will decrease $h_{max}$, since the beam will descend into the atmosphere until it encounters a sufficient number of atoms to completely absorb its energy.

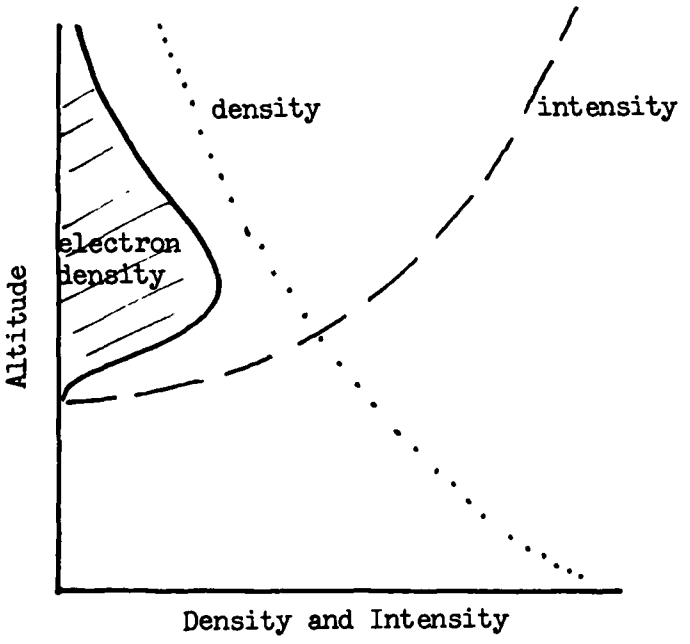



Figure 11.12 Combined Effect of Density and Radiation Variation with Altitude - a Chapman Layer.

### 11.2.2 Ionospheric Layers

The real ionosphere is a sort of composite Chapman layer. This is a consequence of (1) solar radiation and cosmic rays are not monochromatic but possess a wide range of energies and (2) the atmosphere is neither homgeneous nor isotropic--it contains many elements which vary in concentration with both altitude and location. Each combination of ionizing agent (radiation or particles) and atmospheric constituent will produce its own EDP. Since electrons are generally indistinguishable (except by energy, spin, etc.), the result of superimposing several EDPs at a single point is a convoluted EDP revealing several layers (relative maxima) due to predominant elements and ionizing agents (see Figure 11.13).

The vertical structure of the real ionosphere is usually divided into the D, E, F1, and F2, and topside layers. Table 11.1 details the salient features of these layers.

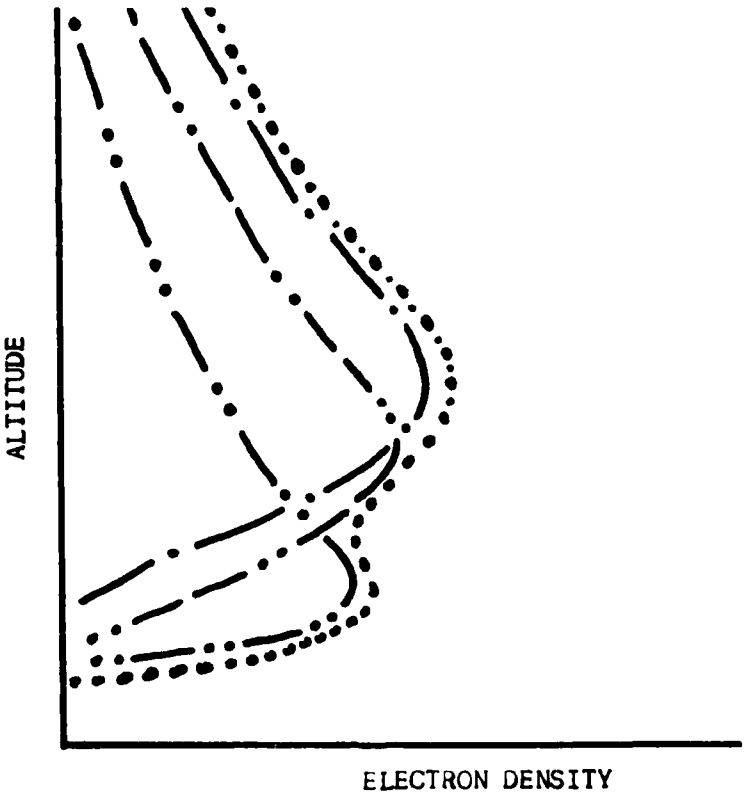



Figure 11.13 Composite Electron Density Profile for a Multicomponent Atmosphere.

| Layer | Height | $N_e(cm^{-3})$ Day | Night | Ion | Source |
|---|---|---|---|---|---|
| D | 75 to 90 km peak 80 km | $10^3$ | - | $NO^+$ | X-rays ( 10Å) Lyman Alpha Cosmic Rays |
| E | 90 to 150 km | $10^5$ | $10^3$ | $O_2^+$ | X-ray/EUV (10-100Å) |
| F | 150 to 1000 km | | | $O^+$ | |
| $F_1$ | peak 160 km | $5x10^5$ | - | | EUV ($10^2$-$10^3$ Å) |
| $F_2$ | peak 280 km | $10^6$ | $10^5$ | | Transport |
| Topside | 280 km - $4R_E$ | $10^3$ | $10^3$ | $H^+$ | EUV and Transport |

Table 11.1 Characteristics of the Major Ionospheric Layers.

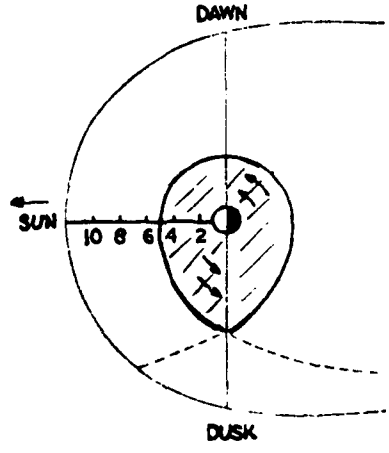



Figure 11.14 Variation of Plasmaspheric Morphology (from Chappell, et. al. 1971).

Of the ionospheric layers, only the E layer approaches Chapman theory regarding strict sun angle control. The D layer is strongly influenced by particle precipitation, particularly at the higher latitudes. The F layer bifurcates during the daylight hours into the F1 (or production) layer and the F2 layer. The F2 layer is maintained by transport of electron density from above or below. The topside is generally taken to begin at $h_{max}$ and end at about 4 $R_E$. In addition to the picture provided by Table 11.1, the plasmaspheric portion of the topside is shaped by solar wind and rotational effects as detailed in Figure 11.14. The morning depression probably results from cooling and recombination during the night and solar wind pressure on the magnetosphere. The afternoon expansion is a consequence of the delayed effect of solar heating, geomagnetic field constraint, and centripetal acceleration.

Our understanding of the ionosphere derives from the theory of a Chapman layer, but it must be modified by many other effects. These modifications yield the quiet or baseline ionosphere.

## 11.3 Quiet Ionospheric Climatology

A truly quiet ionosphere rarely exists. There are nearly always small scale disturbances in progress at various points within the ionosphere. These localized disturbances are characterized by periods of a few minutes (usually less than 10) and wavelengths of 100 km or less. The existence of such phenomena severely limits the distances over which we can accurately infer ionospheric conditions based on a single point observation. In practice, the quiet ionospheric is defined as a sort of theoretical average of a large number of observations made during geomagnetically quiet (but not too quiet--$A_p$ = 8 is ideal) periods. Under such conditions, we often extrapolate middle latitude observations as much as 2900 km in longitude and 1800 km in latitude. Such an average "quiet" condition is taken as a baseline for a given month and level of solar activity. (Notice that the last statement implies, correctly, that "quiet" varies with position in the solar cycle and season.) Day-to-day variations of $\pm$ 20% (in total electron content below about 1000 km, TEC) about baseline values are common. They result from the small scale disturbances already mentioned, and primarily from luni-solar tidal effects. Just as the oceans respond to tidal forces, so, too, does the earth's atmosphere. The "sloshing" of electrons moves available electron density into regions of higher or lower loss rates. The result is a variation in electron density measured at a given point.

### 11.3.1 Latitudinal Ionospheric Regimes

The ionosphere is ordered both horizontally (in geomagnetic latitude) and vertically by its interactions with the magnetosphere and solar radiation. Each of the resulting ionospheric regimes is driven by slightly different phenomena. The vertical structure (the D, E, F1, and F2 layers) results from a variation in density and atomic species with altitude. The latitudinal differences result from electrodynamic interactions as well. The resulting regimes are loosely termed high latitudes (above about 55°), middle latitudes (25°-55°), and low or equatorial latitudes ($\pm$ 25° of the geomagnetic equator).

#### 11.3.1.1 The High Latitude Ionosphere

The high latitude ionosphere is primarily influenced by a considerable influx of energy from the magnetosphere. During geomagnetic storms, the magnetosphere may inject $10^{10}$ - $10^{12}$ watts into the high latitudes. This closely approximates the $10^{11}$ watts deposited by solar EM radiation above 90 km. About 30% of this energy is particle energy (precipitating KeV particles) the remaining 70% is due to Joule heating resulting from the induced currents flowing in and near the auroral zone. These vast amounts of energy may dominate the high latitude ionosphere and strongly influence the middle latitudes as well.

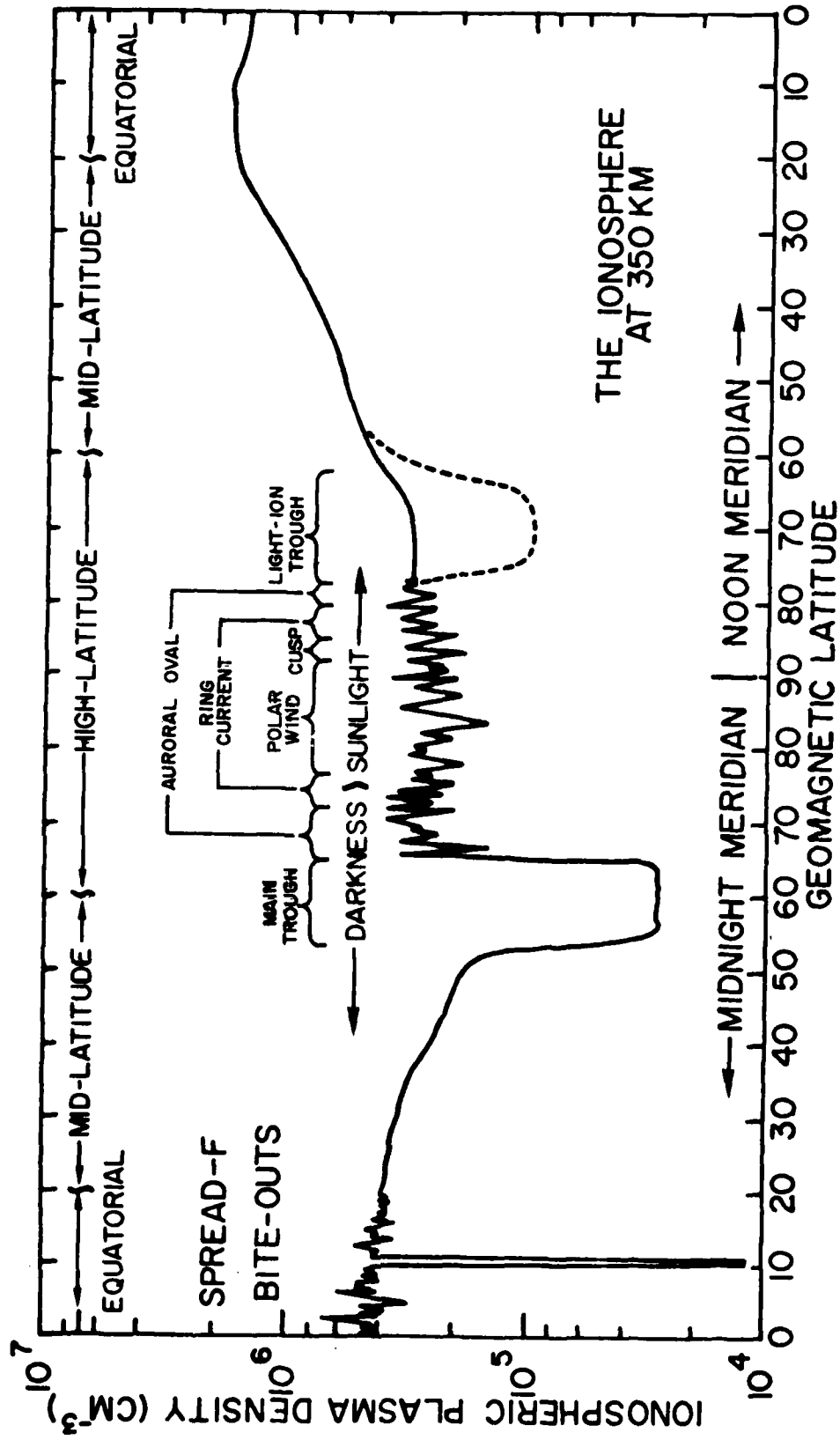



**Figure 11.15** A meridional cross section of electron density with emphasis on the high latitude ionosphere. Values are inferred from spacecraft observations at 350 km.

The high latitude regime is really three separate ionospheric segments: the auroral oval, the polar cap, and the subauroral trough. The geographical location of each segment is defined by the auroral oval, latitudinally the central of the three segments (Figure 11.15). The auroral oval is the instantaneous location of particle precipitation from the magnetotail and from the interplanetary medium, and is itself divided into four zones: evening, midnight, morning, and cusp.

The local evening auroral oval is typically dominated by sporadic E. Sporadic E, as the name implies, is a large but transient increase in the E-layer electron density. Typically, sporadic E exists in thin (1-2 km is typical) dense slabs and often doubles or triples the "quiet" E layer densities. It may alter columnar TEC by as much as 30% at a particular location.

The auroral oval typically displays its greatest latitudinal width ($7^{o}$ or more) in the local midnight zone. Auroral substorms are common here, and the resulting ionosphere is anything but stable. This lack of stability often extends into the morning zone as well.

The fourth zone, the cusp precipitation zone, is located near the noon meridian. This region of the oval is driven by solar wind plasma injected via the dayside cusp or magnetospheric cleft. The predominant feature of this region is a weak/low density F layer, perhaps due to a lack of magnetospheric containment overhead and the absence of an overlying plasmasphere.

Circumscribed by the auroral oval, the polar ionosphere is an entirely separate entity from the remainder of the earth's ionosphere. No conjugate field lines connect the two polar ionospheres. Rather, geomagnetic field lines separate the polar ionosphere from the lower latitudes like a wall. The polar ionosphere is maintained by cosmic rays and particle precipitation as much as by solar EM emissions. The variability in both flux and penetration depth of the ionizing particles results in considerable structural variation in the F layer of the polar ionosphere. This three dimensional variation produces bubbles of abnormally low electron density, unusual tilts, and discontinuities in the EDP which move in time. The overall term for this phenomenon is spread F.

The subauroral trough is the third major feature of the high latitude ionosphere. Sometimes referred to as the mid-latitude trough, this region of abnormally low electron density is located equatorward of the nighttime auroral oval. Figure 11.16 provides a comparison of the trough position by season and level of geomagnetic activity. For a given $K_p$, there is little seasonal variation. A change in $K_p$ does alter the trough position: it moves about $2^{o}$ equatorward for an increase of 1 in $K_p$. The core of the trough is $1^{o}$-$3^{o}$ wide and extends from the late evening to the sunrise meridian at invariant (geomagnetic) latitudes of $55^{o}$-$70^{o}$. Near the trough, densities may change by a factor of two in ten kilometers. Moreover, there is considerable fine structure apparent from spacecraft observations (Figure 11.17).

The subauroral trough is connected (geomagnetically) to a portion of the magnetosphere between the plasmasphere and the plasmasheet horns. This region is often known as the particle cusp. Intense particle precipitation is common

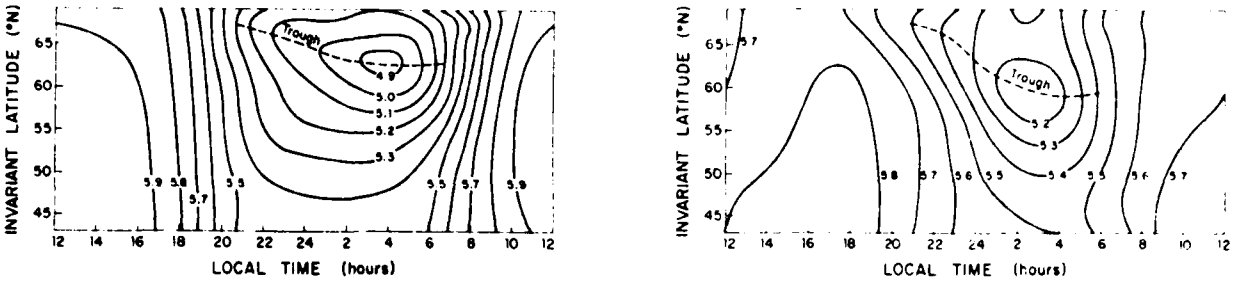



Figure 11.16 Contours of log $_{10}N_{max}$ for Winter with $K_p$=1 (left) and Summer with $K_p$=3 (right) (from Wand and Evans, 1975).

just poleward of the trough, and the associated heating and upward diffusion from the D-region is probably at least partially responsible for the electron density trough. The plasmasphere is the primary source of ionization for the nighttime middle and low latitudes, and cosmic radiation and particle precipitation are primarily responsible for the auroral and polar ionospheres. None of these mechanisms are available to the trough ionosphere, so collisional loss slowly depletes the trough electron density during the hours of darkness. Low solar altitudes during daylight significantly restrict daytime electron densities as well. The trough is of major importance in ionospheric radio wave propagation because of the low densities and steep gradients associated with it.

11.3.1.2 The Middle Latitude Ionosphere

The middle latitude ionosphere is the most classic of the three latitudinal regimes. With the exception of the terminators there are no prevailing ridge or trough structures to mark this region, and solar EM radiation is of sufficient strength to assure electron density stability most of the time. Electric fields and neutral winds are the primary influences on the middle latitude ionosphere. Solar radiation results in generally poleward winds during the daylight hours due to the increased heating and lifting at lower latitudes. An equatorward drift predominates at night. The earth's axial tilt also affects the neutral winds, producing an equatorward component throughout the summer hemisphere (continuously illuminated polar cap) and a poleward wind in the winter hemisphere.

The nighttime middle latitude electron densities are maintained by downward diffusion from the overlying plasmasphere. There is some indication of latitudinal drift across the terminators, but the primary motion seems to be downward. This flow reverses during the daylight hours when the F1 layer electron production replenishes the plasmasphere. This electron exchange between ionosphere and plasmasphere actually involves a bit more--namely, the conjugate ionosphere. Conjugate effects seem to be important in the middle latitudes, with the two hemispheres connected by geomagnetic field lines (flux tubes) which pass through the plasmasphere. The analogy to a siphon constantly primed by the plasmasphere is conceptually appropriate.

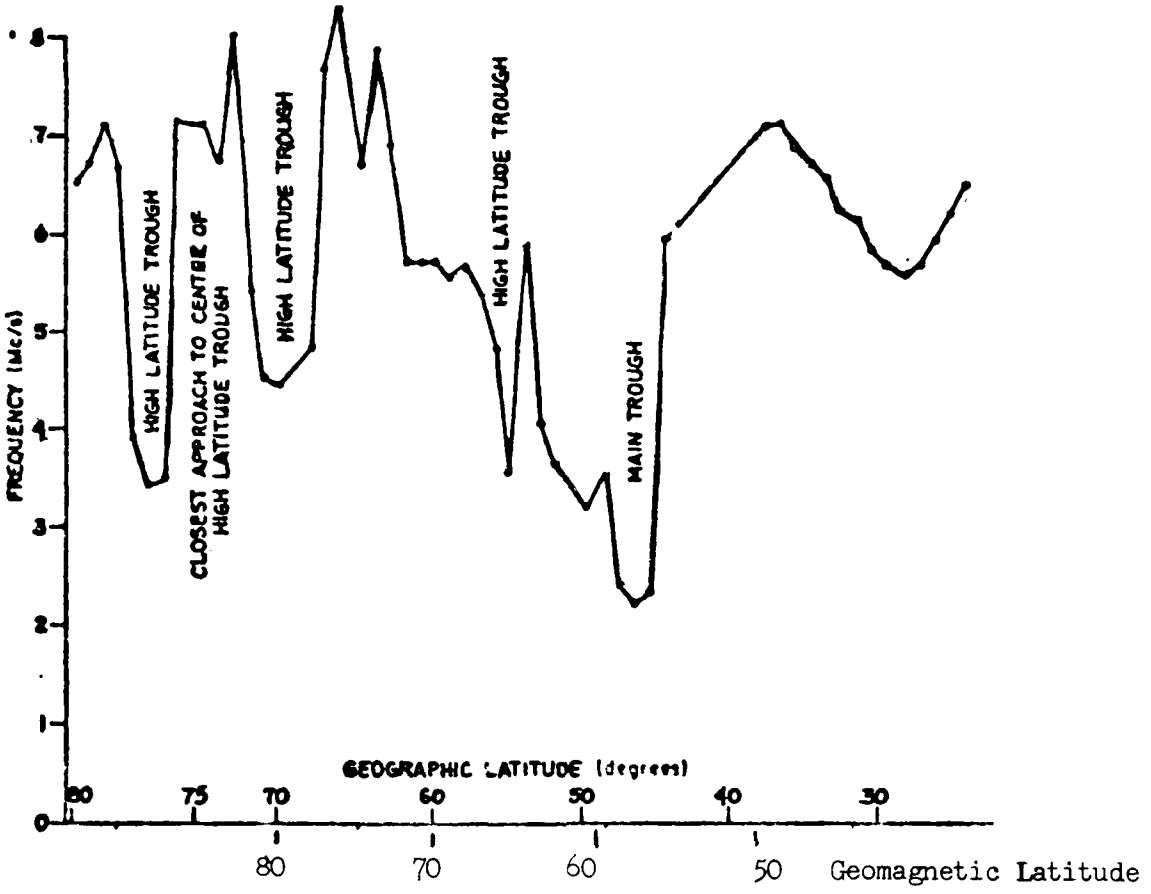



Figure 11.17 Subauroral Trough Morphology from Spacecraft Measurements (from Muldrew, 1965).

Since sunrise occurs first in the summer hemisphere, plasmasphere replenishment begins first here. The relatively lower densities at the winter conjugate point draw additional density from the plasmasphere. After sunrise at the winter conjugate point, the downward flux of electron density begins to slow and finally stops. The reverse situation occurs at sunset. The winter hemisphere overshoots its balance point (electron density equal to that at its summer conjugate) because of the time and damping factor introduced by the plasmasphere. Consequently, there is apparently some back flow to the summer hemisphere sunset). Figure 11.18 provides the qualitative result of this conjugate transport on climatological electron densities. The F-layer critical frequency, $f_oF2$, is related to the peak electron density in the F layer by

$$f_oF2 \simeq 9\ N_e^{0.5}\ \text{KHz}.$$

Figure 11.18 relates ionospheric effects in terms of $f_oF2$ variability. Notice that the winter hemisphere experiences a single, midday (near 1300L) peak in critical frequency which is higher than any value attained during the summer. Conversely, summer densities show a double peak. The first occurs in mid morning and the second just prior to ionospheric sunset (following sunset at the winter conjugate). Conjugate transport easily explains this otherwise perplexing variation.

The plasmasphere's interaction with the middle latitude ionosphere has an additional consequence. The height of maximum electron density, $h_{max}$, is 50-100 km higher near 0000L (when the plasmasphere is supporting the ionosphere) than near 1200L, when solar EM effects approach maximum. $H_{max}$ is usually greatest shortly before ionospheric sunrise in the middle latitudes and falls rapidly at sunrise. This means that sharp gradients are likely near the sunrise and sunset (where the reverse effect occurs) terminators.

As we approach the magnetic equator, the vertical component of the geomagnetic field decreases steadily. The predominantly horizontal nature of the low latitude geomagnetic field minimizes conjugate effects.

11.3.1.3 Low Latitude Ionosphere

The equatorial electrojet and electromagnetically induced drifts are the dominant influences on ionospheric structure below about 25° geomagnetic latitude. This region also receives some energy injection from dissipation of the geomagnetic storm ring current and the more intense solar EM radiation prevalent in the low latitudes.

The equatorial electrojet is an intense ($9 \times 10^{-6}$ amp/$m^2$) easterly current associated with the sunlit geomagnetic equator. The current is centered near an altitude of 110 km and onsets near the 1000L meridian. It extends to the sunset meridian, and the entire current system is sunsynchronous. There are indications of a weaker, westward electrojet in the equatorial evening sector. Figure 11.19 is an observational cross-section of the equatorial electrojet system. Its concentration in the E region is probably a consequence of the large number of charge carriers available (5-50 tons of meteor dust--much of it metallic ions--are deposited daily in the E layer).

The electrical field associated with the electrojet combines with the earth's magnetic field (E x B) to produce vertical drift of electrons from the E layer. Rapid north-south motion through the overlying (conjugate connecting) flux tubes may accelerate electron upwelling from the E layer. Sunlight provides the ionization and heating to facilitate this. At some distance above the electrojet, neutral winds push this "bubbled-up" electron density north and south away from the geomagnetic equator along magnetic field lines. This motion depletes electron densities over the electrojet and produces significant enhancements north and south of the electrojet. Figure 11.20 provides a meridional cross-section of electron densities by altitude near the electrojet.

Notice that the resulting subequatorial electron density "ridges" are not symmetric. Greatest densities are found in the summer hemisphere. The ridges follow the sun but lag the subsolar point somewhat. They are most intense near sunset and dissipate completely by local midnight. Since the ridges are aligned with the geomagnetic equator, they change geographic latitude during the course of each day. Their relative intensities also vary during ionospheric storms, and they have a significant impact on radio wave propagation near or across the geomagnetic equator.

The latitudinal structure of the ionosphere is controlled by the geomagnetic field. The primary features of the high, middle, and low latitude regimes result from geomagnetic and electrodynamic features. The

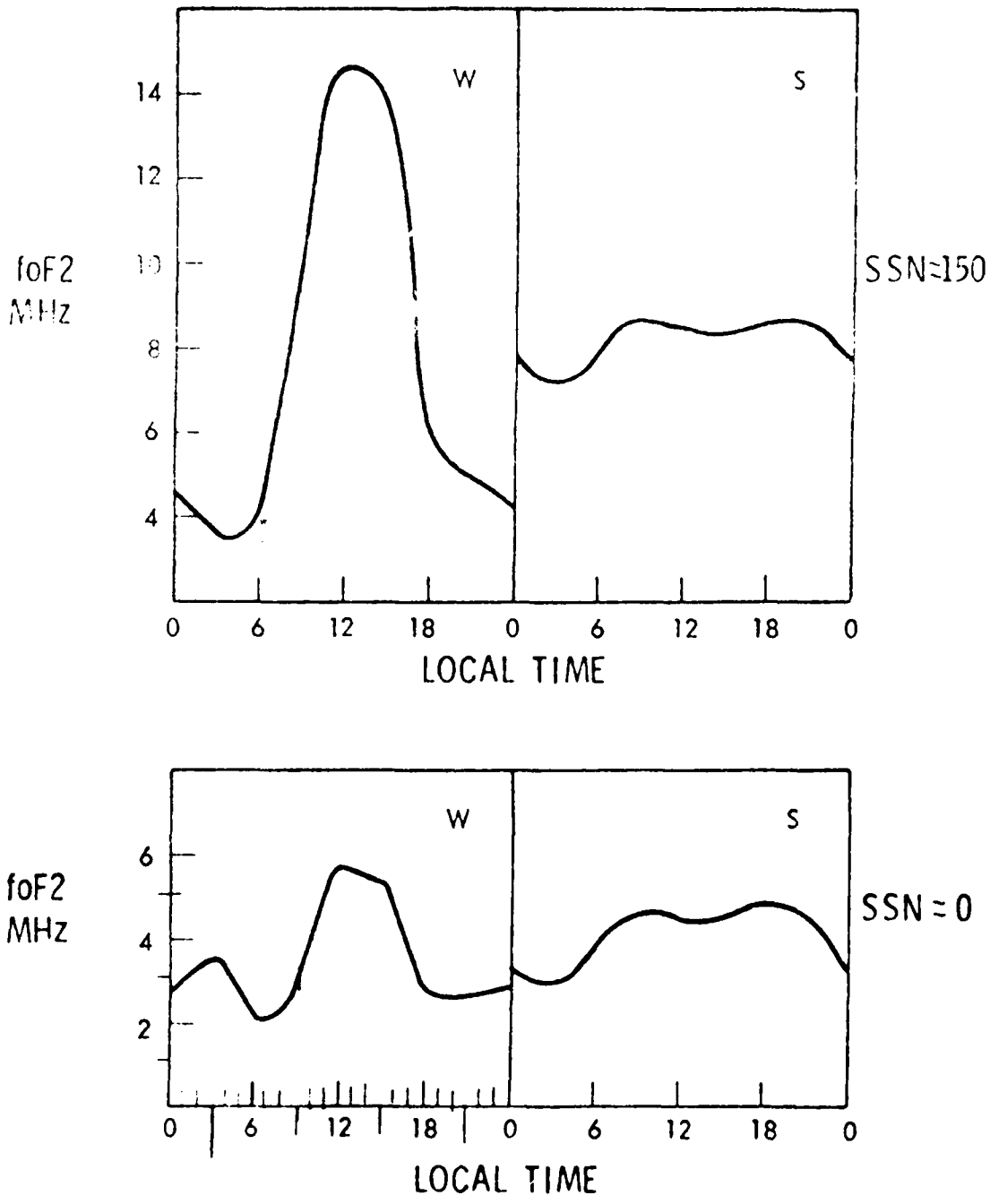



**Figure 11.18** *Diurnal Variation of the F-layer Critical Frequency with Season at Two Sunspot Numbers.*

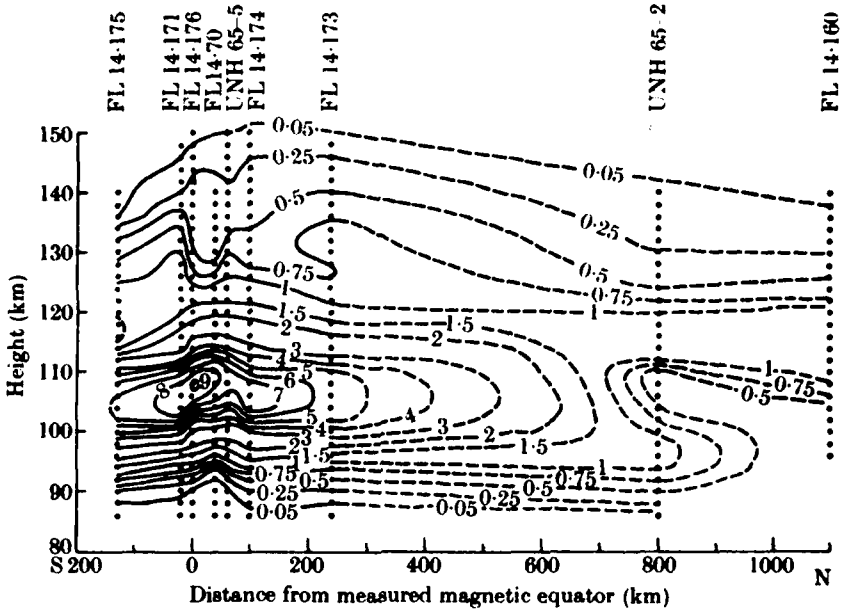



Figure 11.19 Distribution of the Equatorial Electrojet System (from Akasofu and Chapman, 1972). Units are $10^{-6}$A m.$^{-2}$

noon-midnight meridional cross-section in Figure 11.21 displays most of the features discussed thus far. It is organized in geomagnetic coordinates, and contours are of constant plasma frequency (MHz). The height of maximum density varies significantly between the equator and pole. The troughs and ridges are also readily apparent. Equally important are the large number of strong horizontal density gradients disrupting the more theoretically pleasing vertical stratification found only in the middle latitudes.

### 11.3.2 Ionospheric Height Regimes

Vertical stratification of ionospheric electron density is a natural consequence of Chapman theory. Limiting consideration to electrons (neglecting ion species) is much more simple (all electrons look alike) and probably more meaningful (the electron plasma frequency will generally exceed that of the ions by a significant margin). A typical electron density profile (EDP) such as Figure 11.13 shows no sharp layer boundaries, but rather a smooth variation with height. Ionospheric layers are theoretical constructs based primarily on the ionosphere's varying response to outside influence. The consequence is considerable flexibility in specifying the height and thickness of each layer. Indeed, these parameters vary with time, location, and ionospheric condition.

The largest scale vertical division of the ionosphere is based on the level of maximum electron density. This height varies, but it will always exist. As such, it divides the ionosphere into the bottomside and topside components. This division is a physical one, established by sensing constraints. A radio wave of sufficiently high frequency to penetrate the

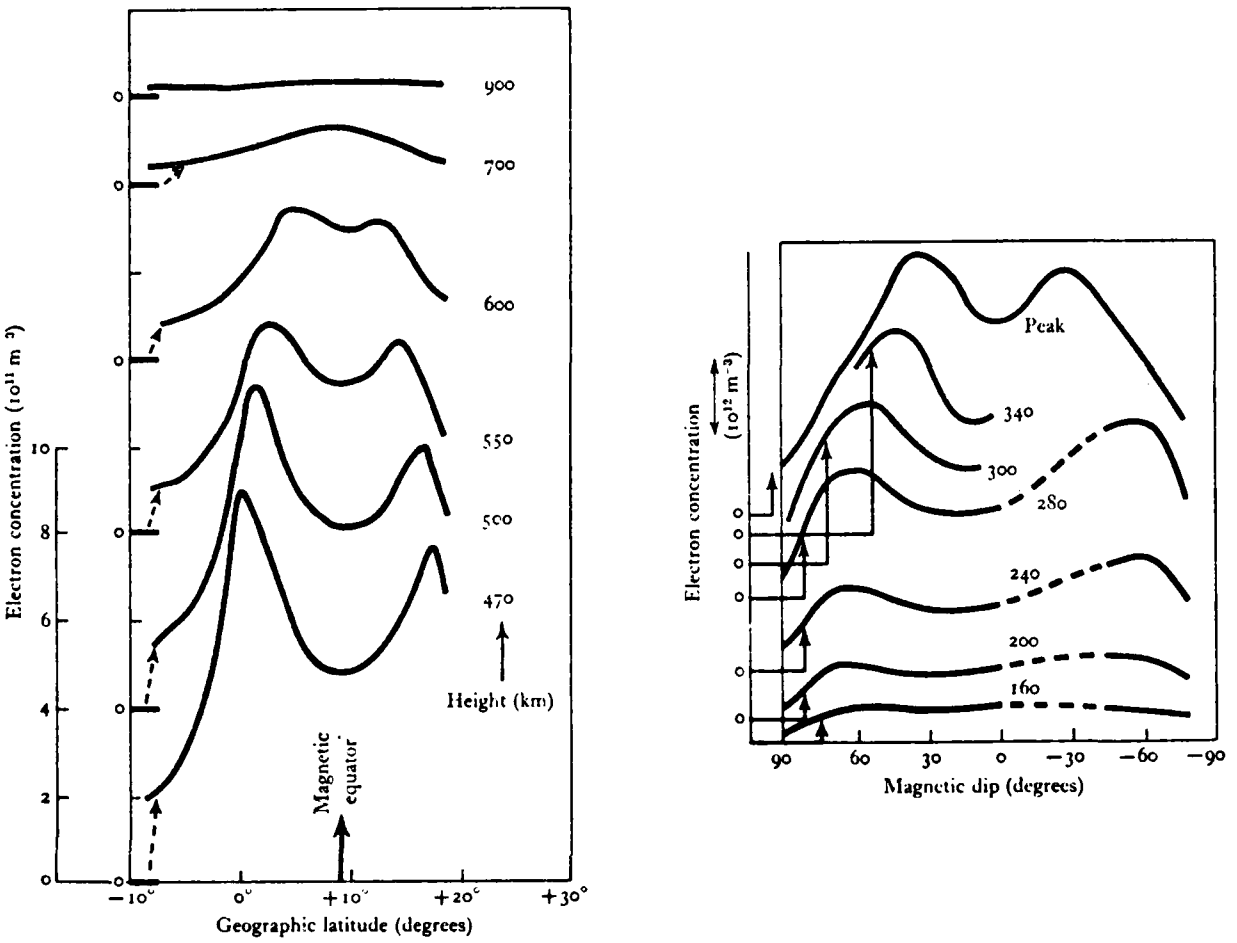



Figure 11.20 Electron Concentration Above (left) and Below (right) $H_{max}$ Near the Geomagnetic Equator (after Ratcliffe, 1972).

plasma at $h_{max}$ will not typically be reflected by any other level of the ionosphere. Lower frequencies will be unable to cross this boundary. Since radio waves are the primary means now employed to probe or use the ionosphere, it would seem that for a transmitter located below $h_{max}$ only the bottomside is accessible. A similar situation applies to the topside. There are a few, limited exceptions, but the operational division into bottom and topside ionospheres is of considerable value in analyzing the ionosphere.

11.3.2.1 Topside Ionosphere

The ionosphere above $h_{max}$ is the least well understood of all ionospheric regimes. This lack of understanding stems primarily from its inaccessibility to ground-based probes. Only in the past decade has it been possible to study even a fraction of the topside structure. Spacecraft and sounding rockets are the primary tools.

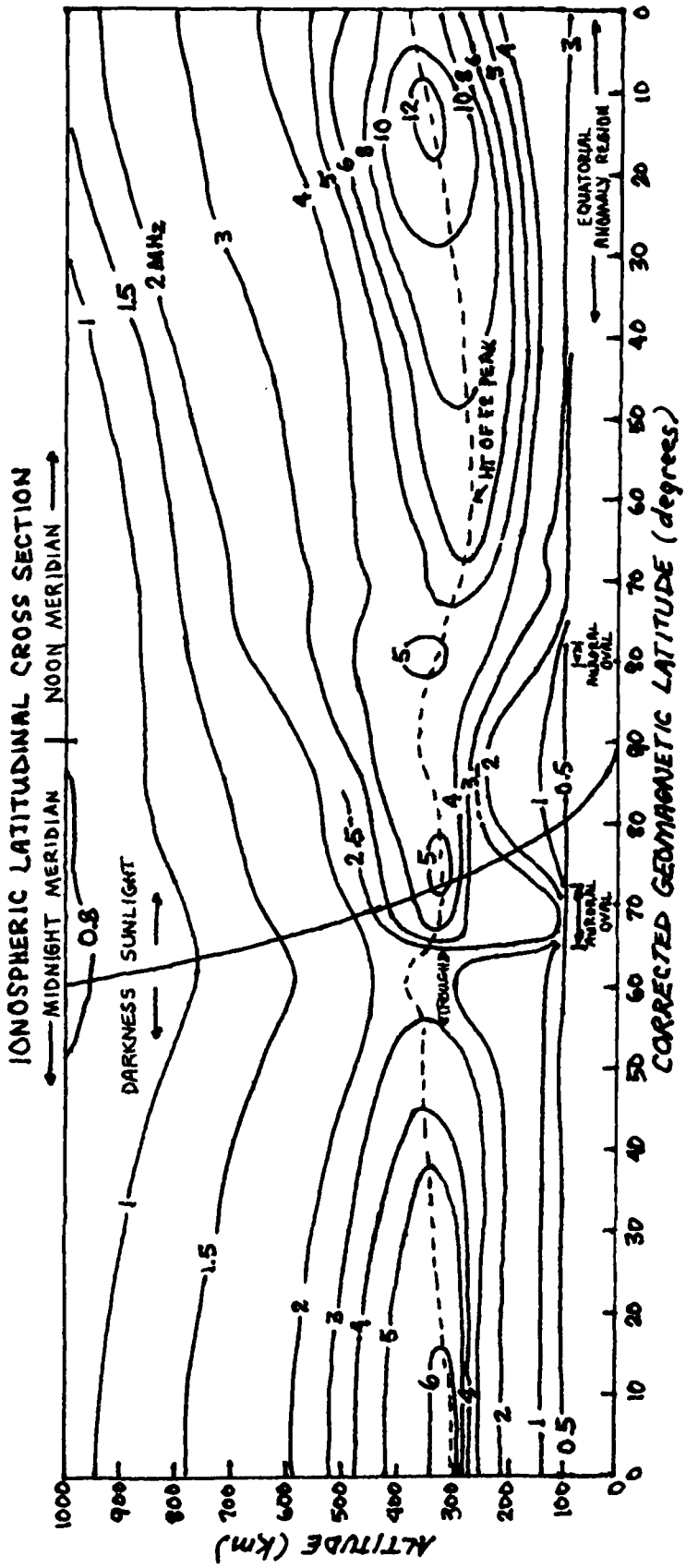



Figure 11.21 Ionospheric Cross Section Along the Local Noon-Midnight Meridian. Contours are of Constant Plasma Frequency in MHz. (After Carrigan and Skrivanek, 1974).

In spite of our limited knowledge, it is certain that the topside ionosphere is almost totally dominated by the geomagnetic field. Half or more of the total ionospheric electron content resides above $h_{max}$. The plasmasphere is a part of the topside ionosphere. It reveals considerable temporal and spatial variability. The ratio of plasmaspheric electron density to that below about 1000 km ($h_{max}$ ranges from 250-400 km) varies from 0.5 before dawn to near 0.15 in the early afternoon (both for middle latitudes). The subauroral trough extends through the topside ionosphere, but the equatorial trough/subequatorial ridges are nearly invisible in the topside structure.

Topside structure reveals numerous transient holes and ridges which are not apparent in the bottomside structure. Considerable longitudinal variability is also apparent above $h_{max}$. The minimum variation seems to exist over eastern Asia (greatest stability). There are some indications (Davies, 1980) that topside variations are oppositely directed over Europe by comparison to those at the same local time over North America. Moreover, topside variations are almost certainly out of phase with those in the bottomside.

11.3.2.2 Bottomside Ionosphere

The bottomside ionosphere is the best studied of the two vertical regimes. For most purposes, it is also the most significant ionospheric regime. It is strongly influenced by compositional changes and neutral winds. Strong electric currents (electrojets) found here influence certain segments of the bottomside structure. The compositional stratification of the heterosphere leads to a layered structure in the bottomside extending from the lowest, or D-layer to the level containing $h_{max}$ - the F2 layer.

11.3.2.2.a D Layer

The D layer, extending from as low as 50 km to about 100 km, is the lowest of the major ionospheric layers (there is evidence for a C-layer, but most effects of significance to DOD result from the D or higher layers). The D-layer is solar-controlled and nearly disappears at night. A vestigial D-region persists in the polar cap and near the auroral oval in response to particle precipitation and cosmic ray bombardment. Maximum electron densities usually occur in the early afternoon and in conjunction with large solar flares.

D region electron densities vary both smoothly and randomly over long periods of time. Sunspot maximum yields higher average densities than are found at cycle minimum due to increased solar emission. Similarly, densities are usually higher in the summer hemisphere due to the more direct sunlight.

The winter anomaly provides a significant exception to the summer peaks. The winter anomaly usually results in a significant increase in electron density between 60-90 km over an area of as much as $100^{o}$ in longitudinal extent in the winter hemisphere. Several theories have been suggested to explain this ionization. Low altitude compositional changes enhance concentrations of NO and $O_2$ while depleting $H_2O^+$. This change results in increased ionization while reducing recombination.

The winter anomaly has also been associated with the onset of large geomagnetic disturbances, but it is constrained to the winter hemisphere. Such a disturbance (geomagnetic) produces particle precipitation in both hemispheres

(both conjugate points), but the winter ionosphere is cooler. Precipitating electrons can penetrate to a much lower altitude in a cool atmosphere. Higher densities at low altitudes permit much greater ionization for the same particle energy flux. The effect is a lowering and intensification of the D region. The anomaly is often first apparent between a few hours and two days following the sudden commencement or sudden impulse; is most significant in the daytime middle latitudes; and may persist for several days over an area of 100° or more in longitudinal extent.

Stratospheric warming is sometimes also suggested as a cause of the winter D region anomaly. Certainly the effect of each is similar (i.e., an increase in D region electron density), and the areal extent is comparable. Stratospheric warming is marked by a sudden (several days) increase in the temperature at and possibly above about 30 km by tens of degrees. It typically occurs in November-March in the northern hemisphere and is often observed over Siberia or the North Atlantic. It typically covers 15° or more in latitude and longitude and often expands with time to covers the entire area poleward of of H0° geomagnetic latitude.

Stratospheric warming results from the atmospheric circulation in the winter hemisphere. Large verticle motions cause NO and $O_2$ to be transported into the lower D region. A decrease in $H_3O^+$ and increased NO means increased electron production and reduced recombination are both possible in the D layer. D region electron density increases in the affected area, and the base of the D region may extend to a lower altitude.

Electron density changes are often measured by their impact on selected frequency bands (usually selected by operational interest). The high frequency (HF: 3-30 MHz) band is probably in the most widespread use for probing and employing the ionosphere. Increased D region electron density generally produces increased absorption on HF frequencies. The effect decreases with increasing radio wave frequency. It results from the high neutral densities collocated with the free electrons. The electrons move in response to the wave energy, but many collide with neutral particles before retransmitting the wave energy. This means that wave energy is dissipated (usually as heat) in the D region, and received signal strength is reduced. Solar flares produce a similar effect.

Precipitating particles also generate increased HF absorption. They are most common in and near the auroral oval. Here, several distinct phenomena have been identified. They include auroral zone absorption (AZA), relativistic electron precipitation (REP), and auroral absorption spikes. The three differ slightly in origin, duration, and extent.

A band of increased absorption is found equatorward of the auroral oval. The absorption is particularly intense in a 5° wide band extending roughly from the 0500L - 1200L meridians. This long-lived increase in absorption is often but not always associated with a geomagnetic disturbance. It results from precipitating KeV energy electrons depositing the bulk of their energy near 80 km, and is called auroral zone absorption (AZA).

The REP is smaller in time and space and somewhat stronger in effect than an AZA. REPs often occur during the recovery phase of a geomagnetic storm and may persist for 1-6 hours. They are most common in and near the oval between

the 0600L-1800L meridians. A REP is thought to result from a localized dumping of relativistic (greater than 400 KeV) electrons and the bremsstrahlung produced x-rays of these particles. These x-rays may penetrate to altitudes as low as 70 km, thereby markedly lowering a small segment of the D-region.

Auroral absorption spikes are brief (5-10 minutes) localized enhancements of an AZA pattern. They result from soft (16-300 KeV) electron precipitation in a 10-50 km wide ribbon shaped band several hundred kilometers in east-west extent. The softer energy spectrum results in the electron density increase occurring near 90 km. Absorption spikes usually occur in conjunction with a geosynchronous plasma injection or the onset of a substorm and are usually situated along the poleward boundary of the ambient radar aurora.

The quiet D region is solar-controlled, but its strict sun angle dependence is blurred by its susceptibility to particle ionization. Chemical composition changes can also be significant because of lower altitude variations. The high neutral densities in the D region make it particularly receptive to sudden disturbance by solar activitiy, such as solar flares, of which more is said elsewhere.

11.3.2.2.b E layer

The quiet E layer is the most ideal Chapman layer of the ionosphere. It displays a strict sun angle response, with summer daytime ionization peaks and a diurnal curve closely aligned with solar zenith distance. Some stratification is apparent, and, on occasion, there are indications of E layer bifurcation. Neutral winds and the auroral and equatorial electrojets maintain some residual E level ionization during the nighttime hours, an effort facilitated by the lower rate of recombination in comparison to the D region. Winds and electrojets not only maintain the darkened E layer, they also account for E layer variability.

Sporadic E is the primary variable feature of the quiet E region. Typically found between 90-120 km, sporadic E is a transient, dense slab of ionization. It is usually 1-2 km thick and ranges from tens to hundreds of kilometers in diameter. Electron densities in these clouds are often two to three times that of the ambient E layer--sufficient to noticeably alter the electron density profile at the point of occurrence. Following the cloud analogy, sporadic E may move in time. It is often found near the electrojets, the auroral oval, or in areas of intense particle precipitation. In the middle latitudes, sporadic E is often associated with meteor showers and intense thunderstorms (and squall lines). The exact connection with the latter is uncertain because of the vast difference in altitudes, but may result from electrical activity in the storm system.

Sporadic E may be thick or thin, blanketing or transparent. The relative transparency depends not only on the electron density but also on the probe frequency and the angle of incidence. It may vary with time. Likewise, the exact alignment (tilt and position) of the sporadic E cloud may vary. A thick sporadic E layer may significantly alter an incident radio wave. An HF radio wave may spend as much as a third of its trajectory (assumed oblique) in a thick sporadic E layer as a consequence of refraction. This changes the geometry of the oblique path by raising the virtual height (based on transit time of a radio wave and may differ from the physical height) of reflection.

Altering the virtual height may, in turn, result in as much as a 10% depression of the maximum usable frequency (MUF) on an oblique, E mode path in comparison to propagation off a thin sporadic E layer. (This variation is due to the change in the height of reflection which is given by the M-factor. The MUF = (M-factor) x (critical frequency of the layer, or plasma frequency).)

The climatology of sporadic E is best considered by geomagnetic latitude. At low ($25^{o}$ or less) geomagnetic latitudes, sporadic E is primarily a daytime phenomenon showing little seasonal variation. It is often found in association with the equatorial electrojet or near the geomagnetic anomalies (Southeast Asian and South Atlantic). Sporadic E is least common in the middle latitudes ($25^{o}$-$55^{o}$ geomagnetic). It seems to be a summer daytime phenomenon, but is also found associated with meteor showers, strong vertical wind shears, and squall lines. High latitude (above $55^{o}$) sporadic E is usually found near the auroral oval/electrojet system. It results (probably) from particle precipitation and shows strong correlation with geomagnetic activity. Little seasonal variation is apparent. The most intense sporadic E (electron plasma frequencies at or above 8 MHz) is often found in association with intense aurora. Plasma frequencies of sporadic E associated with the diffuse aurora seldom exceed 2-3 MHz. (Recall, plasma frequency = 9 $N_e^{0.5}$ KHz.) Sporadic E is also found in the polar cap. It probably results from the expansive phase of an auroral substorm, and is also common during quiet geomagnetic times as a consequence of particle precipitation. When $B_z$ turns southward, this precipitation ceases.

11.3.2.2.c F Layer

Peak ionospheric electron densities usually occur in the F layer. It is the most complicated, variable, and important of the bottomside layers. During daylight hours, it bifurcates into the F1 and F2 layers.

The F1 layer is the primary production layer of the ionosphere. Its existence is limited to daylight hours. At night, it merges into the F2 layer. It is most pronounced during the summer months, near solar cycle maximum, and during negative ionospheric storms. The height of peak production seems to be near 150 km. Horizontal and vertical transport seem unimportant in the F1 layer.

Of all ionospheric layers, only the F2 layer seems to lack a significant in-layer source of ionization. It is sustained by vertical transport from the F1 layer (daytime) and the plasmasphere (night). Upward transport (from the F1) begins about 2 hours after ionospheric sunrise and reaches a maximum in the summer hemisphere about 2 hours after conjugate sunrise. Transport becomes predominantly downward (drainage) in the summer hemisphere following the peak of upward transport in the winter hemisphere. This conjugate effect dominates the middle latitudes and accounts for the climatological variations between the winter and summer hemispheres. Conversely, the high latitude F layer shows very little diurnal variation (except for the subauroral trough) due to the nearly constant sun angle during much of each day.

The low latitude F region is influenced by rapid motion through the overhead flux tubes during the daytime. This, combined with vertical transport from the E and F1 layers produces subequatorial ridges of electron density during the day and a relative trough along the geomagnetic equator.

Low latitude electron densities and the heights of these maximum densities both show a single peak near 1900L (local sunset). $H_{max}$ is typically 100 km lower at 0000L than at local noon. The shut off of overhead flux (conjugate transport) and a negligible or slightly westward electrojet in the nighttime tends to force equatorial electron densities to lower altitudes.

The F layer shows a strong solar cycle variation. Figure 11.22 compares F2 plasma frequency variations to those of the F1 and E layers for a given upper middle latitude site. Notice that the F2 plasma frequency (and, by inference, electron density) increases steadily up to a sunspot number of about 150 and seems to level slightly for higher levels of solar activity. This probably results from an increase in F layer collisional recombination (due to increased heating increasing neutral densities at F layer heights) partially offsetting the increased ionization. It is important to remember that the solar sunspot number (SSN) has no causal relationship with F layer electron density. The ionization results from EUV and x-ray emission. SSN is a convenient index assumed representative of solar EM emission. While this is generally true, it may not hold in specific cases. Even the commonly used F10 index is not always representative of ionizing emission levels at a given time. Extreme care must be taken in the interpretation of correlations relating F10 or SSN to ionospheric response. Figure 11.23 summarizes the effect of seasonal and solar cycle effects on the ionospheric EDP.

The F2 layer's dependence on outside sources of ionization produces a strong susceptibility to transient variations. Often termed anomalies, these features are predictable consequences of F2 layer physics. Conversely, spread F is not as easily predictable. Rather, its occurrence is generally handled statistically.

The winter F layer anomaly is most apparent in the upper middle ($45^{o}$-$50^{o}$) latitudes of the winter hemisphere. Midday electron densities may exceed those in the summer hemisphere by a factor of four. As a consequence, the diurnal variation in F layer electron density is a maximum in the winter hemisphere. This effect, a consequence of conjugate transport, nearly disappears during solar minimum. Conversely, maximum nighttime F layer densities are observed in the summer hemisphere because of the longer period of ionization (daylight).

The December anomaly reinforces the winter anomaly in the northern hemisphere. It results in a worldwide increase in midday electron density. A maximum increase of as much as 20% is observed between $35^{o}$S-$50^{o}$N. The December anomaly results from a 6% increase in solar radiation due to solar proximity (perihelion).

The offset of the geomagnetic field results in anomalous areas in Southeast Asia and over the South Atlantic. The lower field strength for a given altitude in the South Atlantic produces in lower (than elsewhere) trapping altitudes. This means trapped particles encounter higher neutral densities here than elsewhere, and particle precipitation is increased. Since trapped electrons drift westward, they precipitate mainly along the western edge of the South Atlantic anomaly. Protons, drifting eastward, are precipitated along the eastern anomaly boundary. This dumping tends to accentuate F region anomalies (and, if energies are sufficient, may increase

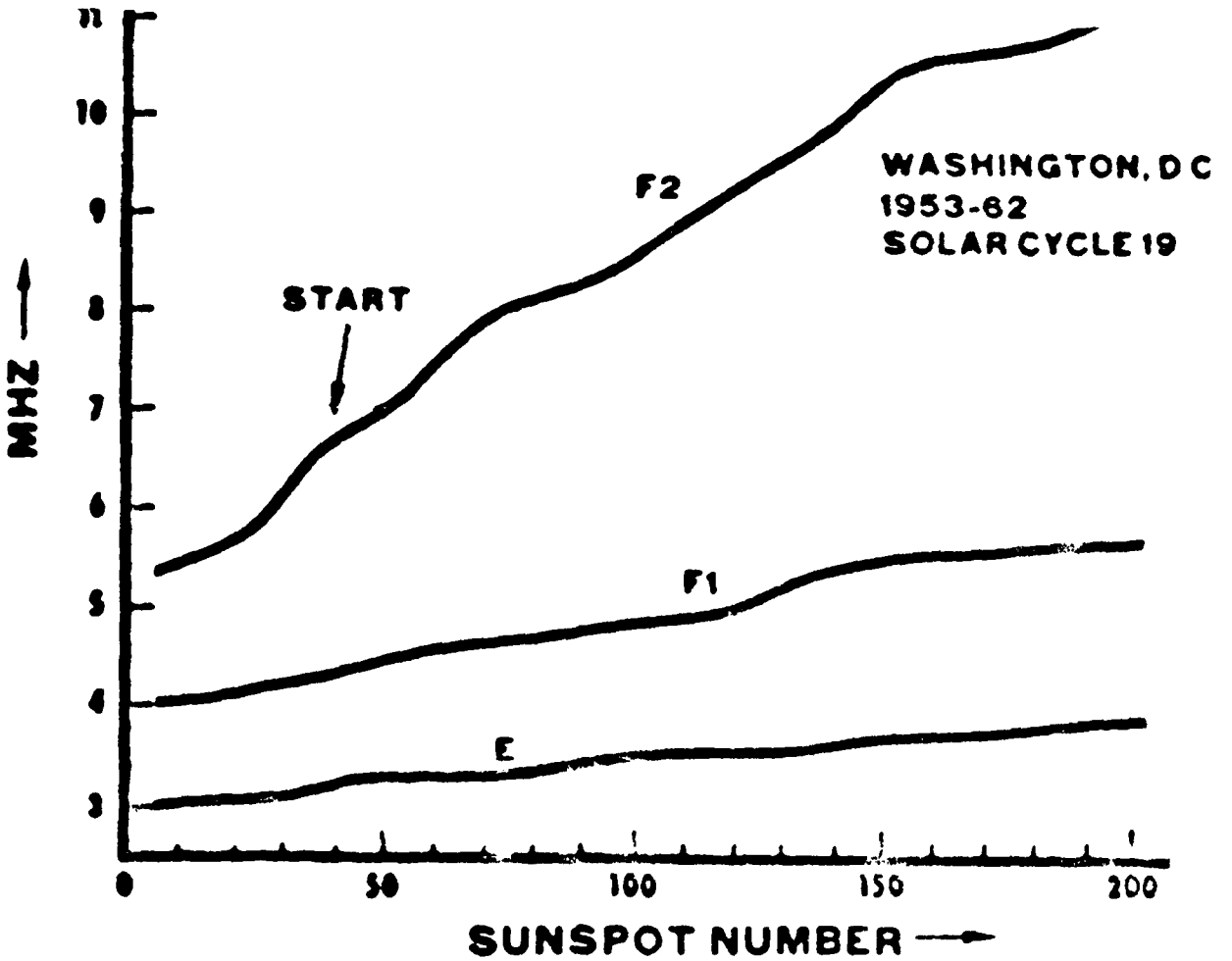



Figure 11.22 Ionospheric Response to Varying Levels of Solar Emission as Measured by Solar Sunspot Number.

D region absorption). The Southeast Asian anomaly is the reverse of the Atlantic situation, with stronger trapping fields present at a given altitude. This means that anything (e.g. a geomagnetic disturbance) which disturbs the trapping regions will precipitate higher energy electrons in the Asian area. Moreover, since the Asian anomaly traps higher energy particles at a given altitude, their precipitation will produce effects at lower altitudes. (Their higher energy yields deeper penetration, greater ionization, or both.) D region absorption is a likely result of particle dumping in the Southeast Asian area. As with many other portions of the ionosphere, the areal anomalies are still not well-monitored. Additional work in both monitoring and theory is needed.

Considerable research has begun to reveal some details of the phenomenon known as spread F. Spread F is the name given to abnormal variations in the F region electron density profile with height. Given time, spread F structures may produce unusual ionospheric density tilts or result in nearly constant density over a considerable range of height. Several factors are involved in the production of spread F, and extensive statistical correlations exist by latitude as well. Its name is a consequence of its impact on a probing radio wave. Inhomogeneities in the electron density profile near $h_{max}$ produce

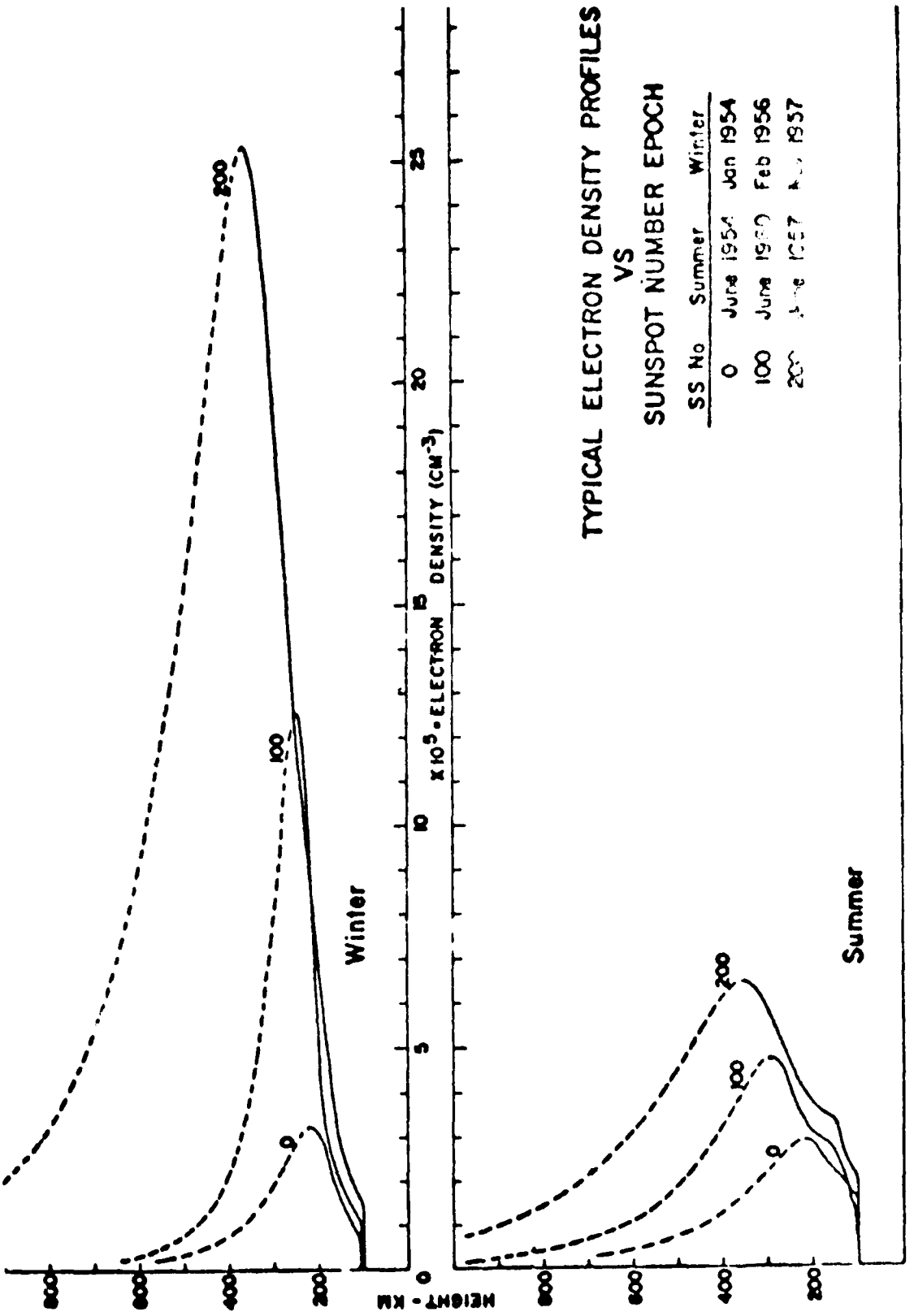



Figure 11.22 Seasonal and Solar Cycle Variations in the Electron Density Profile (after Wright, 1962).

overlapping, fuzzy echoes from radio waves used to probe the electron density structure. Spread F effectively smears the electron density profile over a small area and may significantly degrade radio wave propagation. A typical discontinuity may extend several hundred kilometers vertically and 300-600 km horizontally. Initiation and occurrence appear to vary somewhat with latitude, with low latitudes often being the most severely impacted. Anything which results in uneven or rapidly varying F1 illumination (ionization) seems to result in spread F.

Equatorial spread F is often preceded by a rapid rise in $h_{max}$. Field aligned rods or sheets of abnormally low electron density form low in the ionosphere and rise rapidly (100 km/sec). As they rise, these "underdense plumes" spread north and south along geomagnetic field lines. Underdense plumes seem to form 2 hours or less east of the sunset terminator, possibly due to the abrupt reversal of the equatorial electrojet (and associated E x B forces) near the terminator. Initially, they drift rapidly (500 m/sec) westward as they rise and increase to maximum size. They then drift more slowly (100-200 m/sec) eastward. A fully developed plume actually approximates a flattened cylinder along a geomagnetic field line. The largest density discontinuities (by comparison to ambient value) occur between 225-450 km altitude, but they may extend to an altitude of 1000 km. The fully developed anomaly is 50-300 km thick (vertical), and extends 100-1500 km east-west and 2000 km or more north-south. It will persist for 2-3 hours. Similar structures are observed to form after midnight, often in association with increased geomagnetic activity.

Several mechanisms have been proposed to explain equatorial spread F in addition to the reversal and damping of the equatorial electrojet. The rapid transit through overhead flux tubes may create an area of relatively "low pressure" near the magnetic equator at higher altitudes. Particle precipitation near the two anomalies, gravity waves, and the sharp radiation discontinuity across the terminator are also suggested causes. At subequatorial latitudes, spread F is observed to occur simultaneously at magnetic conjugate points.

Two types of spread F are identified in terms of their effects on an HF probe signal (such as a vertical incidence ionosonde). Frequency spread F is most common at lower altitudes. Different frequencies appear to be reflected from the same height. Of course, "height" is really virtual height--a distance predicated on travel time of the wave to and from its reflection point. Frequency spread F may result from very strong electron density gradients (bubbles) capable of bending the incident ray, thereby decreasing its angle of incidence. The more nearly normal the wave is to the reflecting medium, the lower will be the maximum frequency reflected. Since refraction varies inversely with frequency, higher frequencies could be bent to nearly grazing incidence. By this means, a large range of frequencies might be reflected from the same virtual height.

Range spread F, more common at higher altitudes, probably results from a field aligned corrugation of the electron density. Such a corrugation means that, within a small horizontal distance, a particular electron density may exist at a variety of altitudes. Since the probe beam is of finite size when it hits the ionosphere, a portion of the wavefront will be returned from several different altitudes. Range spread seems most common prior to local

midnight, while frequency spread dominates near local midnight. On geomagnetically disturbed days, a slight secondary maximum in the occurrence of range spread F exists near 0500L. Spread F is most common near the equinox (probably due to the greater day-night disparity in ionizing flux and the shorter twilight). It is least common near the solstice.

Spread F is an unusual occurrence in the middle ($25^{o}$-$55^{o}$ geomagnetic) latitudes. When observed, it is most common after local midnight and during the summer. Geomagnetic substorming produces a 20% increase in early morning spread F. During these periods, the spread F is observed to form and drift westward from the 0300L meridian.

Above $55^{o}$, spread F is relatively common. It is a nearly permanent feature of the quiet polar ionosphere. With quiet geomagnetic conditions, spread F is nearly twice as intense in the summer as it is during the winter months. It seems to increase with geomagnetic activity, darkness (more direct access of magneto-tail plasma), and winter. Spread F extends slightly equatorward of the oval in response to the sharp ionospheric gradients common in these regions (e.g., the subauroral trough).

Cosmic rays and precipitating particles are a major source of high latitude F layer ionization. This is particularly true in the winter hemisphere, where solar EM emission has little impact on the ionosphere. The non-uniform nature of cosmic radiation and precipitating particle flux (in terms of areal and temporal uniformity as well as energy spectrum) results in considerable horizontal and vertical variation in the high latitude ionosphere. These variations are thought to account for high latitude spread F. They also explain its peculiar nature. The low density bubbles responsible for spread F lie along the geomagnetic field lines in the auroral oval and seem to be nearly vertical within the polar cap.

The occurrence of spread F is often masked by D region absorption or blanketing sporadic E, but it seems to reach a maximum in a rough annulus which includes the auroral oval, but extends both equatorward and poleward of the oval. Spread F seems to be slightly less common very near the geomagnetic pole. At these latitudes, spread F is often present during both quiet and geomagnetically disturbed times.

## 11.4 Disturbed Ionospheric Variations

A disturbance of the magnetosphere usually results in an ionospheric disturbance. The magnitude of the ionospheric response is apparently unrelated to the magnitude of the geomagnetic disturbance in many cases. An exact analysis is difficult. The effects are dependent on the method of observation, and are thought to result from storm energy injected into the atmosphere and resulting changes in chemical composition (Prolss, 1980). Other models relate physical changes in the magnetospheric/ionospheric system to observed ionospheric effects.

Considerable energy is injected into the auroral oval during a geomagnetic disturbance. A third of the energy is due to precipitating particles. The remainder is due to joule heating by storm-associated currents.
This energy, injected near the turbopause (105 km), influences the chemical composition of the thermosphere. Molecular nitrogen and oxygen bubble up into

the F region and decrease the available atomic oxygen. Atomic oxygen is a major source of photoelectrons. $N_2$ and $O_2$ (by way of $NO^+$, $O_2^+$, and $N_2^+$) combine with available free electrons and reduce ambient electron density. Meanwhile, O is transported into the topside ionosphere. Neutral winds then displace these compositional changes to lower latitudes (in the summer hemisphere and at night) or constrain them to the higher latitudes (during winter and in the sunlit hemisphere). The latitudinal variation is $20^o$ (maximum equatorward expansion) to $60^o$ (minimum expansion). The intensity of the resulting decrease in electron density depends on degree of constraint and the amount of energy injected. Maximum depletion at a point is likely with minimum expansion. (The greatest depletion should occur in the winter daytime.) The compositional changes spread as rapidly as 300 m/sec horizontally. The convective bubbles open the flux tubes in the plasmasphere. Plasmaspheric electron density then drains into the F layer to replace recombination losses and is, in turn, lost to recombination.

Overall, ionospheric electron density will decrease to some minimum value and then slowly recover. Successive storms may cause additional reduction in electron density, but some absolute minimum value will exist at any point. The storm may alter recombination rates but it cannot turn off electron production, since that depends primarily on sunlight. During the recovery phase, which may last a day or less at solar maximum to a month at solar minimum, daytime increases will initially overbalance nighttime loss as the ionosphere refills. After a week or less, diurnal variations will be comparable to quiet monthly extremes.

Similar conditions occur in conjunction with magnetospheric substorms except that the resulting density depletions are typically confined to the summer daylight sector. The depletions are identified with those longitudes so aligned at the disturbance onset and located above $20^o$ geomagnetic latitude. The depressions then corotate with the earth until they dissipate.

The effects of a storm or substorm are strongly dependent on station location and local time at disturbance onset. Table 11.2 compares the effect of a minor geomagnetic storm ($A_p$ near 30) on a middle latitude TEC station by local time and season. In this analysis, the disturbance is assumed to begin near local sunrise on day 1. Notice that the effects identified in Table 11.1 are for TEC. TEC data actually corresponds to the total electron content below 1000-2000 km altitude. Since this effectively excludes much of the plasmasphere, variations are somewhat different than for total ionospheric electron content. TEC may show significant enhancements early in the disturbance because of the addition of plasmaspheric electrons to those present in the lower ionosphere. These increases gradually turn to depressions as the plasmasphere is drained and recombination begins to dominate. The largest enhancements will typically be recorded by stations located in sunlight when the disturbance begins. Depressions typically replace enhancements following sunset. The most intense depressions usually occur between 2200L-0800L.

Table 11.2 Seasonal and Local Time Variations in TEC for a Middle Latitude Station and $A_p$ = 30 (from Mendillo, et.al., 1975).

| | SUMMER | | | FALL | |
|---|---|---|---|---|---|
| | Local Time | Change in TEC | | Local Time | Change in TEC |
| day 1 | 1000-1500 | +10% | day 1 | 1200-2000 | +35% |
| | 1600-2200 | +25% | | 2100-0100 | + 5% |
| | 2300-0500 | -10% | | | |
| day 2 | 0600-1800 | -25% | day 2 | 0200-0600 | -20% |
| | 1900-2200 | -20% | | 0700-1800 | -10% |
| | 2300-0500 | -30% | | 1900-0600 | -40% |

| | WINTER | | | SPRING | |
|---|---|---|---|---|---|
| | Local Time | Change in TEC | | Local Time | Change in TEC |
| day 1 | 1000-1500 | +25% | day 1 | 1000-1400 | -10% |
| | 1600-1800 | +45% | | 1500-2000 | +15% |
| | 1900-0100 | +10% | | 2100-0600 | -25% |
| day 2 | 0200-0800 | 0 | day 2 | 0700-2000 | -20% |
| | 0900-1800 | +10% | | 2100-0500 | -25% |
| | 1900-0700 | -20% | | | |

The effects of an ionospheric storm can be particularly insidious at some locations. Near the auroral oval, a station may see normal conditions during quiet times and during very disturbed times. Conversely, a small disturbance may result in a severe depletion of the overhead electron density. A slightly stronger disturbance may markedly increase overhead densities. The culprit here is the shifting position of the auroral oval and associated trough with level of geomagnetic activity. A slight disturbance may move the oval just far enough equatorward to place our imaginary station under the nighttime trough. A further increase places the station under the auroral oval, and a major storm may suffice to extend the polar ionosphere over the station. Since the polar ionosphere is essentially unaffected by a geomagnetic disturbance, only the normal polar spread F would differentiate this condition from the quiet ionosphere normally recorded at the site.

Low latitude sites may also experience surprising results from a geomagentic disturbance. The equatorial trough may fill slightly and the ridges remain little changed to slightly enhanced. Moreover, the ridges may move slightly poleward or equatorward. Weakening of the high altitude geomagnetic field by the storm ring current slows the vertical transport. This permits the equatorial regions to retain a larger portion of the electron density produced there. Compression of the magnetosphere/plasmasphere system forces additional electrons into the low latitudes from above, but, unlike the high latitudes, there is no heating from below and "bubbling-up" of the materials responsible for increased recombination. The plasmaspheric drainage is thus retained as an enhancement at the low latitudes. (Of course, these enhancements are present primarily during day and early evening hours when the ridge/trough structure would normally exist. They nearly disappear by early

morning.) This enhancement gradually dissipates as the disturbance subsides, and vertical transport (due to E x B) returns to pre-storm levels. The absorption of ring current energy in the equatorial ionosphere hastens the recombination process and subsequent return to more normal densities.

## 11.5 Eclipses and Meteors

Highly localized spatial and temporal changes in ionospheric electron content are possible due to eclipses and meteors. Eclipses and the reccurrent meteor showers are forecast with good accuracy. Unfortunately, their exact impact on the ionosphere is not so easily forecast.

Solar eclipses may occur several times each year. They result when a new moon occurs in or near the ecliptic. Since the sun and moon have almost the same angular size when viewed from the earth, the moon can block solar radiation for as much as seven minutes on occasion. The locations on the earth from which the sun appears totally occulted are said to be in the path of totality. The width of this swath is usually measured in tens of miles. The shutoff of insolation means that recombination will reduce electron densities in the D, E, and F1 layers. Since the event is short-lived, the immediate effects are typically brief and confined to the vicinity of the path of totality. However, the sudden change in electron density may trigger a traveling ionospheric disturbance (TID). Such a disturbance may have much greater impact on ionospheric systems than the eclipse itself, because the TID may affect systems remote from the path of totality and at a significant time after the eclipse.

The immediate impact of a meteor on the ionosphere is also small. Typical dimensions range from $10^{-7}$ to $10^{3}$ grams mass and 40 microns to 8 cm in diameter. The smaller, lighter particles are by far the most numerous. Despite their individually small masses, meteors are thought to encounter the earth at a rate of 5-50 tons per day. As these dust grains (often metallic in nature) enter the earth's atmosphere, they are heated by friction to incandescence. The brilliance of a meteor depends on its mass and velocity roughly as $1/2mv^2$. A portion of this energy produces a dense column of ionization about the meteor. The length and size of these columns (or trails) depends on the energy released and the entry angle of the meteor. Lengths may range up to 50 km but probably average 10-20 km for the "average" meteor. Trail radii range from 0 to about 1.2 m optically and 0.5 m - 4.4 m on radar. The smaller trail occurs at lower altitudes where the greater neutral density limits the effective range of the energy released (only a certain number of particles can be ionized, and this number fits into a smaller space at lower, more dense altitudes). Recombination and winds quickly dissipate the trails. Lifetimes of less than a second are common, although a larger meteor's trail may persist for up to a minute. Trail durations approaching an hour have been reported, but they are are exceedingly rare.

Literally millions of meteors occur each day, and, while their trails are short-lived, the resulting metallic ions (from the meteor itself) seem to have very long lifetimes (in comparison to those produced by solar radiation) at E layer altitudes where they are deposited. Sporadic E is thought to be one consequence of dense accumulations of this meteoritic debris.

Largest accumulations of debris in an area usually result from meteor showers as opposed to sporadic meteors. Meteor showers probably result from the earth's passage through a dust cloud in solar orbit. These dust clouds are thought to result from the break up of comets. Indeed, several meteor showers seem to be associated with known cometary orbits. Unlike random meteors, meteor showers are predictable and tend to produce as many as ten times the number of meteors per unit time as are observed outside of showers. Several major showers are recognized, and Table 11.3 records their average occurrence.

Table 11.3 Major Recurrent Meteor Showers (After Allen, 1973).

| Stream | Maximum | Period of Visibility | Average Hourly Counting Rate |
|---|---|---|---|
| Quadrantids | 3 Jan | 2 - 4 Jan | 30 |
| Lyrids | 23 Apr | 20 - 22 Apr | 8 |
| Eta Aquarids | 4 May | 2 - 7 May | 10 |
| Delta Aquarids | 30 Jul | 20 Jul - 14 Aug | 15 |
| Perseids | 12 Aug | 29 Jul - 18 Aug | 40 |
| Orionids | 21 Oct | 17 - 24 Oct | 15 |
| Taurids | 4 Nov | 20 Oct - 25 Nov | 8 |
| Leonids | 16 Nov | 14 - 19 Nov | 6 |
| Geminids | 13 Dec | 8 - 15 Dec | 50 |
| Ursids | 22 Dec | 19 - 23 Dec | 12 |
| Arietids | 8 Jun | 29 May - 17 Jun | 40 |
| Zeta Perseids | 9 Jun | 1 - 15 Jun | 30 |
| Beta Taurids | 30 Jun | 23 Jun - 7 Jul | 20 |

Major showers are named for the constellation from which they appear to emanate. This point in space (found by extrapolating the meteor trails backwards to their apparent origin) is called the shower radiant.

Although major showers occur throughout the year, the greatest shower density occurs between April and September. Theoretical and observational analyses both reveal a marked diurnal variation in both shower and random meteors. Both counting rates and individual brightness are higher between the local midnight-sunrise-noon meridians than in the afternoon and evening sectors. The post midnight and morning sectors are on the leading edge of the earth in its orbit about the sun. Here, the earth sweeps up meteors, adding its revolutionary velocity to that of the meteor. On the trailing portion of the earth, meteors must catch up with the earth in its orbit. Since the ionization and illumination produced by a meteor are proportional the meteor's velocity squared, the morning meteors have much more effect on the ionosphere.

## 11.6 Summary

Depending on the purpose, the earth's atmosphere may be divided into many different, sometimes overlapping, layers. Since the ionosphere's existence depends on the interaction of particular atomic and molecular species with solar radiation and precipitating particles, we are primarily interested in chemical composition regimes. Below the turbopause, the atmosphere regularly undergoes convective mixing resulting in a fairly uniform mixture of all species. This is the homosphere. Above about 100 km, atmospheric

constituents stratify by weight. Since each interacts differently with incoming EM radiation, the number of photoelectrons produced varies with height. Decreasing density with height and decreasing radiation intensity with depth also work to explain the resulting EDP. Chapman theory consolidates these ideas.

Neutral atmospheric density varies in response to varying solar radiation. Proximity of the sun in December, longer illumination periods in summer, and variations in solar emissions (with flares, 27-day, and 11 year periods) all act to alter the density at a given altitude and the EDP.

A consideration of the primary influences on the ionosphere leads us to consider both latitude and height regimes within the ionosphere. Only the middle latitude ionosphere is found to closely obey a simple ionospheric theory. At high latitudes, magnetosphere-ionosphere interactions play a vital role in structuring the ionosphere. The low latitude ionosphere is similarly dominated by the equatorial electrojet and a nearly horizontal geomagnetic field.

The vertical structure of the ionosphere is equally diverse. Although greatest electron densities are found in the F2 layer, it contains no production mechanisms. The topside ionosphere probably contains over half of the ionospheric electron content, maintains the nighttime F layer, and plays a major role in ionospheric storm variations. Despite its importance to ionospheric operations, the topside is poorly observed and even more poorly understood.

Ionospheric storms fare little better, perhaps because of a lack of understanding of the topside. Several theories seem to explain many storm time observations in terms of chemical composition changes. A good model for the magnitude and detailed effects of a storm is still lacking. The full understanding of these ideas must await more complete observation of ionospheric storms and a better definition of what constitutes a "quiet" ionosphere.

# CHAPTER 12

# IONOSPHERIC OBSERVATIONS

Any attempt to analyze the state of the ionosphere and forecast its variations is critically dependent on our ability to accurately observe the ionosphere. Ionospheric measurements have been made since the early 1930s using a variety of instruments. The vertical incidence ionosonde was one of the first and is still among the most widely used instruments. The oblique ionosonde is similar to but more complicated than the vertical incidence ionosonde. Other ground-based instruments include the riometer and the polarimeter. Of these, only the polarimeter is not self-contained. Rather, it depends on a satellite-borne transmitter(s).

The advent of spacecraft permitted measurement of the topside ionosphere. While the polarimeter requires transionospheric signals, other instruments do not. These include the active topside ionosonde, the passive or breakthrough frequency monitor, and the plasma sensor. Topside observations used in conjunction with ground-level bottomside measurements provide the opportunity to observe a complete electron density profile (EDP).

## 12.1 Vertical Incidence Ionosonde

The vertical incidence (VI) sounder is actually a pulsed, variable frequency radar. It generally has a narrow beam-width and is amplitude-modulated (AM). Amplitude modulation permits unique identification of a given pulse. This, in turn, permits timing the travel (out and back) of a given pulse. Knowing the general characteristics of the atmosphere, it is possible to calculate the average speed of an electromagnetic wave. The target's range, or altitude, can then be calculated. The VI sounder's target is the electron density overhead, and heights measured using signal delay time are known as virtual heights. It is important to note that virtual heights may bear little resemblance to actual or true height.

### 12.1.1 Operational Theory

The idealized output of the VI sounder is a profile of virtual height versus frequency. The highest frequency reflected from a given height depends on the electron density at that altitude. In fact, this frequency will be the critical frequency ($f_o$) for that altitude. Since electron density increases, in general, with altitude, we might expect that higher frequencies will be reflected from higher altitudes. This will continue until reflection occurs from the altitude of peak electron density in the overhead ionosphere. Higher frequencies will be transmitted into space. Consequently, the VI sounder measures only the bottomside ionosphere (i.e. that below the F2 layer peak density which occurs at a height known as $h_{max}$). Frequencies typically employed range from 1-15 MHz.

Unfortunately, several details cannot be disregarded. The ionosphere is dispersive (phase speed is a function of frequency). This means that the virtual height measured from time delay is not a smooth function of frequency even if the ionospheric electron density varies smoothly. As the reflection

height approaches the peak density height of each ionospheric layer (probe frequency approaches layer critical frequency) the signal experiences increasing retardation. The result is a cusp at the peak of each layer on an ionogram --the virtual height of a layer is noticeably higher than the true height of reflection of the critical frequency for that layer.

The continuously varying EDP also acts to refract the probing beam. In the absence of significant ionospheric tilts, the problem is not large. Effectively, this means that the VI sounder seldom measures the ionosphere directly overhead. Refraction generally decreases with increasing frequency. Moreover, it is independent of signal amplitude. In the absence of collisions and a magnetic field, the index of refraction is

$$\mu = (1 - k\,(N/f^2))^{1/2}$$

where

$$k = \frac{e^2}{4\pi^2 \epsilon_0 m},$$ and

e = electron charge,
m = electron mass,
f = frequency, and
n = number density.

The phase velocity (v) is proportional to the speed of light (c) divided by the index of refraction;

$$v = \frac{c}{\mu}$$

If measurements are made on the terrestrial ionosphere, the presence of a magnetic field cannot be neglected. It causes the electromagnetic wave to split into two (and, occasionally three) components when it enters the ionosphere. The sounder's beam is typically linearly polarized in a random (with respect to the local geomagnetic field) direction. The geomagnetic field splits this linearly polarized wave into two, such that the polarization vector of one wave (the ordinary wave) lies parallel to the ambient magnetic field vector. The second (extraordinary wave) is polarized perpendicularly to the magnetic field. The extraordinary wave interacts with the geomagnetic field (Lorentz Force) and so suffers increased absorption with respect to the ordinary wave. This interaction also alters the effective index of refraction, hence phase velocity, for the extraordinary wave. The end result is that a given layer (certain electron density) will generally reflect a higher frequency extraordinary wave than ordinary wave. The extraordinary return will also be weaker and less dependable (i.e., more susceptible to small variations in electron density, angle of incidence, etc.) than the ordinary wave. At high geomagnetic latitudes, a third (z wave) component is

also produced. If the critical frequency of the ordinary wave, $f_o$, is much greater than the gyro frequency, $f_g$, it will be related to the extraordinary wave critical frequency, $f_x$, by the equation

$$f_x - f_o = \frac{f_g}{2}$$

### 12.1.2 Ionogram Interpretation

A schematic ionogram is shown in Figure 12.1. Several features are immediately apparent. The trace is height vs frequency. Figure 12.1 shows both the ordinary and extraordinary waves. Closer inspection reveals the presence of two ionospheric layers. These can be inferred from the presence of at least two cusps. A midlatitude daytime ionogram will often show E, F1, and F2 layer returns. The F1 return generally disappears at night. Secondary echoes may result from a portion of the wave energy reflecting ionosphere-ground-ionosphere-VI receiver. The higher altitude for the secondary echo is, of course, a relic of the longer delay time. Additional reflections are occasionally present and indicate a strong, stable ionosphere. The greater the stability the more nearly identical will be the primary and secondary traces in shape and frequencies (but not intensity) of the features observed.

Several parameters are routinely determined from hourly ionograms. Figure 12.2 identifies these parameters on a theoretical ionogram. Note that the reflected and extraordinary traces have been eliminated for simplicity.

The lowest frequency recorded by the ionosonde is defined as $f_{min}$. $F_{min}$ is a function of a number of parameters, not all of which are ionospheric in nature. Equipment alignment, sensitivity, and power output establish a baseline and a lower limit on detectable frequencies from the ionosphere. Local and atmospheric radio noise may further degrade equipment

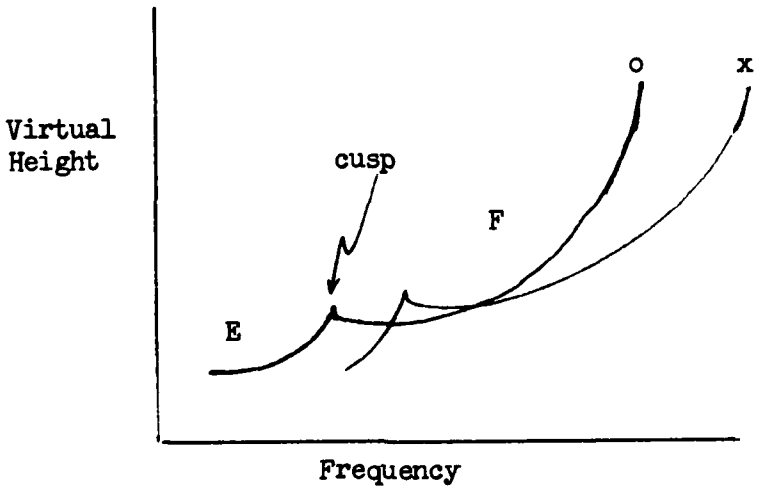



Figure 12.1 Schematic Ionogram Showing the Ordinary (O) and Extraordinary (X) Traces, Cusps, and Ionospheric Layers.

sensitivity. Nondeviative (no path line variation) absorption by the D region is the primary ionospheric influence on $f_{min}$. Significant differences in equipment and ambient noise conditions make inter-site comparisons of $f_{min}$ meaningless for ionospheric analysis. Meaningful ionospheric analysis can be done by comparing day-to-day variations in $f_{min}$ at a given site. Abnormally high $f_{min}$ values are indicative of ionospheric absorption. The extreme condition is total absorption of all frequencies. In this case, no ionospheric returns are present, and an AWS ionosonde observatory would record only the qualifier 2.

While D layer measurements are severely limited, E layer returns are much more descriptive. The frequency at which the ionogram E layer trace first turns vertical (cusps) is known as the critical frequency of the E layer (ordinary wave)--$f_oE$. The horizontal trace extending to frequencies above the $f_oE$ is a consequence of sporadic E. Unlike the normal E layer, it reveals little vertical structure. The maximum frequency at which sporadic E returns are received is the $f_oES$.

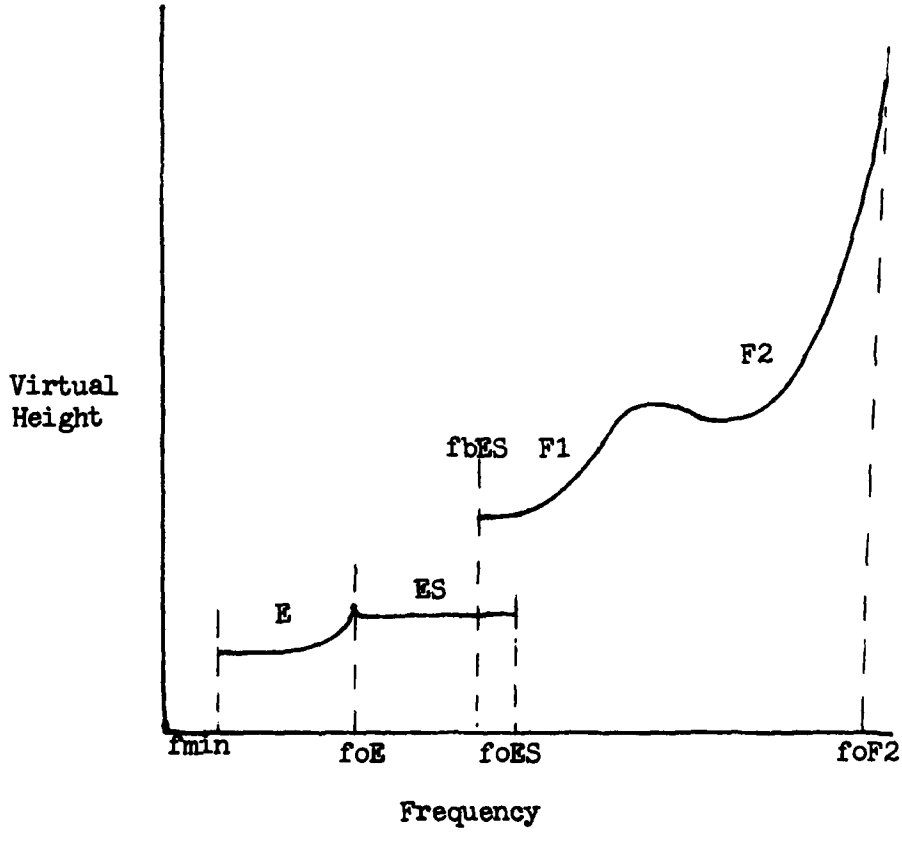



Figure 12.2 Theoretical Ionogram Showing Typical Parameters.

Occasionally, the sporadic E layer electron densities will be sufficiently high to reflect all sounder energy at certain frequencies. No return will be observed from the F region at these frequencies (see Figure 12.3), and the E layer is termed blanketing. The highest frequency at which blanketing sporadic E occurs is $f_bES$. It is not routinely reported. When the blanketing frequency is high enough to preclude returns from the F layer at any frequency, AWS observatories encode a qualifier 1. Under these conditions, the $f_oES$ transmitted may be equivalent to $f_bES$. International observatories may transmit a qualifier of 1 (or 2, in the case of absorption) even when some portion of the F layer return is visible. Blanketing sporadic E is highly variable. An indication of its variability is available by comparing the initial and reflection sporadic E echoes in position and frequency.

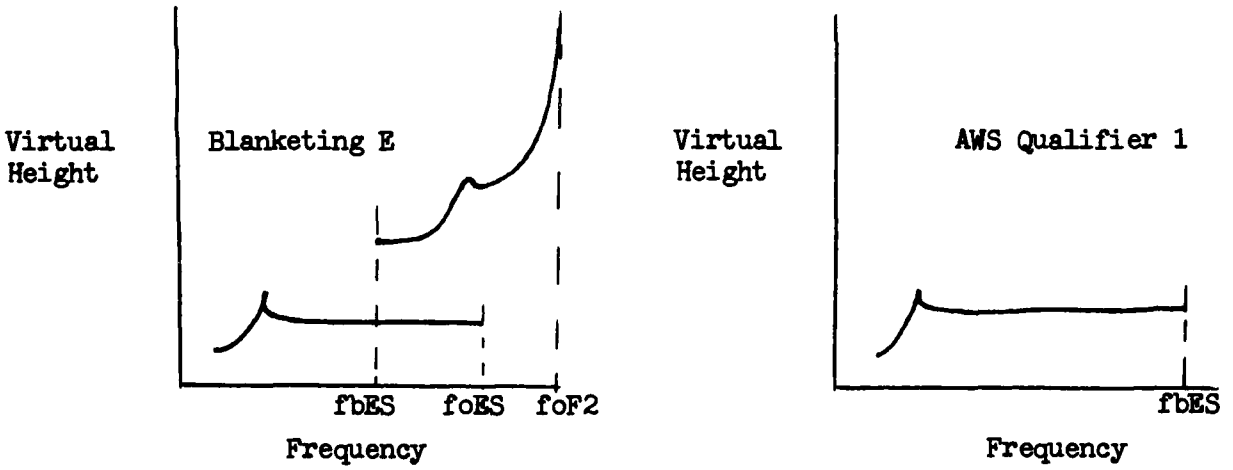



Figure 12.3 Blanketing Sporadic E vs Qualifier 1.

F layer analysis is generally similar to that for the E layer. Frequencies defined at the F1 and F2 cusps are the $f_oF1$ and $f_oF2$ (ordinary wave), respectively. The two returns can usually be differentiated by the height at which they occur. This is particularly important when F1 densities exceed F2 densities. Under these conditions, no F2 layer returns will occur, and only height considerations will provide a key. A qualifier of 7 is generally used in this case.

The F2 layer critical frequency is determined as the frequency at which the F2 layer return becomes vertical on the ionogram. It is the highest frequency reflected from the ionosphere at vertical incidence. A different critical frequency is measured depending on which trace (ordinary - $f_o$, extraordinary - $f_x$) is used. The extraordinary wave will generally yield a higher frequency. This permits differentiation of the two waves.

The two waves also sample different parts of the ionosphere. Just as water refracts light, so too does the ionosphere refract radio waves. This makes long range radio communication possible, but it complicates vertical incidence ionosonde measurements. As the radio wave enters the ionosphere, it is bent, or refracted parallel to the magnetic meridian. Moreover, the index of refraction changes with altitude, because the electron density and magnetic field strength change. The greatest deviation from a true vertical path occurs near the height of reflection and is generally greater for the extraordinary wave. For an ordinary wave under quiet ionospheric conditions, the deviation may be as great as 50-60 km for frequencies near the $f_oF2$. The deviation falls off rapidly above the $f_oF2$. Theoretically, the paths of the upcoming and downcoming waves should be identifical. Electron collisions modify this slightly; the modification becoming most pronounced for frequencies near the gyro frequency.

Scattering or defocusing of the ionosonde signal may also result from tilts or unusual stratification in the F region electron density profile. Such phenomena are particularly common in areas where ionospheric gradients are strong. These include the terminators and the auroral ovals. The result may be deviative aborption--qualifier 9. Figure 12.4 shows two examples of qualifier 9 ionograms. In one case, the F layer cusp simply does not occur. The return just gradually ends. In the second case, several frequencies are returned with the same delay time, and no cusp occurs. This is probably a consequence of oblique reflection and shows a sharp frequency cutoff. Note that qualifier 4 (frequency exceeds equipment limit) is very similar to 9 except for the baseline extension. With condition 4, the F layer trace extends off the high frequency edge of the ionogram without cusping. Comparison of the features of non-deviative (D region) absorption with deviative (F region) absorption suggests the basic difference: non-deviative absorption degrades a certain frequency range (effect is frequency dependent) while deviative absorption (resulting from scattering or defocusing) affects a given height range. Deviative absorption generally results from a geometry consideration, while non-deviative absorption is a function of electron density.

A height-related phenomena is the Lacuna (see Figure 12.5). It is not encoded, but is important ionospherically. This type of trace identifies an EDP discontinuity or bubble developing within the ionosphere. Such bubbles are thought to be one cause of spread F.

The F layer is also susceptible to spread F. Figure 12.6 provides examples of the two primary types of spread F--range and frequency spread. Routinely transmitted VI observations do not distinguish between the two, but their impact and occurrence are different.

The virtual heights of the electron density peak of each layer are also recorded. It should be recognized that, no matter how it is done, these values will be approximate. The typical height parameter is the ratio of the 3000 km MUF (MUF 3000) to the $f_oF2$. This ratio is known as the MFAC, M3000, or multiplicative factor. Figure 12.7 shows the relationship between the virtual height of the F2 peak density ($h_{max}$) and the MFAC. The MFAC is determined by identifying the highest frequency transmission curve tangent to the F2 layer return. A sample transmission curve overlay is shown in Figure 12.8. It is used by aligning it on the ionogram and, if necessary, interpolating to identify the appropriate transmission curve. Dividing the transmission curve frequency by the $f_oF2$ yields the MFAC. The relationship between $h_{max}$ and MFAC shown graphically in Figure 12.7 is:

$$h_{max} = \frac{1490}{MFAC} - 176$$

where $h_{max}$ is in kilometers.

It does not apply to the F1 layer. The transmission curves presently in use are derived from this relationship. It is a result of a 1942 compromise, and is not universally used. Several stations today use different relationships. This accounts for some station-to-station differences observed in comparing international data to AWS data. Note that for frequencies below the critical frequency, a given transmission curve may intersect the ionosonde return at two or more points. These points indicate various propagation modes are possible at oblique incidence. There is only one possible mode for the MUF.

Vertical incidence ionograms can provide a detailed analysis of the bottomside ionosphere. Their limitations in this realm are primarily due to their reliance on virtual as opposed to true height of reflection. Then, too, ionograms require considerable interpretation. While they adequately define the F and, to a lesser degree the E regions, little information is available on the D-layer.

## 12.2 Oblique Ionosondes

It is not always possible to locate an ionosonde directly under the control point of a given HF path. This is particularly true for systems like the Over-the-Horizon (OTH) radars or ships at sea. Oblique sounders, either forward or backscatter, are used in this and similar cases. An oblique, backscatter system is similar to a VI sounder, but the beam is adjusted to strike the ionosphere at oblique incidence. Transmitter and receiver are collocated. The forward scatter oblique sounder is essentially an HF communications system. The transmitter and receiver are located at opposite ends of the path to be sounded and are kept in phase electronically (hopefully). The Navy's Prophet system is such a sounder.

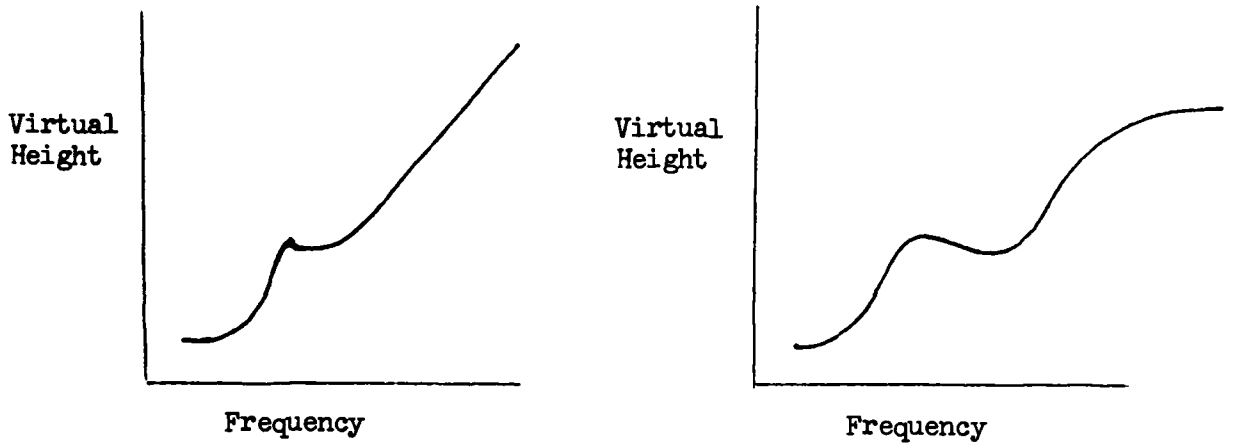



Figure 12.4 Two Types of Deviative Absorption.

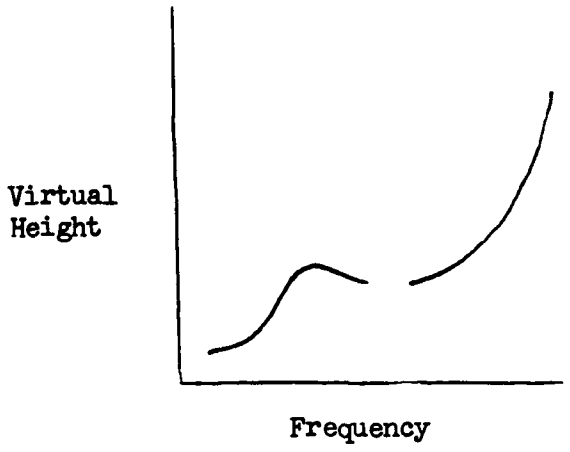



Figure 12.5 Lacuna Due to Mid-level Irregularities.

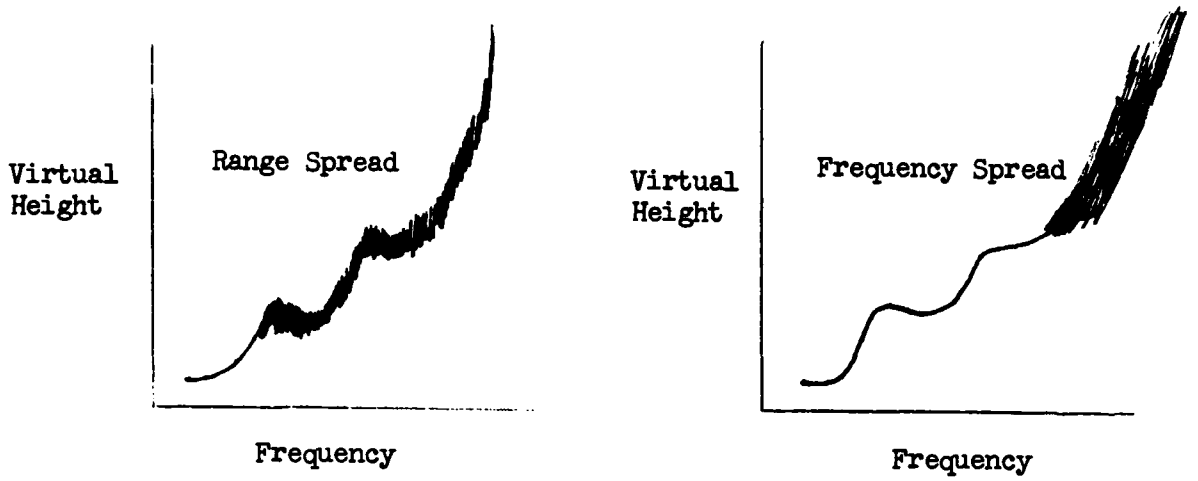



Figure 12.6 Range vs Frequency Spread F.

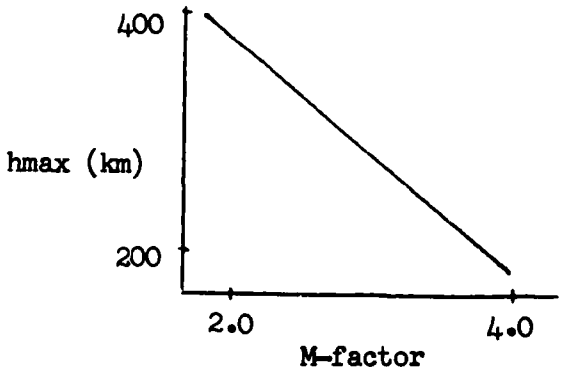



Figure 12.7 Relationship between MFAC and $h_{max}$.

### 12.2.1 Theory of Operation

Oblique sounders provide highly accurate propagation information on specific paths in addition to the information provided by a VI sounder. The oblique ionosonde's advantage is that its beam follows the same path as an HF system operating over this region--ionospheric anomalies have similar effects (ideally) on both. This is a distinct disadvantage when attempting to combine oblique and VI data to model a portion of the ionosphere. Path vagarities introduce considerable uncertainty as to the exact location for which the oblique ionosonde data is applicable. This problem is particularly severe near the terminators and the auroral oval. The oblique ionogram can provide

considerably more information than the VI ionogram, if properly interpreted. Extrapolating the data to another path may not, however, be valid.

The maximum frequency which will reflect from a given altitude in the ionosphere (a given electron density) depends on the signal's angle of incidence (as well as the electron density, of course). The minimum value ($f_v$) will occur at vertical or normal incidence. For the $h_{max}$ level, we earlier termed this the $f_oF2$. Higher frequencies can be reflected (refracted) at oblique incidence. The oblique frequency ($f_{ob}$) is related to the angle of incidence, j (= 90° for vertical incidence), by

$$f_{ob} = f_v \sec j.$$

Notice that a given value of $f_{ob}$ can be produced in several ways by various combinations of $f_v$ and j. In other words, a given frequency can propagate by several different modes. The more nearly vertical is the ray, the more reflections (hops) the ray will require to go from transmitter to receiver. The more hops (greater j) a ray requires, the higher the altitude in the ionosphere from which the signal is reflected. (The use of transmission curves (Figure 12.8) on a VI ionogram reveals several possible modes for a given frequency by the intersection of a transmission curve with the ionogram at several different points.)

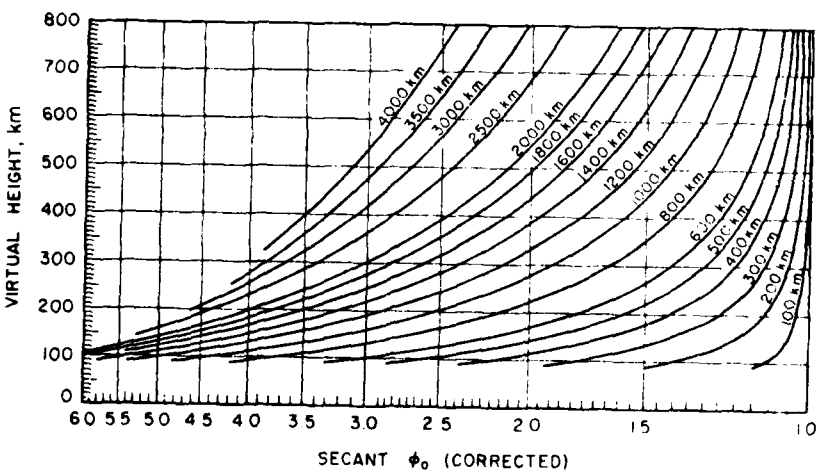



Figure 12.8 Transmission Curves (Davies, 1965).

## 12.2.2 Oblique Ionograms

Figure 12.9 compares a VI and a forward scatter oblique ionogram. The oblique ionogram also provides a plot of virtual height versus frequency. The letters on the ionograms provide a comparison of points--A corresponds to A', etc. The higher traces on the oblique ionogram are not echoes. Rather, they result from multihop paths over the same circuit. This (in Figure 12.9) comparison assumes that the VI sounding is made at the control point of the one-hop path. For this reason, there is no comparison between the VI ionogram and the multihop mode for the oblique ionogram. Notice also that the maximum frequency on the oblique ionogram is not reflected from the height of maximum electron density but from a lower altitude. Moreover, the reflection altitude decreases with increasing obliquity. The extreme is propagation tangent to the ionosphere--line of sight. The highest frequencies will propagate at the tangent point. The other extreme is that of the VI sounder-- vertical incidence. This leads to additional terminology for oblique ionograms.

This terminology is shown in Figure 12.9. The critical frequency of the F2 layer is related to the maximum electron density and 90° incidence angle. The lowest observed frequency (LOF) depends on the mode (layer used for reflection), number of hops, equipment characteristics, and can be qualified. Without qualification, the LOF is conceptually similar to the $f_{min}$. The MOF (maximum observed frequency) is the highest observed frequency on the trace. It, too, can be qualified by the number of hops and the layer from wich it is reflected (E, F, F2, etc.). The junction frequency (JF, in general, but may be qualified) has no direct analog on the VI ionogram. It is the classical MUF for the path. Notice that the maximum observed frequency is not *reflected from the height of maximum* electron density, but from a height which decreases with increasing obliquity.

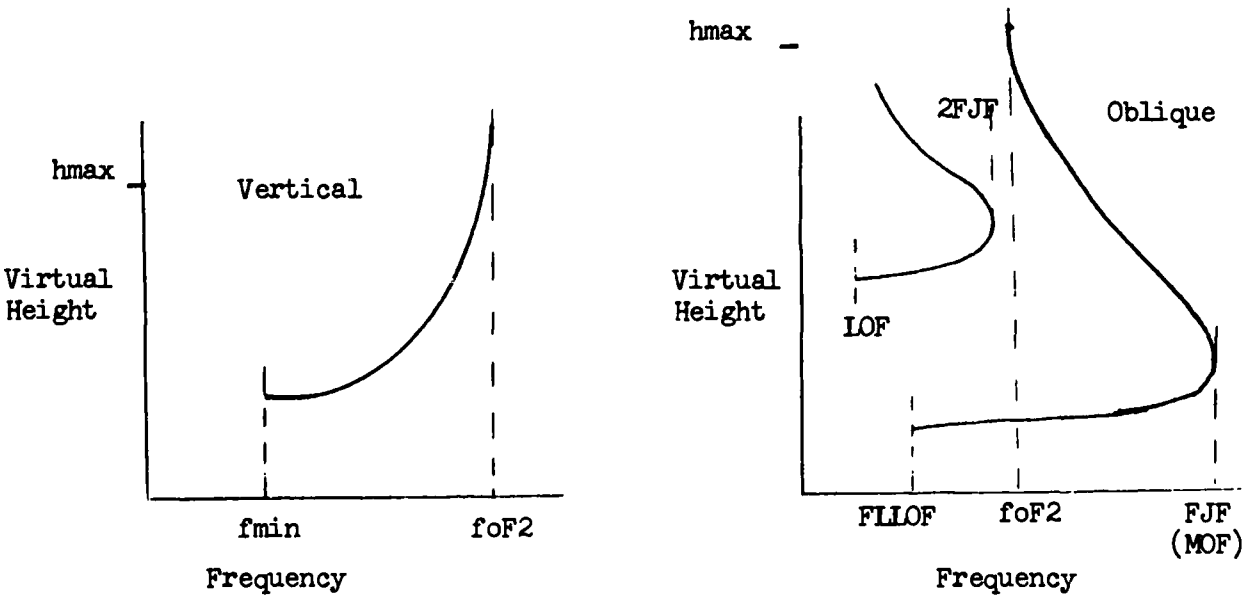



**Figure 12.9 Corresponding Vertical and Forward Scatter Oblique Ionograms.**

The junction frequency may be less than the MOF due to ionospheric scattering. The difference may be as much as 8% during the winter daytime at middle latitudes and on equinox evenings near the magnetic equator. Although both MOF and LOF are somewhat equipment dependent, the so-called MOF extension seems to correlate with times of ionospheric irregularities (spread F, for instance--notice the times when each is most common). Technically, the junction frequency for a given mode is the frequency at which the high and low angle rays join. Only one mode is possible at this frequency. For frequencies below the MUF (or JF), two (or more) different angles of incidence may be used to produce the same $f_{ob}$ from a given ionospheric layer. The extremes (maximum and minimum angles of incidence) for given layer and hop combination (1 hop, F2) define the high and low rays, respectively. Both rays, and many in between, are usually present in ionospheric radio propagation. Sometimes they are more significant than others. As can be seen from Figure 12.9, the high ray is reflected from a higher altitude and employs a greater angle of incidence than the low ray. Figure 12.10 is a pictorial representation of the high and low ray phenomenon for the F1 and F2 layers.

The letters H and L are used to differentiate between High and Low rays. Several examples are:

1. 2F2LOF = lowest observed frequency propagated by two reflections from the F2 layer.

2. 2F2MOF = highest observed frequency in a two hop F2 trace.

3. F2HLOF = lowest observed frequency of the high angle ray via a one-hop F2 path. H and L have no meaning with respect to MOF or JF. The number of hops and the reflecting layer do.

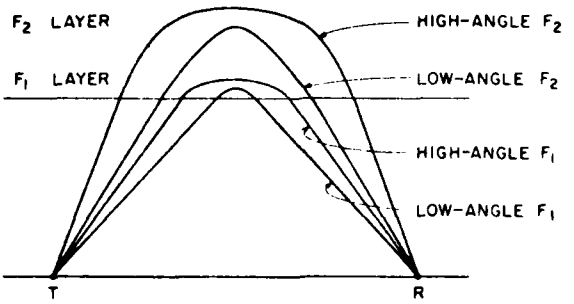



Figure 12.10 High and Low Ray Paths (from Davies, 1965).

Blanketing sporadic E and spread F are also apparent on an oblique ionogram. ES (sporadic E) reflections will occur at a lower altitude than F1 or F2 layer reflections. Since an ES layer is typically very thin, no bending (or refraction) occurs in this layer, and there is only one ray for a given number of hops (angle of incidence). In other words, there are no high and low mode ES rays. Hence, there is no JF. Spread F is apparent on F layer paths of an oblique ionogram in much the same manner that it appears on VI traces. The oblique sounding usually affords a more complete interpretation of the phenomenon:

As with the VI sounder, the oblique instrument provides little information on the D region. The D region provides the ionospheric input to the LUF and LOF, but equipment parameters make it difficult to extract the D region effects from everything else.

## 12.3 Riometer

The relative ionospheric opacity meter, riometer, is intended to sense variations in D region electron density. The measurement is indirect and somewhat low in sensitivity due to the method of operation and dependence on a quiet day curve (QDC). It is used to define the occurrence of a PCA (polar cap absorption), and auroral zone events such as the REP (relativistic electron precipitation) and AZA (auroral zone absorption). For pure D layer analysis, VLF equipment has considerably greater precision. The riometer is cheaper and simpler to use and so continues to experience widespread use.

### 12.3.1 Theory of Operation

Riometer observations are based on the assumption that radio noise reaching the earth from a given direction in space is a constant. This means that galactic noise recorded at the earth's surface should be a function of sidereal time (technically, the local hour angle of the vernal equinox; sidereal time is indicative of the direction the station faces into space) and ionospheric absorption. Radio noise should vary slowly throughout the day as the earth's rotation causes the antenna to slowly sweep across the sky. This slow variation in the galactic "transmitter" establishes the theoretical QDC. The diurnal variation for a middle latitude station is about 1.0 dB.

In reality, the QDC is based on the average of many day's observations. This multiplicity is meant to smooth the impact of ionospheric variations. The D and F regions each account for about 1 decibel (dB) of quiet daytime absorption at 30 MHz. Of course, this varies from day to day and hour to hour even in the absence of ionospheric disturbances, since the electron density of these layers varies continuously. Changes in sun angle, solar output, and atmospheric tides are the primary environmental factors. Manmade and equipment noise may also produce day to day changes.

The occurrence of a PCA, AZA, or other disturbance should produce a marked increase in the absorption of galactic radio noise. This results from large increases in the D region electron density. This increased absorption is measured by comparison to the QDC to specify the magnitude of the event. The QDC is of major importance in event definition. Care must be taken to ensure that large absorption events do not unduly influence the QDC itself. This is usually done by basing the QDC on a very long time period. At the other

extreme, seasonal variations limit the maximum useful time period over which to build QDCs. Too long a period smooths seasonal effects and results in seasonal variations causing unusual changes. Too short an averaging period results in undue susceptibility to disturbances, equipment vagaries, and local noise problems. The absorption, in dB, is calculated by comparing the quiet and disturbed signal strengths, $I_q$ and $I_d$, respectively.

$$\text{absorption} = 10 \log (I_q/I_d).$$

12.3.2 Equipment Design

The typical riometer is a wideband, HF receiver with a vertically-aligned antenna. Since it is designed to monitor galactic radio noise, it must operate at frequencies above the F layer critical frequency. Frequencies between 30-50 MHz are common. The use of such high frequencies, while necessary, severely limits the D region sensitivity of a riometer. Nondeviative absorption is inversely related to frequency, and 30-50 MHz are considerable higher than the D layer critical frequency. While the riometer may have an instrumental accuracy of $\pm$ 0.1 dB, it is sensitive only to moderately large changes in D layer ionization. The physical beam width of a typical riometer at the D layer is tens of kilometers. This further limits instrument sensitivity by averaging out ionosphere irregularities of comparable or smaller scale. Interference from nearby radio emitters can also hamper a riometer and affect the validity of the QDC. A partial solution results from using a wide frequency bandwidth (100 kHz or more) and measuring in the quietest portion of the band.

12.4 Polarimeter

Two somewhat different devices are loosely termed polarimeters. The dual frequency polarimeter and the Faraday rotation polarimeter are both used to measure the total electron content (TEC) of the overhead ionosphere. Only the Faraday rotation instrument is actually a polarimeter, and it is somewhat less accurate as a TEC measuring device. It is, however, considerably cheaper and simpler to operate.

Both instruments depend on one or more transmitters located above the ionosphere. Since the signal must pass through the ionosphere with minimal absorption, frequencies well above the F layer critical frequency are required. VHF (30-300 MHz) frequencies are commonly used. This frequency range permits probing the ionosphere while experiencing minimal absorption. Columnar TEC is observed, and values are typically in the range of $10^{16}$-$10^{17}$ electrons/meter$^2$. Geostationary satellites are commonly used, since any other platform would require tracking capability at the ground site. Likewise, the column being measured would never be the same. Using geostationary vehicles means that, for most ground stations, the column will not be vertical, but rather a slant column. For satellite altitudes below 15° (zenith distance greater than 75°), the variation between slant and vertical TEC is excessive. (Most AWS polarimeters are designed so that satellite altitude will not be less than 15°.) In fact, much smaller deviations from vertical often produce significant variability in comparison to a true vertical measurement. Finally, the measurement is not an EDP but simply a content value.

### 12.4.1 Dual Frequency Polarimeter

The dual frequency instrument provides an accurate, unambiguous determination of column electron contents. Since EM radiation speed depends on frequency in a dispersive medium such as the ionosphere, we can measure the electron density by noticing the speed difference of two closely-spaced frequencies. The difference in speed (or retardation) is actually a function of the index of refraction, $\mu$:

$$\mu = (1 - k\,(N/f^2))^{1/2}$$

where k is a constant, N the electron density, and f the frequency. Commonly used frequencies are near 140 MHz and 340 MHz. Any wider spacing makes measurement difficult.

The primary difficulties in dual-frequency measurements revolve around the cost and complexity of a properly instrumented ground site. Then, too, the paucity of suitable satellites (with dual frequency beacons in geostationary orbit) further complicates the matter. Regardless, the dual frequency measurement yields a slant TEC directly.

### 12.4.2 Faraday Rotation Polarimeter

The more commonly employed Faraday rotation polarimeter determines TEC indirectly. It measures changes in (NOT the absolute value of) the polarization of the satellite beacon's signal. Geostationary satellites with linearly polarized VHF beacons are used in conjuction with a ground-based crossed-Yagi antenna (see Figure 12.11). Such a system is capable of sensing polarization changes as small as 0.1%. Unfortunately, such a system responds only to changes in TEC which occur below about 1000 km. This is a consequence of the weaker magnetic field at higher altitudes and its importance to the polarimeter's operation. Plasmasphere variations are not accurately sensed and can only be estimated. The resulting errors may range from perhaps 5% near midday to nearly 50% at night.

For a linearly polarized E-M wave, we can visualize the polarization vector as lying along the electric field vector (this is perpendicular to the direction of wave propagation). When the wave enters a plasma, free electrons (and protons) will be set in motion by the wave's electric field. This motion will be parallel to the electric field vector (hence to polarization vector). The motion of the free electrons thus defines the wave polarization, and anything which alters the direction of electron motion also alters the wave polarization. If the plasma is permeated by a magnetic field (geomagnetic field in the ionosphere), the Lorentz force provides a means for altering the electron motion.

Actually, two factors are involved in generating Faraday rotation. First, the polarization vector will not, in general, be parallel to the ambient magnetic field at the point where the wave enters the ionosphere. The wave is decomposed into two components: a component with its electric field vector

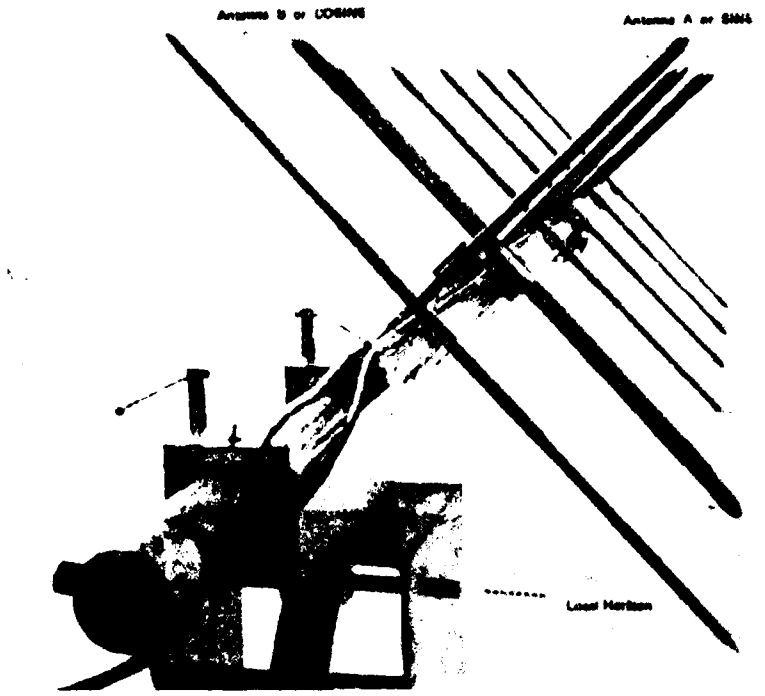


Figure 12.11 Typical Crossed-Yagi Antenna.

parallel to the magnetic field vector (ordinary wave) and a component with the electric field perpendicular to the magnetic field (extraordinary wave). The resulting components are unlikely to be of equal amplitude. The wave polarization at any point along the wave path is the vector sum of the electric field vectors of the two components at that point. This sum will change, because the ionosphere affects these two waves differently. In fact, they experience a different index of refraction, . This results in the two components having different phase velocities, v, since approximately

$$v = c/\mu .$$

By the time the two waves reach the polarimeter antenna and are recombined, they will have been somewhat displaced with respect to one another. The sum of their electric field vectors (polarization) differs from what it was at the instant of their decomposition at the top of the ionosphere. For VHF frequencies, the time delay, hence the variation in polarization is generally small.

A second factor contributes to the rotation in the observed polarization vector. The index of refraction at each altitude in the ionosphere depends not only on the ambient magnetic field (as we've seen above) and wave frequency (the ionosphere is a dispersive medium), but also on the electron density, N, at that altitude. The direction of wave propagation (approximately the antenna to satellite line of sight; it is perpendicular to the polarization vector) will make some angle, , to the ambient geomagnetic field. For a satellite directly overhead a station located on the geomagnetic equator, would be approximately 90°. (The geomagnetic field lines are essentially parallel to the earth's surface here.) The EM wave sets electrons

into motion perpendicular to the direction of wave propagation (a transverse wave) when it enters the ionosphere. For less than 90°, some Lorentz force will exist which deflects the electron path and alters the electric field vector of the E-M wave in the plane of the wavefront. This changes the wave polarization. Since the ordinary and extraordinary waves spend a different time at each level and follow nearly the same paths, both will experience this rotation of their polarization vector. The amount of Faraday rotation resulting, in radians (2 pi radians = 360°), is given by

$$\Omega = \frac{K}{f^2} \int B \cos \theta \; Ndl$$

where $K = 2.36 \times 10^{-5}$ (a constant),
$f$ = wave frequency (in Hz),
$B$ = geomagnetic field strength (in gamma),
and
$\int Ndl$ = TEC along the line of sight (in electrons/$m^2$).

Note that this equation breaks down for ground stations located near the geomagnetic equator ( $\theta$ near 90°).

The polarization of a VHF satellite signal is altered by the earth's magnetic field and the electron density of the ionosphere. Theory permits us to separate out the effects of the geomagnetic field. Observations of polarization yield TEC. The results of wave decomposition are generally minimal by comparison with the impact of the columnar electron density. Interpreting the polarimeter output is the key to effective analysis.

### 12.4.3 Faraday Rotation Analysis

A number of types of polarimeters have been devised. Most AWS observations use an electronically rotated antenna system. Also known as a crossed-Yagi antenna (see Figure 12.11), this system uses two antennas aligned 90° to each other. This permits electronic simulation of four antennas with 45° spacing. Measuring the relative signal strength on each axis permits a determination of the received polarization (with an accuracy of $\pm$ 180°, since the head and tail of the polarization vector are electronically indistinguishable). This system is designed to be mounted on the ground to minimize spurious, ground-reflected polarization changes. A roof mount is feasible only for satellite altitudes in excess of 40°.

The polarimeter produces three output traces (see Figure 12.12): a satellite signal amplitude trace, and two identical but out of phase (90°) polarization traces. The amplitude trace permits monitoring of local interference and ionospheric scintillation. The polarization traces permit accurate monitoring of polarization changes. Two out-of-phase traces are used to simplify analysis when polarization is changing rapidly or reversing direction. Figure 12.13 graphically illustrates the relationship between the polarization trace and the actual polarization vector. Each cross-chart excursion (on the strip chart) represents 180° of rotation in the polarization vector. The similarity in appearance of these excursions to ramps leads to calling them "pi ramps", since 180° = pi radians.

The sample traces show that polarization changes most rapidly near sunrise and sunset. This might be expected, since these are the times of maximum variability in ionospheric electron content. In addition to diurnal variations, TEC may vary by as much as 15% in response to lunar and solar tidal effects on the ionosphere. Normally, these quiet variations are easy to anticipate and interpret. If need be, recorder speed can be increased to spread out a change for easier analysis. Some geophysical phenomena are not so easily predictable.

Figure 12.14 shows a SITEC near the center of both polarization traces. Notice the associated drop in signal amplitude on the top (amplitude) trace. This is a consequence of ionospheric absorption of the VHF signal (similar to an HF SWF). Both result from increased D region electron density caused by flare EUV and X-ray emission. The increase typically occurs over a period of a few minutes, may be as large 15% of background, and may remain elevated for tens of minutes before fading into the background.

The polarimeter is capable of recording changes of as much as a ramp every few seconds if recorder speed is a set high enough (say 5 mm/sec) to permit analysis. The most rapid changes yet recorded are about 1 pi every 30 seconds.

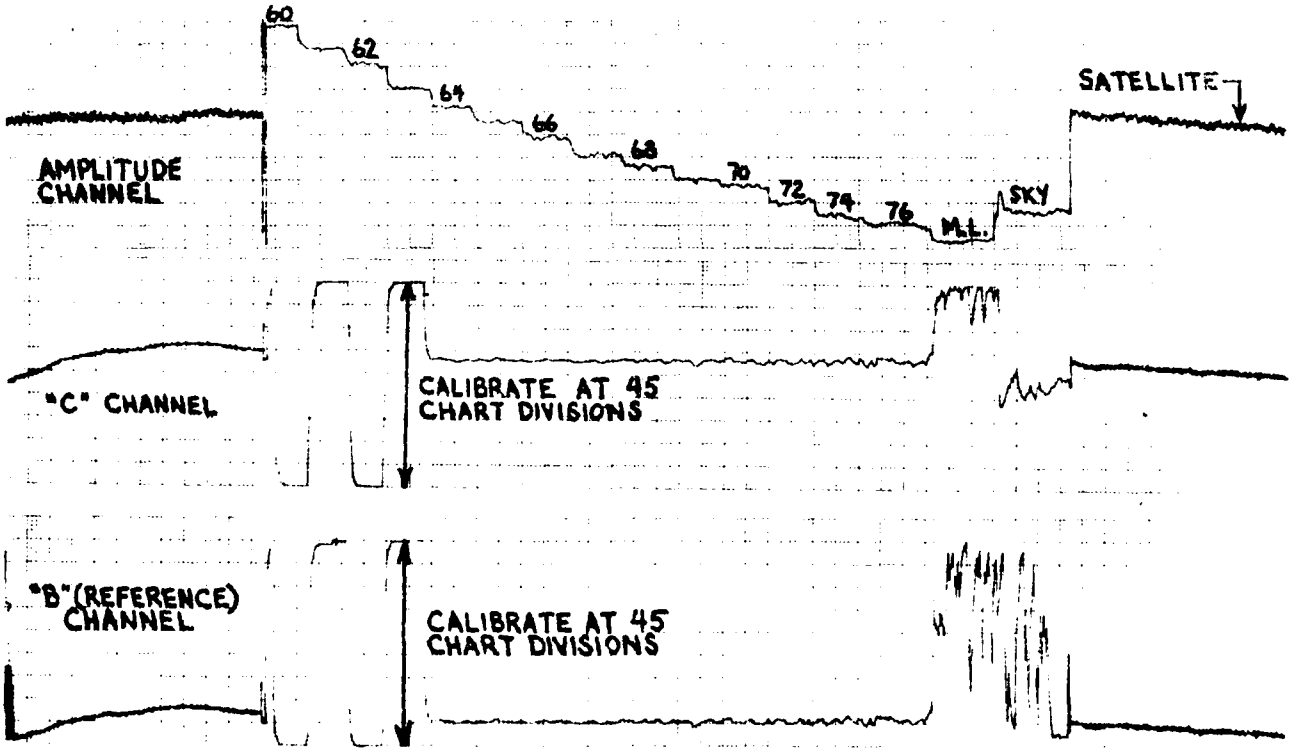



Figure 12.12 Sample polarimeter trace showing amplitude, alternate, and reference channels during calibration. (from Eis, et.al., 1977).

Although diurnal variations and SITECs often produce rapid TEC (hence, polarization) changes, ionospheric storms typically produce the most rapid changes. Figure 12.15 shows the effect of auroral precipitation (associated with a geomagnetic and ionospheric storm) on the Goose Bay polarimeter trace. Note the intense amplitude scintillation visible on the bottom or amplitude trace.

### 12.4.4 Polarimeter Calibration

The Faraday rotation polarimeter responds only to changes in electron content. It is incapable of an absolute measurement of electron density. This fault necessitates a periodic determination of the actual electron density along a polarimeter's line of sight. Ideally, this would be done with a dual frequency polarimeter. Since few of these instruments are available, a more theoretical (somewhat less accurate) method must be used.

The critical or plasma frequency of the F2 layer is a function of the peak electron density in the F2 layer. "Knowing" the value of the peak electron density (from the $f_oF2$), it is possible to infer the electron density profile by making various theoretical and climatological assumptions. From the EDP, we can calculate the total electron content of the ionosphere above the ionosonde at the instant of measurement (it is the area under the EDP curve). Since the impact of electron density on a polarimeter depends only on the total electron content (TEC) and not its distribution with height (loosely speaking), we can equate the effect of the ionosphere to that of a homogeneous slab of electrons between the polarimeter and the satellite.

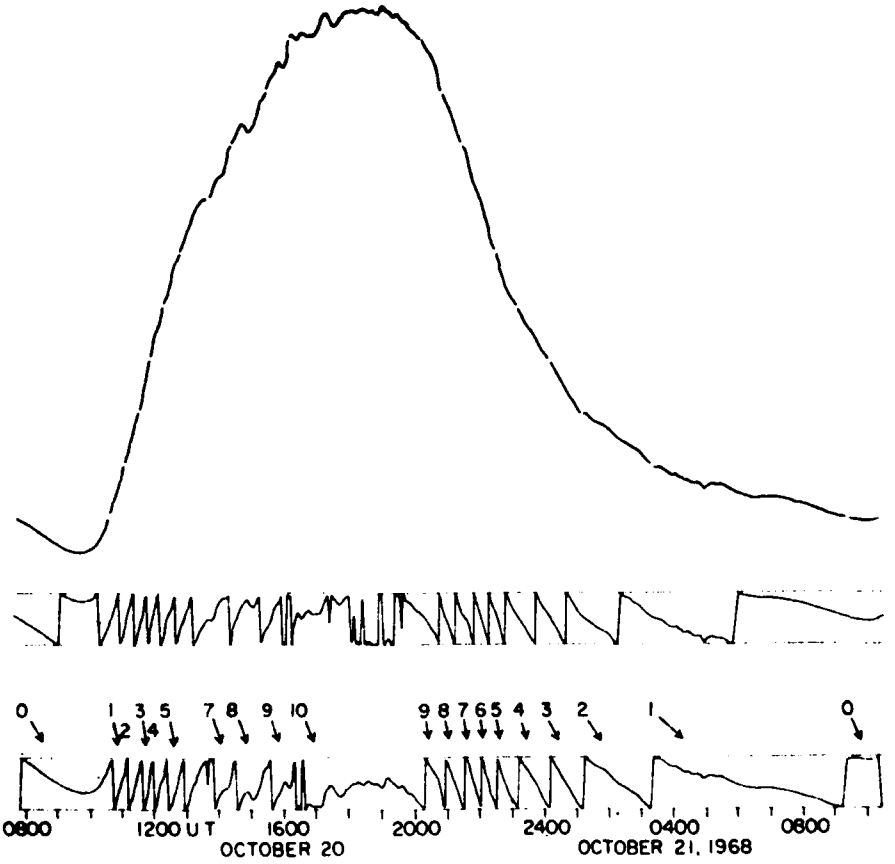



**Figure 12.13** Diurnal variation in polarization comparing a normal graph of a day's change (top) with the two out-of-phase strip chart recordings (bottom). Notice the complementary nature of the out-of-phase traces (from Eis, et. al., 1977).

Theory suggests that the thickness of such a slab should be 200-300 km when electron density is a minimum (shortly before ionospheric sunrise). By definition, the slab thickness, S, is given by

$$S = \frac{\text{TEC (electrons/m}^2)}{N_{max} \text{ (electrons/m}^3)}$$

where $N_{max}$ is the electron density at the maximum point (i.e. in the F2 layer). The critical frequency of the F2 layer is related to the peak electron density by

$$f_oF2 = 9 \quad (N_{max})^{1/2}.$$

Since $N_{max}$ is normally in electrons/cm$^3$, we can correct for the units and substitute observable quantities into the equation for slab thickness. Doing so yields

$$S = \frac{80.6 \text{ TEC}}{(f_oF2)^2}$$

where S is in km; TEC is in el/m$^2$, and $f_oF2$ is in MHz. Figure 12.16 is a nomogram developed from this relationship.

The nomogram is used by selecting an ionosonde measurement of $f_oF2$ representative of the polarimeter column measure and for the appropriate time (minimum electron density). In practice, we attempt to choose an ionosonde

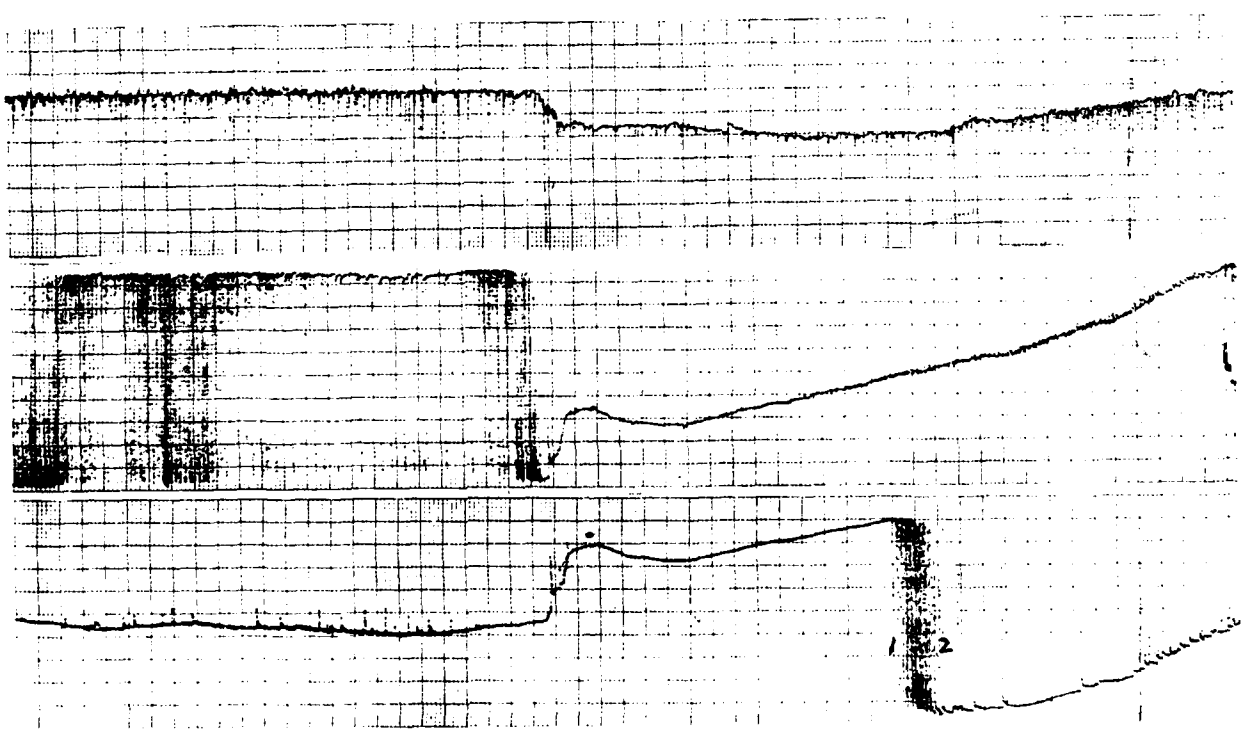


Figure 12.14 A SITEC near center of data record (from Eis, et. al., 1977).

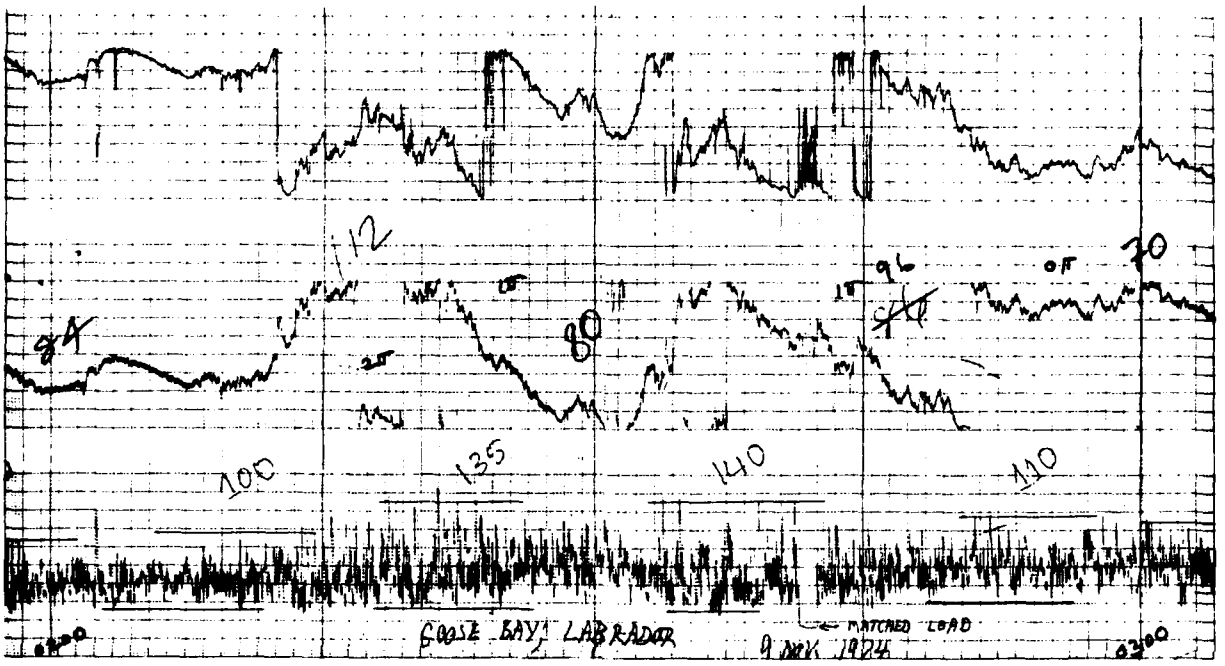



Figure 12.15 Rapid polarization changes and scintillation in response to auroral precipitation (from Eis, et. al. 1977).

which is located below the ionospheric penetration point (IPP) of the polarimeter to be calibrated. The IPP is a theoretical construct. It is the point at which the satellite-polarimeter link passes through an altitude of 300-400 km. This altitude is chosen since the peak electron density in the predawn ionosphere typically occurs at about this altitude. The maximum change in polarization will occur near this altitude.

Since geostationary satellites are used, ionosonde calibration is progressively less accurate for higher latitude polarimeters. (Such paths are progressively less vertical and more slant measurements, while the ionosonde is always a vertical measurement.) Ionospheric variability further complicates this calibration. Tilts may result in the ionosonde measuring the $f_0F2$ of a portion of the ionosphere not directly overhead. Moreover, the ionosphere may so refract the satellite signal as to place the actual IPP some distance from the geometrically determined IPP.

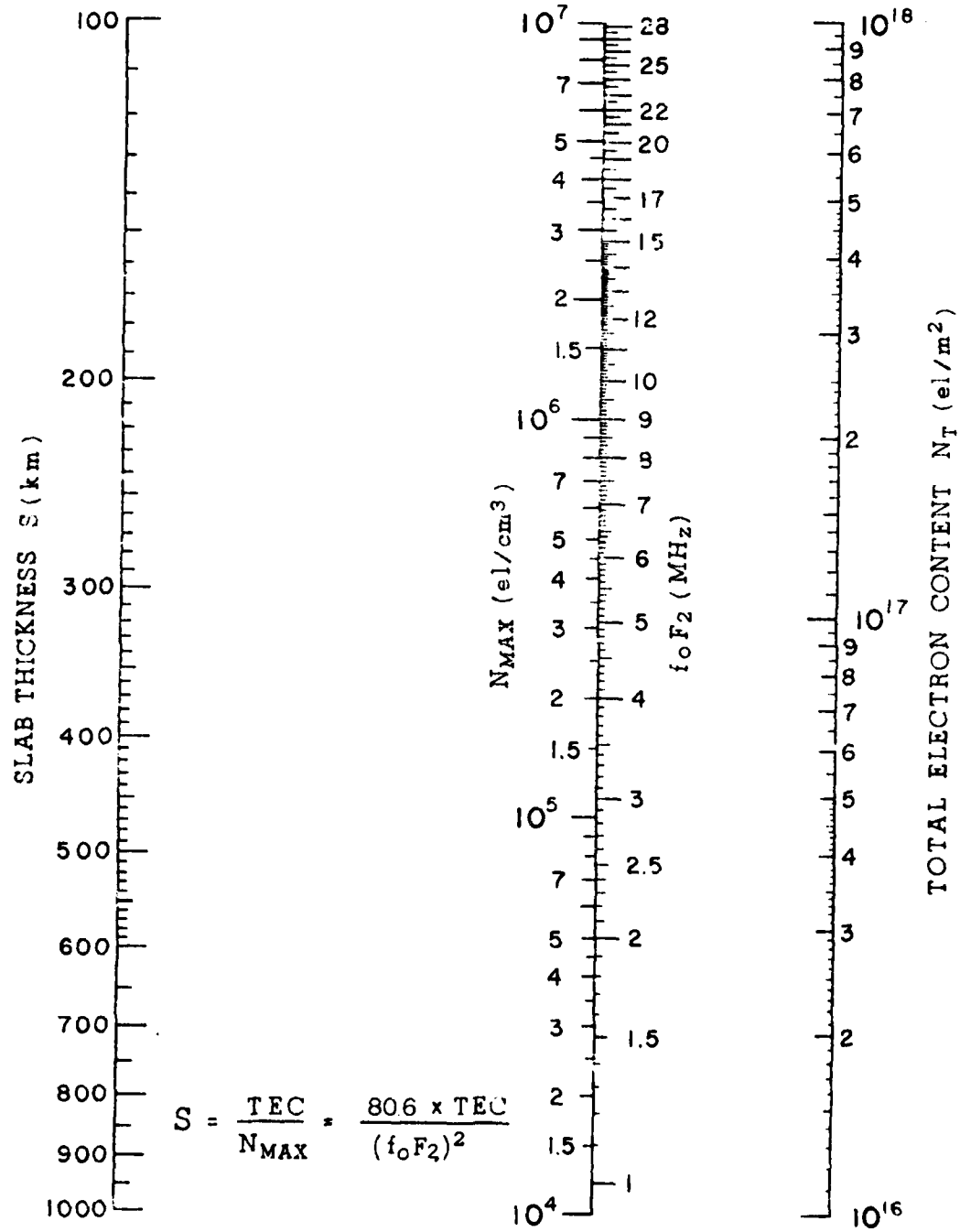



**Figure 12.16 Slab Thickness Nomogram (from Eis, et. al., 1977).**

While ionospheric variability does degrade the accuracy of an already dubious calibration, the sparse ionosonde network provides the largest handicap. Ideally, the calibration ionosonde should be located equatorward of the polarimeter and within $\pm$ 15° of the polarimeter's geomagnetic longitude. In fact, we would like to have the ionosonde within 50-100 km of the actual IPP (this is the wavelength of some ionospheric storms). This ionosonde measurement should be made within 5 minutes of the polarimeter observation with which it is to be used. (Five minutes is the typical minimum period of ionospheric disturbances.) As you might imagine, this host of conditions is seldom met.

Under perfectly quiet, stable ionospheric conditions, ionosonde data can be extrapolated as much as $\pm$ 2900 km in longitude and $\pm$ 1800 km in latitude. If the option exists, the difference in geomagnetic latitude between the ionosonde and IPP should be minimized first, longitudinal differences second. The ionosphere is ordered in geomagnetic coordinates, and most of the large scale variations occur latitudinally. The values given above are averages. Near the auroral zone, much smaller limits should be applied whenever possible. Of course, the sparsity of ionosonde data makes it impossible to unequivocably determine the existence (or absence) of a quiet, stable ionosphere. Again, assumptions must be made based on available ionospheric, magnetospheric, and climatological data.

It would be nice to say that absolute TEC calibrations (note that these are not the same as the routine equipment calibrations performed daily at each polarimeter site) should only be made during absolute ionospheric quiet. Unfortunately, the need for recalibration is greatest during ionospheric storms when ramp changes occur rapidly, and observer errors are most likely. Storm conditions will not, typically, alter the choice of ionosondes for comparison. Such conditions do place severe constraints on the data to be used. The slab concept assumes that the $f_0F2$ is representative of both the $N_{max}$ overhead and a certain theoretical electron density profile (namely a quiet profile with the same $N_{max}$, existing at the IPP at the time of observation).

The $f_0F2$ will not be representative of the overhead or IPP $N_{max}$ when spread F or F2 region (deviative) absorption conditions exist. Rapid M-factor variations suggest potential spread F conditions. If the auroral oval, subauroral trough, subequatorial ridge, or Atlantic or Pacific Anomalies lie over the ionosonde but not over the IPP (or vice versa) the slab technique cannot be used reliably. Ionosonde observations will not be representative of IPP conditions when the ionosonde or IPP is in the vicinity of the sunrise or sunset terminator. Terminator errors cannot be eliminated with certainty even if the data compared are for the same local time (which it must be, as a minimum - zulu time is not relevant unless ionosonde and IPP have the same geomagnetic longitude). This is a result of the variation in sunrise time (ionospheric sunrise, of course) with latitude. Sometimes, it's necessary to interpolate between two observations to get "observations" for the same local time. If there is 7 1/2° longitudinal difference between IPP and ionosonde, they are 30 minutes different in local time (approximately, everything really should be in geomagnetic coordinates and geomagnetic local time but. . .). This may be significant. Likewise, it's necessary to acknowledge the potential error of interpolating between two observations or failing to do so!

The final caveat on TEC calibration deals with ionospheric storms. The forecaster must select times when the $f_oF2$ is truly representative of the overlying electron density profile. This may not be the case during the onset phase of a storm or during rapid storm variations. At these times, the bottomside electron density may be artificially enhanced at the topside's expense. This produces an abnormal electron density profile which cannot be correctly represented by an $f_oF2$ measurement. Similarly, the occurrence of intense sporadic E may significantly alter TEC while producing no effect on the $f_oF2$. Storm displacements of quiescent phenomena such as the subequatorial ridges or the electojets may produce similar effects. Only a careful review of all available ionosonde, polarimeter, and magnetometer data can begin to suggest the feasibility of using the slab technique. Such a review will sometimes reveal tiny, traveling disturbances which can invalidate a calibration on even a truly (otherwise) quiet day. Just checking the $A_p$ and selecting the nearest ionosonde is <u>not</u> sufficient. In fact, it may be counterproductive.

## 12.5 Space-borne Instruments

Of the equipment now available, only rockets and space-borne instruments are capable of consistently measuring the topside ionosphere. VHF satellite beacons used in conjunction with polarimeters have yielded rather general information on topside electron densities. They do not yield information on the density profile. Spacecraft sensors run the gamut from complex sounders designed to provide a full topside profile to passive receivers.

The most complex topside sensor is the active topside sounder. This is essentially a vertical incidence ionosonde directed downwards. It is capable of providing information on density and height down to the peak electron density point. Here, the data must be mated with bottomside measurements to establish a complete profile. When such instruments are flown on other than geostationary vehicles, there is considerable difficulty in determining depth, because the spacecraft's motion relative to the ionospheric segment sampled must be considered. The conversion of time-delay to distance is simpler for a fixed vehicle, but still not easy because of the variable atmospheric density and resulting changes in the ambient speed of light. As a consequence, scale heights often replace true heights in topside soundings. Finally, ionograms must generally be scaled electronically on board the spacecraft. Doing so requires considerable sophistication of spacecraft software.

Passive on-board instrumentation can measure electron density at the spacecraft's altitude. (This isn't much help if the satellite is geosynchronous. Why?) This provides a single point on the topside curve. Combined with TEC, VI data, and a bit of theory, a moderately precise profile can be drawn through the topside ionosphere. Topside measures by spacecraft also have the advantage of using the same piece of equipment to measure many parts of the ionosphere. The data comparability gained from using only one instrument is partially negated by the non-simultaneous nature of these observations. The significant increase in observational coverage remains a distinct, positive feature of spacecraft-type systems.

The simplest of all satellite sensors is the breakthrough sensor. It is simply a multi-frequency receiver intended to determine the lowest frequency just discernable from a ground-based transmitter. This would provide direct measurements of the $f_oF2$ at a number of widely scattered points over a short period of time. It provides no significant information on the topside ionosphere, but would greatly increase the density and extent of available $f_oF2$ data assuming negligible oblique signals.

## 12.6 Summary

Ionospheric measurement is still in its infancy. Sensors are sparse and often not ideally located to optimize their output. Three are of particular importance to SESS: the ionosonde, the riometer, and the Faraday rotation polarimeter. Other instruments, such as the magnetometer, can be used to make general inferences concerning the ionosphere. The goal of ionospheric measurement is to establish an electron density profile continuous in time and location. The VI sounder accurately measures the portion of this profile up to and including the electron density peak, i.e., the bottomside ionosphere. The oblique ionosonde provides similar data over a given path but with less positional accuracy. Ionosonde measurements of the D-region are severely limited, because it does not reflect HF signals. The riometer is a simple instrument designed primarily to detect changes in D-region electron density. While the ionosonde is similar to a radar, the riometer is essentially a radio receiver for cosmic noise. The polarimeter is also a receiver, but it depends on a satellite beacon, typically VHF, for its electromagnetic probe. It measures only electron content and provides no profile information. Although generally restricted to the area below 1000 km by its sensitivity, it does provide limited information about the topside ionosphere. It responds to changes in radio wave polarization caused by changes in electron density. The polarimeter must be calibrated periodically to provide information on total electron content. This is generally done by means of nearby ionosondes and the slab technique. Complete measurements of the topside ionosphere require spacecraft sounders or density measuring devices.

# CHAPTER 13

# IONOSPHERIC RADIO WAVE PROPAGATION

Traveling at the speed of light, photons are nature's speediest beasties. They come in many colors, but all travel at the same speed in a vacuum. The situation changes when photons leave a vacuum. Light speed changes and may even become a function of wavelength (or frequency) in a dispersive medium. Additional changes are possible if the medium is a plasma such as the ionosphere. For frequencies near the plasma frequency, photon beams (also known as electromagnetic-EM-waves) may be reflected or refracted by the plasma. The EM waves known as radio waves are affected by the ionosphere.

## 13.1 Bands and Modes

Radio waves are subdivided into bands according to their frequency or wavelength. Commonly used terminology is summarized in Table 13.1.

Table 13.1 Radio Wave Bands

| FREQUENCY RANGE | WAVELENGTH RANGE | COMMON NAME |
|---|---|---|
| 20Hz-3KHz | Greater then 100 | (ELF) Extremely Low Frequency |
| 3KHz-30KHz | 100km-10km | (VLF) Very Low Frequency |
| 30KHz-300KHz | 10km-1km | (LF) Low Frequency/Long wave |
| 300KHz-3MHz | 1km-100m | (MF) Medium Frequency/Medium Wave |
| 3MHz-30MHz | 100m-10m | (HF) High Frequency/Short Wave |
| 30MHz-300MHz | 10m-1m | (VHF) Very High Frequency |
| 300MHz-3GHz | 1m-10cm | (UHF) Ultra High Frequency |
| 3GHz-30GHz | 10cm-1cm | (SHF) Super High Frequency |
| 30GHz-300GHz | 1cm-1mm | (EHF) Extremely High Frequency |

Microwave is a general term applying all frequences above about 200 MHz.

### 13.1.1 Propagation Modes

Because of their varying interactions with the ionosphere, radio waves may propagate in one or more different manners, or modes. Moreover, a given frequency may propagate by several modes simultaneously. Which modes predominate are determined by the ionosphere, the transmitter, and receiver. Figure 13.1 provides examples of the primary propagation modes for radio waves. These include the surface wave, ground reflected wave, direct wave, tropospheric wave, and sky wave.

The surface wave is the component which propagates along the atmosphere-earth junction. The efficiency of this mode of propagation depends on the conductivity of the surface over which it is passing. Water is a relatively good conductor, and this mode is important for overwater propagation when both antennas are close to the surface. Some frequency bands will penetrate the surface (greatest penetration occurs for the lowest frequencies) and be rapidly attenuated. Other frequency ranges are reflected in varying degrees at this boundary. Surface wave range depends on both surface conductivity and radio wave frequency.

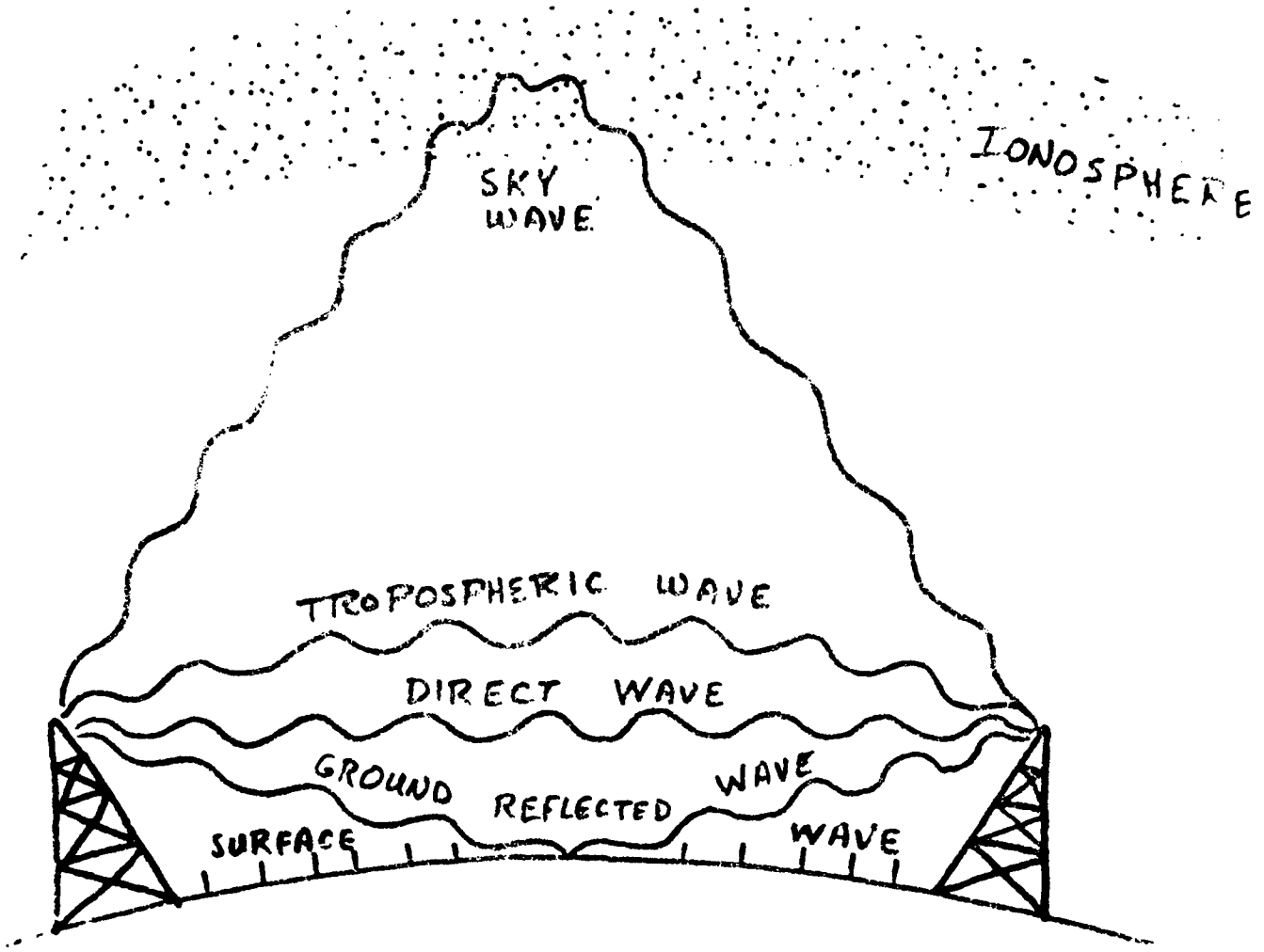



Figure 13.1 Radio Wave Propagation Modes.

Generally, neither the ground reflected wave nor the direct wave is effective over significant distances. The ground reflected wave has an effective range, depending on antenna height, of a few miles. For some bands, it may create a significant source of interference. By comparison, the direct (or line-of- sight) wave is probably the most reliable mode of radio wave propagation. Curvature of the earth effectively limits its range to tens of miles.

The tropospheric wave is, at first glance, apparently similar to the sky wave. Its refraction in the atmosphere usually occurs well below the ionosphere and is unrelated to electron densities. Steep vertical temperature or humidity gradients (often inversions) are the primary refracting medium. This propagation mode produces the so-called radar ducting. It is most effective for higher frequencies, but even here, range is limited. Tropospheric scattering can be a nuisance when path geometry is important (as in radar systems). In other situations, it is a primary mode of propagation. For certain frequency bands and conditions, the region between the ionosphere and the earth's surface may function as a sort of waveguide. Each is a

conductor and suffice to contain a radio wave, particularly in the LF, VLF, and ELF bands. These frequencies may propagate over considerable distances. Curvature of the earth (hence, the waveguide), changing D region height, and changing conductivity can significantly affect waveguide propagation over a short distance.

Meteors are a consequence of frictional heating of meteoroids (dust particles) which encounter the earth's atmosphere. A meteor is commonly defined by a trail of light, but an ionization trail is also produced. Although short-lived, these trails are typically regions of higher than normal electron density. As such, they are capable of reflecting radio waves incident on them. The large number of meteors occurring daily make meteor burst propagation a useful, albeit variable mode of propagation for higher frequencies. It may also significantly alter expected path geometry for normal sky wave operations.

### 13.1.2 Sky Wave Propagation

Long range propagation (in excess of about 1500 km) without intermediate repeater stations must use sky wave. Sky wave propagation relies on the refractive properties of the ionospheric plasma to bend the radio wave back to earth beyond the earth's curvature. The range and path geometry of sky wave propagation depend on transmitter takeoff angle (celestial altitude towards which the antenna is directed) and a number of ionospheric parameters. The maximum single hop range occurs for a zero takeoff angle (antenna aimed parallel to earth's surface) and is approximately 4000 km.

The sky wave is refracted by the ionosphere at the path control point. The signal will also be reflected by the earth's surface, with sea water providing a stronger reflection and forested areas tending to absorb more radio energy.

This "bounce" permits a range extension beyond 4000 km by means of a multi-hop path (see Figure 13.2). Radio energy returns to the ionosphere where it is again refracted back to the earth. Each bounce/refraction will scatter the incident radio energy. Some will be sent forward, in the direction of propagation, and some will be reflected backwards, towards the transmitter. A small, but non-zero fraction will be scattered to the left and right of the desired path. Finally, some of the energy will be absorbed in the surface and the ionosphere, and some will escape the earth's atmosphere. Most radio energy will propagate along a great-circle path from the transmitter in a direction dependent on the antenna azimuth. Ionospheric vagaries (tilts, or horizontal gradients in the ionospheric electron density) may produce non-great-circle propagation modes. The more hops a path contains, the greater the potential for path irregularities. (Any deviation from a great-circle route is considered an "irregularity" or non-great circle propagation.)

Path irregularities, often termed non-great circle propagation, usually occur in regions of strong horizontal (as opposed to the normal, vertical) electron density changes. Such variations are common near the auroral oval/subauroral trough and in the subequatorial regions. In these regions, strong horizontal gradients persist over long periods of time. Short-lived discontinuities are thought to appear and disappear throughout the ionosphere. Many

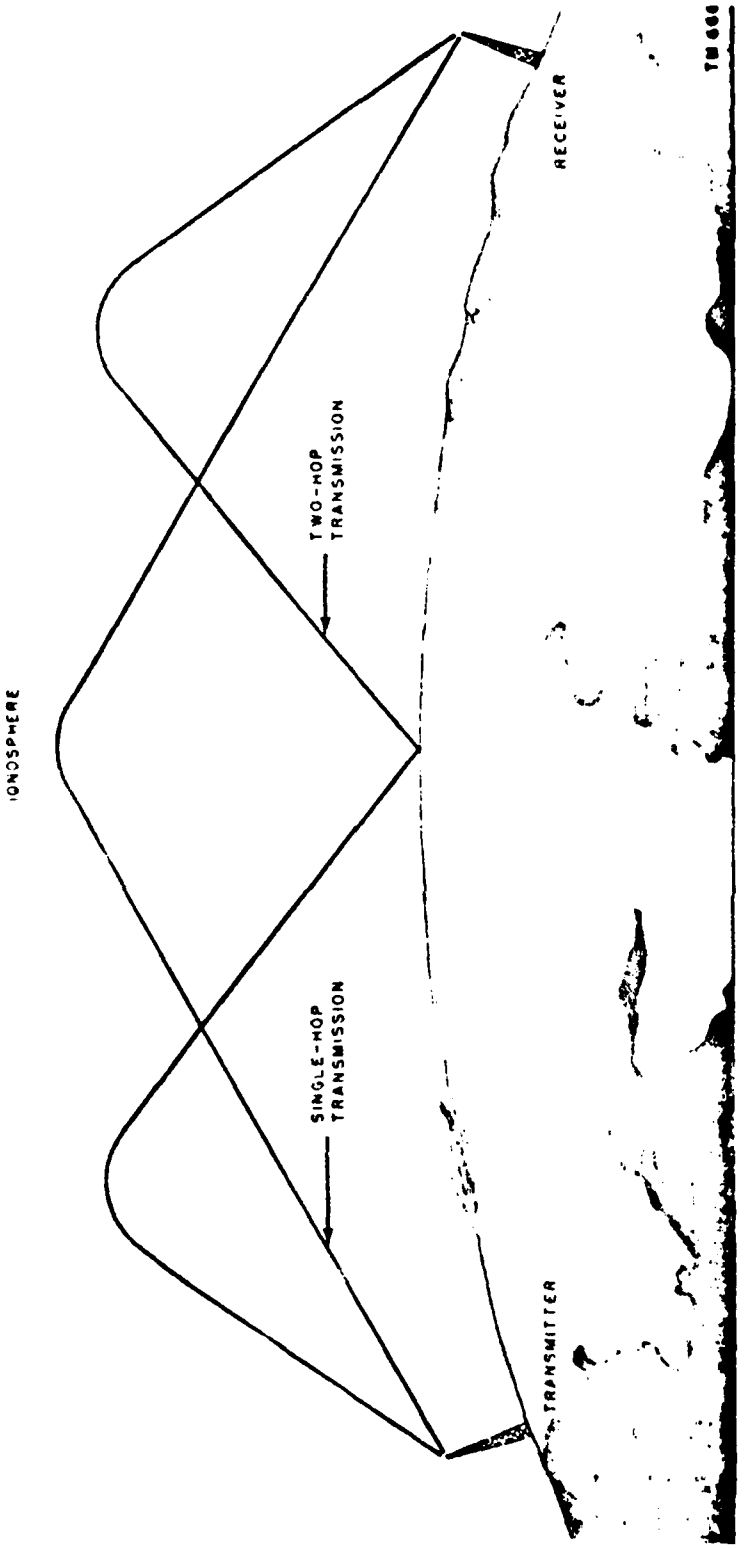



Figure 13.2 Single and Multi-Hop Paths.

probably last no more than 1-2 hours. They are often associated with the sunrise and sunset terminators and the South Atlantic and Southeast Asian anomalies. Solar flares, eclipses, auroral substorms, stratospheric warming, strong atmospheric storm systems, earthquakes (may result in a tsunami at sea), and volcanic eruptions are also thought to generate traveling waves of electron density. They are often called TIDs (Traveling Ionospheric Disturbances).

A radio wave doesn't differentiate in terms of the geometry of the electron density gradient it encounters. If the gradient is not perpendicular to the earth's surface (as it would be for stratification parallel to the surface) the radio wave will be deflected away from its original groundtrack. To confuse the results even further, a portion of the wave energy may maintain its original ground track, while the other portion is deflected in another direction. These unintentional path changes can be used to the operator's benefit if he is aware of their existence and structure.

Antenna pointing altitude and azimuth and frequency selection are the primary user inputs in determining sky wave geometry. In large measure, they do determine the path configuration, but selection accuracy is limited by our lack of knowledge concerning real-time ionospheric geometry. Altering the takeoff angle will vary the angle of incidence at the control point though not always by the expected amount. The minimum range results from vertical or normal incidence; the maximum range occurs for a minimum angle of incidence. Under certain conditions, grazing incidence can even be achieved.

Most frequencies have several possible sky waves. The extremes in angle of incidence for a given frequency define what are called the high and low mode waves. Figure 13.3 compares the various modes of sky wave propagation. The high mode (called the Pederson ray) may have equivalent range to the low mode wave due to ionospheric refraction at a level where the wave and plasma frequencies are nearly equal. Under these conditions, propagation is generally assumed to be by low mode wave, as the high mode wave experiences considerable attenuation and is not consistently dependable for the full distance range of the low mode.

### 13.1.3 Skip Zone

There is generally a minimum distance (frequency and takeoff angle dependent) for which sky wave will provide reliable communications. This minimum range will vary with time (due to ionospheric variability) and is defined as the skip distance. The takeoff angle which generates the skip distance wave is the skip angle. The highest angle at which a particular frequency will return to the earth is the critical angle for that frequency. Sky wave returns are generally strongest just beyond the skip distance. The region between the maximum effective range of the surface wave and the skip distance is the skip zone (Figure 13.4). Radio wave communications between the transmitter and any point within the skip zone are impossible.

## 13.2 Radio Wave Transmission

Radio waves are used for many different purposes, but the underlying purpose is to convey information from one point to another without establishing a

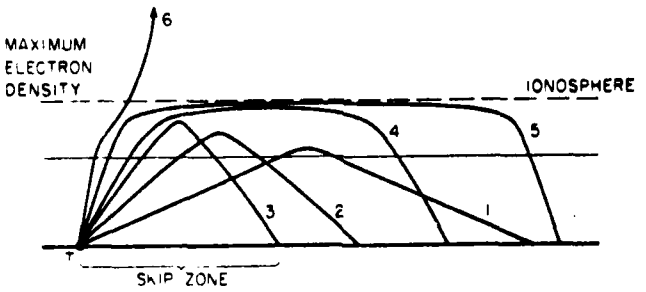



Figure 13.3 Effect of Varying Takeoff Angle on Sky Wave Propagation (from Davies, 1965).

solid link between the points. The information may be music, a television picture, an encrypted set of numbers, or just a series of dots and dashes. A transmitter normally operates at a given frequency. This is the carrier frequency and would be the nominal frequency at which you tune in the station. Much of what follows is generally applicable to radio waves, but some segments emphasize HF operations because of their wide use and considerable importance.

### 13.2.1 Bandwidth

Unfortunately, very little information could be transmitted at a single frequency. A simple audio signal may cover a range of 100 Hz-5 KHz. In order to transmit information, a channel is assigned centered on the carrier frequency and consisting of a small band of frequencies on either side of the carrier. These bands are termed the upper and lower side bands. The total bandwidth of the channel is the sum of the frequency range of the two side bands. Figure 13.5 graphically portrays bandwidth for an audio broadcast station operating at 15 MHz with a 10 KHz bandwidth. If this is an amplitude modulated (AM) station, its audio range would be 1 Hz to 5000 Hz. LF, MF, and HF stations are generally AM.

The upper and lower sidebands contain essentially the same information. A conversion to use of a single sideband (SSB) would provide the same amount of information while permitting twice as many channels. Unfortunately many radios are not designed for SSB. The other option is to reduce the bandwidth by reducing the stations audio range to perhaps 4000 Hz. Depending on the station's objective, this may not be feasible. It means a reduction in the amount of information which can be transmitted. Consider an audio signal consisting of music--a normal response range of a good stereo is 40-15,000 Hz. Obviously, this could not be fitted into a channel having a bandwidth of 8000 Hz (remember, that means a possible audio range of 0-4000 Hz--just half the bandwidth). SSB seems to be the direction of choice, but growth in this direction is slow.

### 13.2.2 Interference

The finite bandwidth of each radio wave channel automatically limits the number of users of a given frequency band. If more than one user is assigned the same channel, the potential for interference will exist. The degree of interference depends on the range of the frequency band in use and the proximity of the transmitters to the possible receivers.

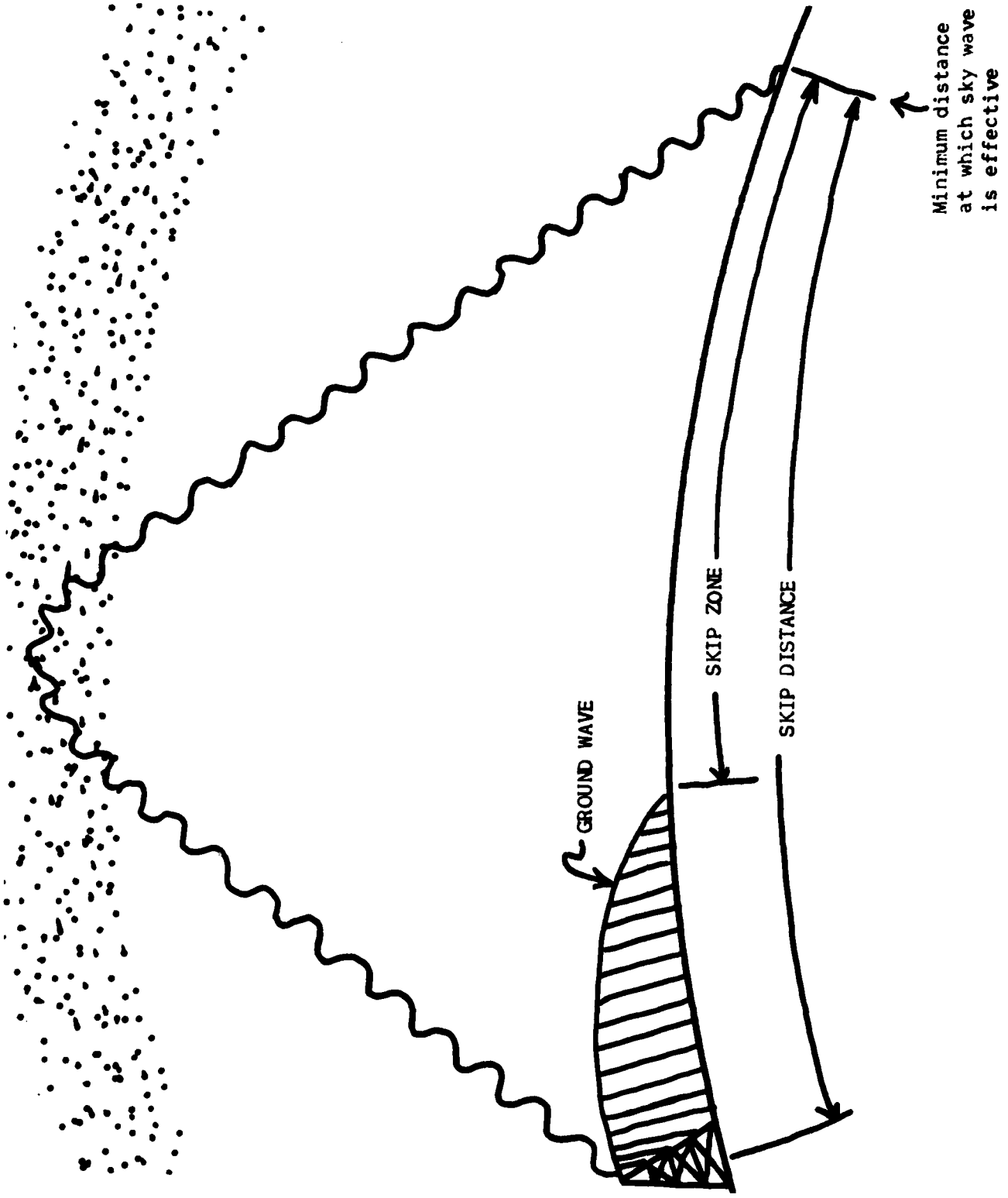



Figure 13.4 Skip Zone.

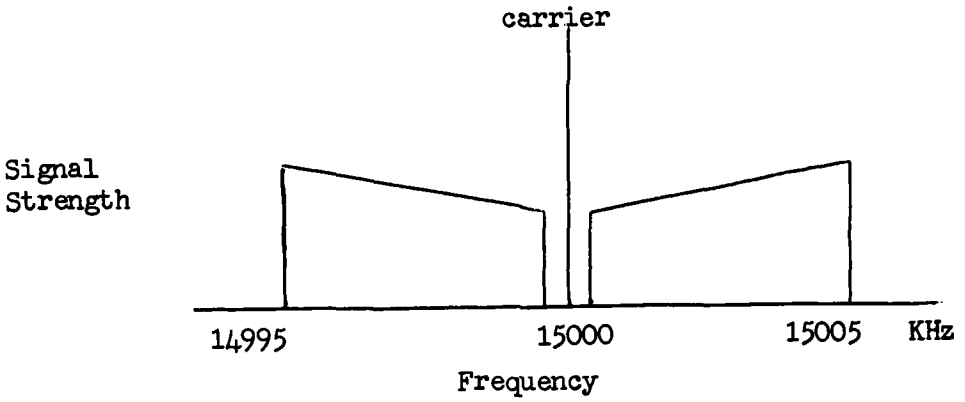



Figure 13.5 A Sample 10 KHz Channel in the HF Band (after Vastenhoud, 1980).

Interference also results from limited selectivity of the radio receiver. If the receiver passband exceeds the 10 KHz channel spacing (of this example) or is not precisely tuned on band center adjacent channels will also be sensed. Selectivity defines a receiver's ability to separate adjacent channels. Channel stability is also dependent on the transmitter's ability to control bandwidth and carrier wave frequency. Not all do, and interference will result from closely spaced adjacent channels.

13.2.3 Signal and Noise

While bandwidth limits the amount of information which can be transmitted by radio wave, the signal-to-noise ratio (S/N) determines what portion will be received. The larger this ratio, the easier it is to differentiate between signal and background noise. Radio wave systems require a minimum S/N in order to operate successfully. S/N can be increased by increasing signal, reducing noise, or both.

Signal strength is initially dependent on transmitter power and antenna efficiency (both receiver and transmitter). Antenna efficiency can be increased by using a directional antenna. Steerability and directionality are both important to concentrate available power into the desired direction. Selection of antenna size is also an important part of efficiency. At least one element of the antenna is, ideally, equal in length to a full or half wavelength of the radio wave to be used. Greatest power and efficiency usually combine in a fixed station. Portable stations usually have the least flexibility. Radio wave noise has considerable impact on the design and operation of radio systems. If it could be eliminated, infinite amplification (receiver) would be possible and infinitesimal transmitter power would suffice. Since amplification also increases noise, there are practical limits in its use. Selection of a receiving site and frequency band are of major importance in ensuring a desired S/N, since both man made and environmental radio noise are location dependent.

Man made noise sources can at least be identified. Most are associated with interference, electric motors, or electronic noise internal to the radio

itself. The latter type is often the dominant noise source on frequencies above about 300 MHz. Narrow bandwidth, directional antennas, and high power are the only real solutions to man made noise other than eliminating it.

Environmental noise cannot be eliminated. Noise power is generally proportional to bandwidth (due to increased exposure). Noise from thunderstorms, rain, and dust storms is concentrated in the LF and MF bands, although some overlap into neighboring bands does occur. Radio noise is also produced by the sun, other stars, and the fabric of space itself. Noise originating below the ionosphere will often be most intense here due to the trapping effect of the ionosphere. The ionosphere and earth's surface then provide a wave guide which ducts this noise worldwide within a short time. This would seem to imply a certain amount of homogeneity in the distribution of environmental noise. Figure 13.6 suggests that radio noise is not evenly distributed about the globe. Noise power depends on location, local time, season, frequency, and bandwidth. Ambient noise levels generally increase with frequency.

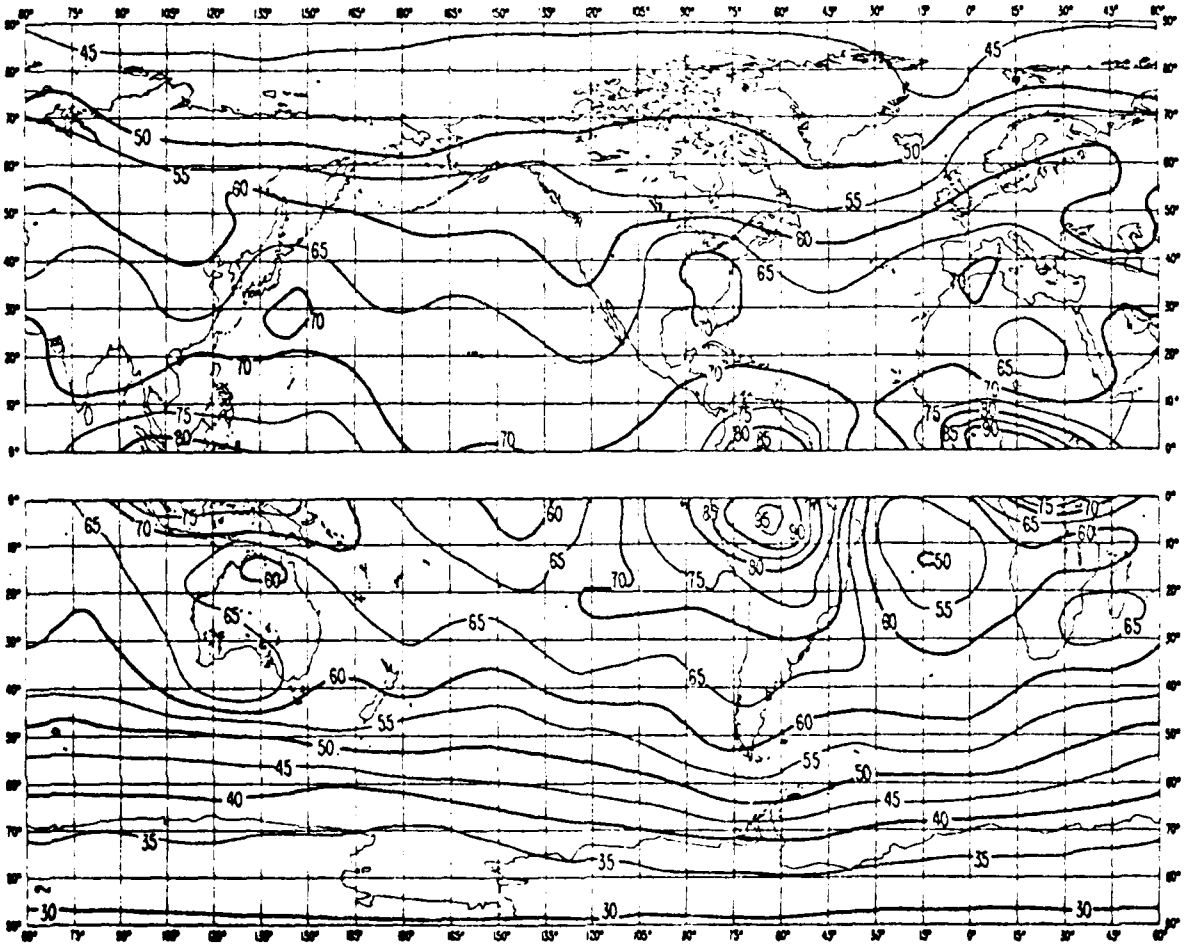


Figure 13.6 Radio Noise Distribution, Local Winter Morning (from Manley, 1981).

### 13.2.4 Fading and Absorption

The environment can alter the signal and noise levels in other ways as well, and sky waves are by far the most vulnerable. The conductivity of the earth's surface near the ground "bounce" point of a radio wave and the ionospheric conductivity near the control point both affect the amount of wave energy forwarded to the receiver. Likewise, absorption and fading can reduce the received signal and noise levels. These effects are generally frequency dependent. Fading usually results from path variations, but absorption is generally a consequence of conversion of radio energy into heat in the ionosphere. Absorption reduces both signal and environmental noise. Fading may affect both, but is more often a periodic change in signal strength alone. Fading is usually faster (shorter period, quicker drop-out) on higher frequencies. The duration of the fade depends on the causal mechanism and the frequency band involved.

Fading commonly results from changes in path geometry, but other origins are also possible. The nature of the fading can occasionally be used to infer its origin. Then, too, certain fading mechanisms apply only to a few frequency bands. Interference fading results when a signal from a given transmitter is made to follow two or more separate paths to the receiver. Horizontal electron density gradients (ionospheric tilts) may cause this by splitting the radio wave energy into several signals. Since the separate paths are often of different length, the signals will arrive slightly out of phase with one another. Unless electronic sophistication is employed, the receiver unknowingly accepts both signals. The resulting signal strength may vary due to destructive and constructive interference alternating as the paths slowly change. This type of fading often has a period of seconds and is known as multipath. Polarization fading is similar in period and origin. Differing polarization is imparted to the signals by the ionosphere because of the different manner in which they are reflected. When recombined, fading again results. Both polarization and interference fading are often lumped together as polar flutter when circuits involved pass near or through the high latitudes. Ionospheric tilts are common here, and the resulting fading is often identified with its region of occurrence.

Focus fading is usually associated with those frequency bands near HF and occurs for receivers located near the skip distance or at the extreme ground wave/line-of-sight range for a given transmitter. The skip zone for a given sky wave system is not fixed, but varies in response to changes in the overlying ionosphere. Focus fading usually has a period of 15-30 minutes.

Unlike fading, absorption typically reduces both signal and noise levels at the receiver. Absorption often results from an absolute change in the ionospheric electron density--not its configuration (tilt). In the D region, increased electron density means that incident radio wave energy is distributed among an increased number of tiny repeaters. Ideally this wouldn't matter, as each would rebroadcast the radio wave to another electron and so on until the radio wave departed the ionosphere (note that each electron has a fraction of the energy of the wave--like a number of network affiliates each broadcasting the same material, simultaneously). The D region is not, however, ideal. In addition to its tiny (about 1%) constituent of free electrons, it contains a huge population of neutral particles. Many of the electrons which receive the radio wave's energy collide with neutral particles before they can rebroadcast

the wave energy. These collisions dissipate much of the wave energy as heat, so the energy of the radio wave which departs the D region is significantly less than that of the wave incident on the D region. The closer the wave and plasma frequencies are to equality, the greater the absorption the wave will experience. Absorption is not as significant a problem in the higher levels of the ionosphere (E and F layers) because of decreased neutral particle density. Likewise, it is generally not significant for frequencies above about the middle of the VHF band.

Terminology often confuses the distinction between absorption and fading. Many texts allude to "absorption fading," and the term shortwave fadeout (SWF) is in common use. Stratospheric warming, PCAs, auroral zone absorption (due to increaed D region ionization from energetic particle bombardment), and SIDs (sudden ionospheric disturbance resulting from solar flare emissions) are all absorption events. SWF is the term applied to the absorption of an HF signal in the D region. The potential for confusion is apparent. When using such terminology, it is important to keep the underlying physics in mind. Only in this way is valid analysis possible.

The transmission of all radio waves is affected by the ionosphere. The effects vary somewhat by frequency band; so it is important to review the specific effects in each major band.

## 13.3 The Long Waves - ELF, VLF, LF

D layer plasma frequencies effectively prevent the entry of radio frequencies below about 300 KHz into the ionosphere. In short, these frequency bands are trapped between the earth's surface and the ionosphere. Skywave propagation at these frequencies is by a psuedo waveguide mode between ground and ionosphere. Considerable range and good signal strengths result. This waveguide also insulates these bands from some of the ionosphere's variability. Disturbance associated fading is slower than at higher frequences, and ELF through LF often remain usable throughout an ionospheric storm.

Unfortunately, these advantages are offset by several serious disadvantages. Probably the most serious of these is the limited bandwidth available. Noise levels are high in these bands, so high power levels are required. This is particularly critical for the longest waves. Equipment is not only expensive, but also very large. Some ELF systems require antennas hundreds of meters long, several megawatts of power, and employ vacuum tubes which are individually as large as a truck. The lower the frequency the higher the cost.

### 13.3.1 Employment

The high cost and sometimes ungainly size of these systems limits their use to certain specialized applications. Long range navigation and subsurface (ocean) communication are among their primary uses. These systems provide particularly good range over a good conductor, such as water. Moreover, signal penetration into a conductor is inversely proportional to wavelength. VLF will penetrate 10-15 m into sea water, and ELF can theoretically be received at depths in excess of 100 m. Whle a submarine might not have the space for an ELF transmitter, an ELF receiver and trailing wire antenna would permit

reception of commands while remaining safely submerged. Use of LF or higher frequencies would require at least the antenna mast be at or above the water's surface. VLF serves surface navigation through such systems as LORAN and OMEGA, where D layer stability is essential to positional accuracy.

The great overwater range permits a few VLF stations to provide nearly global navigation capability. LF experiences markedly higher absorption and is used for short range direction finding, particularly by aircraft near airports. This permits lower power LF systems to function with acceptable accuracy. Likewise, several widely separated LF beacons can use the same or similar frequencies without significant interference.

### 13.3.2 Propagation Variability

The stability and structure of the waveguide strongly influence the propagation of radio waves of the ELF, VLF, and LF bands. Changes in waveguide conductivity (surface or ionosphere) alter the phase speed of the signal. Wave speed is greatest during the day and over the ocean. Near midday (when the most D region ionization exists), phase speed sometimes exceeds the speed of light in a vacuum. These speed changes result in phase anomalies along the coasts and across the ionospheric terminators. Absorption of radio wave energy increases over land (surface a poorer conductor than water) and during the day (higher electron density in the D layer). Finally, diurnal changes include a change in the height of the D layer from about 70 km during the day to as much as 90 km at night. On occasion, even this vestigal D region will completely disappear at night. This change in the waveguide diameter typically results in a sunrise phase advance.

Other types of D layer variations *also impact* ELF-LF systems. Meteor showers deposit considerable heat and energy near 100 km (E layer). Since they are most intense between local midnight and noon, meteor showers tend to delay the normal sunrise phase advance. The primary seasonal change results from a lowering and intensification of the D layer in the winter day hemisphere. LF in particular experiences increased amplitude and a gradual phase advance under these conditions.

### 13.3.3 Ionospheric Disturbances

Three types of ionospheric disturbances have significant impact on long waves. Solar flare effects on frequencies below 300 KHz are generally shorter and less drastic than are the effects of a polar cap absorption (PCA) event. Flare effects are limited to the sunlit hemisphere, and PCAs are confined to the polar caps. Geomagnetic storms also affect the long wave bands on high latitude circuits. Effects are similar to those of the PCA.

Three flare effects are routinely identified with long waves. The SPA (Sudden Phase Anomaly/Advance), SES (Sudden Enhancement of Signal), and SEA (Sudden Enhancement of Atmospherics) all result from a lowering and intensification of the sunlit D region. "Intensification" refers to an increase in electron density due to solar x-ray and EUV flux released during the flare. Lowering the D layer reduces the effective path length by constricting the waveguide cavity. Since phase at the receiver is related to path length divided by signal wavelength, a phase advance usually results (some retardations have been observed). Reduction of the effective path length between transmitter and receiver also reduces the normal signal strength loss

($1/r^2$ drop off). A similar enhancement of low frequency atmospheric noise results. The SPA, SES, and SEA are all forms of the SID (Sudden Ionospheric Disturbance). The D layer height may fall to 60 km with a large flare (or about a 7-10 km change) and produce a 30°-60° phase advance. The SPA will onset in 1-5 minutes and decay in 30 minutes-3 hours. The exact impact on a given circuit depends on solar zenith and sunlit path length. Resulting LORAN and OMEGA positional errors may range from 1-12 km.

Polar cap absorption events may alter D layer heights by as much as 20 km and produce a 240° SPA. VLF signal amplitude may decline, while LF amplitude increases slightly. Since particle precipitation varies (in particle density and energy) during the course of the PCA, the resulting SPA will also vary. In fact, VLF systems are probably the most sensitive PCA monitor available. They sense much lower intensity precipitation than do riometers.

Geomagnetic storm effects on ELF-LF are very similar in both location and impact to PCA features. This results in considerable confusion, and high latitude ELF-LF operators often refer to all large SPA-like symptoms as a PCA. This can be frustrating to the analyst lacking other, corroborative data. Effects vary somewhat within these three frequency bands, but rapid, deep fading occasionally accompanies the main phase of the storm at high latitudes. Effects are usually delayed until after sunset and may occasionally be delayed for more than 24 hours after storm onset.

The recovery phase of a geomagnetic storm is often marked by highly localized auroral oval disturbances. The relativistic electron precipitation (REP) is perhaps the most intense of these events. Energetic electrons (100 KeV-1 MeV) are dumped--probably from the tail or trapping regions--into the oval. Collisions between these electrons and other atmospheric constituents near 100 km produce x-rays (bremsstrahlung). These x-rays penetrate to 60-80 km before ionizing surrounding particles. This localized increase in electron density produces a brief (1-3 hours) "bump" in the bottom of the D region. Circuits operating in the vicinity of a REP can experience intense phase anomalies and path variations at a time when conditions elsewhere are showing a gradual return to normal. The effects are all the more significant because they are usually unexpected and transient. REPs can also occur in conjunction with quiet time (geomagnetically) auroral substorms. In either situation, they may produce strong wave-like disturbances in the D region which spread out from the REP site. This permits a REP to influence propagation conditions at locations far removed from the site of occurrence. These TIDs strongly influence the F region, and a more extensive discussion of them appears in the discussion of HF systems.

### 13.4 Medium Frequencies (MF)

The 300 KHz - 3 MHz frequency band is the lowest capable of at least partial penetration into the ionosphere. A comparison of plasma frequencies reveals that MF should pass through the normal D layer and be reflected by or refracted in the E layer. Yet, MF frequencies don't significantly exceed the critical frequency of the D layer. Consequently, MF experiences considerable

non-deviative (D layer) absorption during daylight hours. The D layer is sufficiently weak (low to non-existent electron density) at night as to result in negligible absorption.

### 13.4.1 MF Propagation Modes

These conditions produce considerable variability in MF propagation modes. While sky wave mode always exists, it is not a significant means of MF communication during daylight hours. Even wave-guide propagation is not particularly useful as a consequence of the high D layer absorption of MF frequencies. In fact, surface wave and line-of-sight are the primary means of daytime propagation for MF systems. Long distance (beyond a few hundred miles) communication is impossible for MF systems during the daylight unless transmitter power is very high. At night, a very different situation develops. The turn off of ionizing solar radiation permits recombination to eliminate much of the D region. With the resulting drop in absorption, even low power circuits can communicate successfully over great distances.

Long distance communication is not without its problems. The MF band contains non-directional airport radio beacons at its low end and the commercial (AM) broadcast band at its upper end. Considerable frequency congestion exists in the MF band, and a number of stations are assigned closely spaced or overlapping channels. These stations are widely separated in geographical location, so there is little interference during daylight hours. At night, a very different situation develops. AM broadcast stations can often be received over considerable distances. Distant stations can thus interfere with local stations. The result, at least in the United States, is a complicated allocation of (1) assigned frequency (channel), (2) authorized power, and (3) limited broadcast times.

Certain stations (AM) in the United States are assigned clear channels. These stations are authorized to operate at high power (perhaps 50 kw) 24 hours a day and are assured that no other stations are authorized to use their frequency within the U.S. Consequently, clear channel stations are unlikely to interfere with local stations (or be interfered with by them) and may be received over considerable distance--particularly at night.

Other U.S. stations are not similarly assured of interference-free operation. In fact, several U.S. stations (widely separated) are often assigned the same frequency. In order to minimize interference with stations operating on the same frequency (actually, channel, since a finite bandwidth is assigned), several restrictions or combinations of restrictions are employed. These stations are usually limited in output power. Moreover, they are usually required to operate at even lower power (perhaps only 500 watts) during nighttime hours or go off the air entirely between sunset and sunrise. The latter restriction is often employed with new stations or stations located very near previously existing stations. Even these restrictions don't always ensure interference-free reception due to lack of receiver selectivity or transmitter frequency drift. The Federal Communications Commission (FCC) closely monitors the carrier frequency and effective bandwidth of commercial broadcast stations. Only the consumer can monitor (by his selection of radio) the selectivity of his receiver. The move towards stereo broadcasts on AM stations will increase the already acute frequency congestion in the MF band unless it is accompanied by increased technology (Remember: More information must be transmitted to generate stereo than for monophonic sound; so greater bandwidth is required).

### 13.4.2 Environmental Impact on MF

Disturbances of the ionosphere also affect MF operations. Yet, these effects are certainly limited by comparison with the impact of solar-geophysical activity on the neighboring LF and HF bands. The structure of the ionosphere itself explains this "insulation" of MF frequencies. Since MF penetrates the D region, it is not particularly sensitive to D layer height variations. The increase in D region electron density associated with a solar flare does increase absorption of MF signals, but this absorption is really noticeable only on frequencies near the top end of the MF band operating with considerable power. Lower frequency or power circuits are normally completely absorbed in the D region even in the absence of a solar flare, so increased flare emission is of little concern to most MF systems. Occasionally, the D region enhancement due to a flare will be sufficient to permit the layer to reflect MF signals. For a brief period of time, the affected MF circuits will experience a significant range and signal strength increase due to their exclusion from the D region. The result will often be increased interference by distant systems also using the affected frequencies.

Interference may also result from the radio bursts associated with solar flares. Many solar radio noise storms are concentrated in the upper LF and lower MF bands, and sunlit systems will be seriously degraded by the increase in wideband (large bandwidth) radio noise and consequent decrease in S/N. (Solar LF and MF bursts are usually more powerful than terrestrial transmissions and so penetrate the D layer even when transmitted signals cannot.)

Atmospheric and ionospheric storms can also affect MF operators. The radio noise of thunderstorms, dust storms, etc. so prevalent in the LF band does not ignore the MF spectrum. The effect is somewhat diminished for more distant storms and for point sources of manmade noise by the omnidirectional nature of many MF antennae. Such antennae average signals at a particular frequency from all directions, so the noise contribution from one sector may not be overly significant--depending on its distance and strength. Ionospheric storm effects are somewhat more mixed in their impact. Blanketing sporadic E during the nighttime hours may significantly extend the range and increase the received signal strength of an MF signal. This is a consequence of the much stronger reflecting surface provided by the sporadic E in comparison to the normally anemic nighttime E layer. Conversely, sporadic E may alter MF path geometry and result in intermittent signal fading at the receiver site due to E layer variability. Increased D layer absorption due to PCA or auroral particle precipitation will usually have effects similar to those of solar flare. Some differences will be apparent near the auroral oval. Here, strong horizontal gradients of D layer electron density (from non-existent in the nighttime sector equatorward of the oval to very high in the oval) may produce sideways reflections of MF transmissions and result in long distance, non-great-circle propagation. The ambiguity of storm effects is not, however, limited to the MF band.

## 13.5 Shortwave Radio - the HF Band

The 3-30 MHz frequency range is probably the most extensively used of all radio wave bands. If anything, frequency congestion is even worse on HF than for MF. Markedly greater bandwidth is available to HF operators, and HF

easily provides the means of long distance communication. The basic technology for the employment of this band has long been available, and extensive research continues. Consequently, costs are low. The desirability of HF is tied up in its low cost and low power requirements combined with its truly worldwide range. The same physics which provide great range and good signal strength for low power also yield the major disadvantages of HF: highly variable propagation conditions (resulting in decreased reliability), high interference, and significant path variability.

### 13.5.1 HF Circuitry

A comparison of plasma frequencies and HF frequencies suggests that long range sky wave propagation of HF will be by means of the F region. Ionospheric variability means path variability, however, and other ionospheric layers may briefly alter HF propagation modes. Whenever they are present, the D, E, and F1 layers will refract (and occasionally reflect) the HF radio wave prior to its arrival in the F2 region. This means that more purely geometrical considerations are necessary to determine the exact portion of the ionosphere illuminated by an HF signal. Further confusion is a consequence of the signal "spread" as it moves outward from the antenna. At the F2 layer, a typical HF system will illuminate an area 10-20 km on a side. Ionospheric irregularities within the beam pattern will affect signal continuity by dispersing the signal. This reduces effective signal strength independent of other factors such as D region absorption.

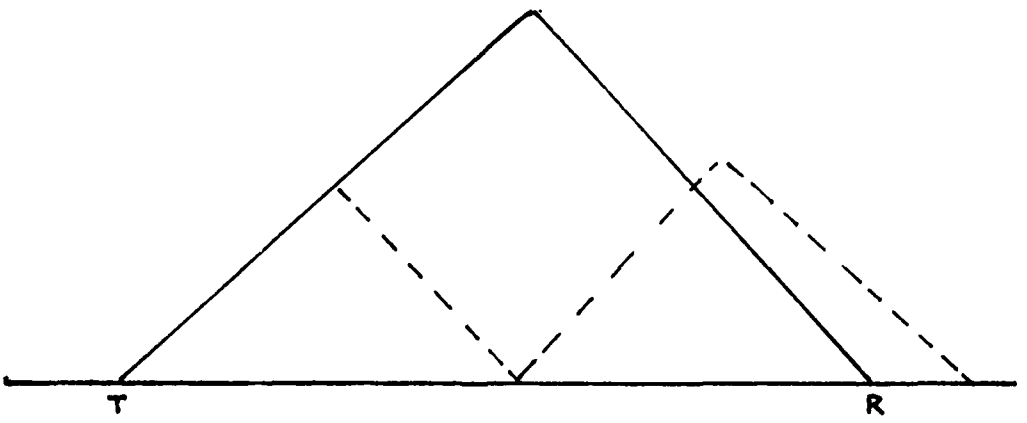
Systems equipped with directional antennas generally assume F2 layer reflection and great-circle propagation. So, too, do many AFGWC models. The ionosphere is not always in agreement. During daylight at solar maximum and occasionally during the summer daytime, F1 electron densities may exceed those of the F2 layer for a few hours. This is also observed during intense ionospheric storms. If elevation angle and azimuth are predicated on F2 geometry, communications may be lost during these periods. Figure 13.7 compares the paths resulting from F2 layer and low layer reflection at a fixed elevation angle. The effect is very similar to skip fading. Proper analysis may be initially difficult, since the control point (or points) location depends on the height of reflection. Equally significant, lowering the control point may convert a single-hop path into a two hop path. If a multi-hop path includes control points at significantly different altitudes the likelihood of encountering ionospheric tilts is increased. This condition is termed "mixed mode" propagation, and requires a dense sensing network to adequately define (due to extreme positional variability of ionospheric conditions). Temporal variability also makes definition difficult, since significant changes may occur in a manner of minutes.

An extreme example of layer-induced variability is known as "skip". A term common to CBers (Citizen Band Radio; CB operates near 27 MHz) and ham operators, skip describes propagation at higher than normal frequencies over greater than normal distances. This is often a consequence of E layer propagation. In particular, blanketing sporadic E may briefly double the range and maximum usable frequency (MUF) on a given path. For other users of the same control point but with slightly different geometry, the occurrence of blanketing sporadic E may eliminate propagation for a time.

Skip has other connotations as well. It may, for example, permit unusually good transequatorial propagation by avoiding the D region for much of what would normally be a multi-hop path. The subequatorial ridges provide strong horizontal density gradients. A judicious choice of path geometry (see Figure 13.8) places the radio wave nearly tangent to the F2 layer over much of its journey. This eliminates all but two passes through the D region. Subsequent reduction in D region absorption may make communications possible where they were not previously. Of course, skip can coexist with "normal" or other mixed propagation modes. The result will usually be multipath fading (if the signal reaches the receiver by more than one path) or a reduction in signal strength (if only one does) since incident energy must be divided among the various modes.

### 13.5.2 Ionospheric Scattering

Ionospheric scattering may be intentional or accidental, but it is almost always present. Likewise, ground scatter occurs on most circuits. In fact, this situation makes both forward and backscatter HF systems possible. Until now, we have tacitly assumed a forward scatter mode was optimum. This is not always the case. A backscatter system permits collocating the transmitter and receiver. Such a system is called monostatic (single-station) as compared to more common bistatic (two-station) systems. While a monostatic system may not be very useful for communications, it is appropriate for a detection system such as a radar. Unlike normal search radars (which operate in the VHF and UHF bands) HF radars are <u>not</u> constrained to line-of-sight detection. Since HF waves are reflected by <u>the</u> ionosphere, over-the-horizon (OTH) detection is possible. HF radars of this sort are often known as OTHB systems since they use backscatter from the target along the transmission path to the transmitter.




13.7 Path Variation Resulting from Altering Control Point Altitude.

Simply collocating the transmitter and receiver and defining the system as a backscatter operation does not alter the ionosphere's effect on it. Generally, most wave energy will propagate in a forward scatter mode. Only a small portion (typically less than 1%) of the transmitted energy will be backscattered. This means that a backscatter system requires considerable power to generate a detectable return. An OTHB system views not only its targets but also the earth's surface. In order to differentiate between the two, radar electronics look for target motion. This use of the Doppler effect does have limitations. The radar's signal will be backscattered from objects in motion other than ships, tanks, or aircraft. In particular, the precipitating particles of the aurora are capable of generating Doppler-shifted returns which may look very much like aircraft or missile launches.

In addition to identifying a target, the radar must identify the target location. If the ionosphere were a mirror, then pure geometry would permit us to locate the surface area being illuminated at the end of the first ionospheric hop. Pure geometry doesn't work very often, particularly near the auroral oval. The ionospheric tilts, refraction, and the finite area illuminated by the radar beam combine to create considerable confusion in interpreting an OTHB return. A wide beam may experience bending and flexing as well as pure displacement. The use of surface transponders (an electronic device which generates a radio signal when it receives a radio signal) or positive identification of fixed surface features can help, but only a multi-dimensional ionospheric model will be truly useful.

Backscatter can also be useful to a forward scatter system. Greatest utility will accrue to a frequency-limited system operating near strong ionospheric discontinuities (e.g. a terminator or the subauroral trough). Figure 13.9 demonstrates a configuration where backscatter can make communication possible by using a control point in a more reliable or stronger portion of the ionosphere. Again, making use of non-standard path geometries requires considerable knowledge of the ionosphere and system flexibility.

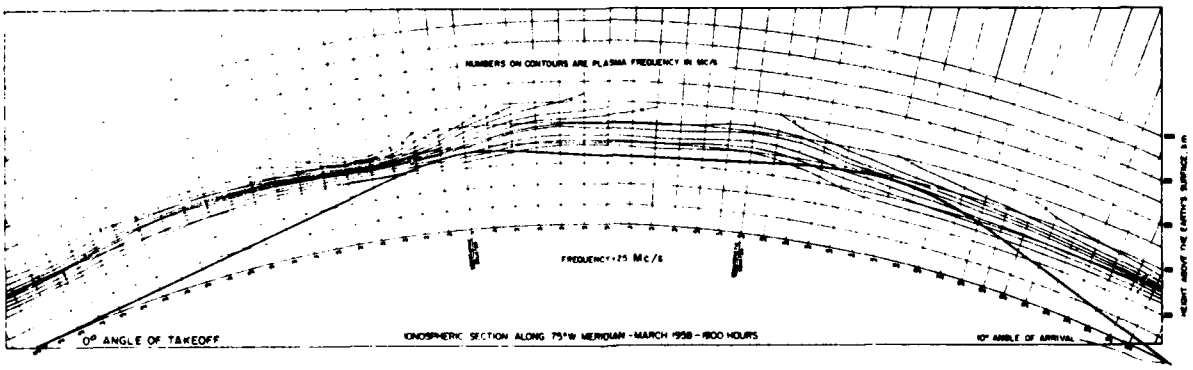



Figure 13.8 Grazing Incidence on Transequatorial Path (after Davies, 1965).

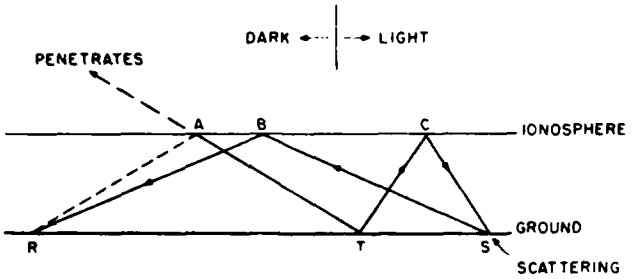



Figure 13.9 Forward Scatter System Using Backscatter Propagation to Avoid a Weaker Control Point (after Davies, 1965).

Scattering may also occur in planes perpendicular to the forward scatter, great circle plane. Such scattering is commonly called non-great-circle propagation and may occasionally be very useful. There are similar left-right variations which are not true scattering but result from the variation in the index of refraction with altitude. The geomagnetic field splits the radio wave into ordinary and extraordinary components. For fairly short paths, they recombine on emergence in a plane parallel to their entry plane. (The incident and emergent planes coincide for exactly east-west paths at all latitudes and for north-south paths lying along magnetic meridians.) For longer paths (in excess of about 400 km), the ordinary and extraordinary waves are deflected south and north of their great-circle centerline by several hundred meters. A horizontal projection of the resulting ray paths is depicted in Figure 13.10 for a generally east-west path.

Since higher frequencies penetrate to higher altitudes in the F2 layer, they experience larger mid-path deflections. North-south propagation does not experience a similar left-right displacement from centerline. Rather, the ordinary and extraordinary wave control points are displaced north and south (respectively, in the northern hemisphere) from the geometrical path control point. Were the ionosphere homogeneous, these displacements would have no impact. It is not, so the ordinary and extraordinary waves will generally experience slightly different propagation conditions enroute to the receiver. This may produce constructive or destructive interference when the two are recombined at the receiver. Problems of this sort are more likely for paths operating close to various ionospheric irregularities. At the least, such effects may cause conditions to <u>appear</u> much better or worse then is truly the case. VI sounder measurements made near the theoretical path control point will often have little relevance to conditions actually experienced by the radio wave. This emphasizes the importance of station density in identifying actual ionospheric conditions.

Another consequence of these apparently small scale effects is that two HF systems which are identical in all ways except transmitter or receiver location will probably encounter different ionospheric conditions even if they use the same control point. Just the different orientation of the paths is sufficient to induce differences. Each path will experience different lower ionosphere conditions and impact the control point from a different direction. Changing the angle of incidence alone can significantly alter ionospheric response to a radio wave.

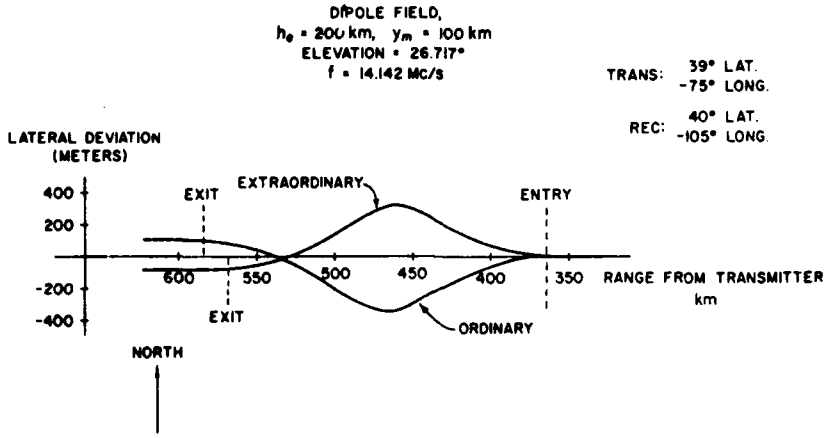



Figure 13.10 Horizontal Projection of East-West Ray Paths Showing Expected Deflections (after Davies, 1965).

### 13.5.3 HF Climatology

Just as certain parameters (e.g. $f_oF2$, MFAC, etc.) are used in vertical ionospheric sounding, so too, do we use a selection of terms to specify HF radio wave propagation conditions. These terms, while related to theoretical parameters, are generally operationally derived. Several can be related to vertical sounding parameters and are often highly path and equipment dependent.

#### 13.5.3.1 HF Parameter Definitions

The maximum usable frequency (MUF) is usually defined as the highest frequency for a given length path using only ionospheric refraction on a great circle route. Theoretically, it is related to the $f_oF2$ by the M factor. It is generally calculated for 3000 km or 4000 km single hop paths. A 4000 km path is the longest single hop path possible under normal conditions. At this frequency, only one ray path is possible. The high and low rays merge at this frequency, and it could be taken as equivalent to the F2 layer junction frequency. Maximum frequency communications will be possible at the MUF 50% of the time. Since there is only one ray path for the MUF, the path length for this frequency is its skip distance. The MUF4000, for example, will not propagate over any distance shorter than 4000 km.

Communicators generally prefer to employ the highest frequency possible for operations involving the F2 layer. This minimizes ionospheric absorption, which decreases as the square of the operating frequency. The MUF, as implied above, does not account for ionospheric and path variability. The result is a 50% success rate using the MUF. A statistical analysis has revealed that a frequency of about .85MUF provides reliable communications on a given path about 90% of the time. This frequency, known as the frequency of optimum transmission (FOT), is taken as a much more usable frequency for routine

operations. It is statistically derived and may not always be usable for a given path because of geometry considerations. For a given path, the FOT might be too high for a two-hop connection and too low for a single-hop, high mode (Pederson) ray.

The gradual replacement of voice communications with high speed data links has introduced yet a third factor into consideration--the multipath reduction factor (MRF). Multipath interference can be particularly severe in the HF band with correspondingly serious impact on data link systems. Multipathing increases in severity as the signal time dispersion increases (i.e. the difference in path lengths increases). Since the higher frequencies have fewer possible propagation modes for a given path (the MUF has only one), increasing frequency decreases time dispersion (compare transit time for the high and low rays as the junction frequency is approached). The MRF is defined as the lowest percentage of the MUF for which the range of multipath propagation times is less than some specified value. It provides the lowest frequency at which multipathing is reduced i.e. below a certain, predetermined level). By determining the permissable time dispersion for a given data link, the operator can determine the MRF. Applying MRF to the MUF will yield the optimum combination of low multipath and high ionospheric reliability.

Ideally, multipathing would be minimized simply by using the highest frequency which is less than or equal to the MUF. Realistically, frequency allocations usually don't permit this. A slightly less desirable alternative can be employed in such a situation. Statistics suggest that multipath interference can be reduced on lower frequencies by selecting a propagation mode such that the frequency used is near the MUF for that mode (e.g. a multihop E mode path instead of a long single-hop, or use of high angle rays where possible). Such adjustments will usually require steerable antennas and operator flexibility. Research indicates that multipathing (signal time dispersion) reaches a maximum for values near .56 MUF on a 1000 km path (Davies, 1965). Similar estimates might be made for longer paths by extrapolation.

While the MUF and its associated factors are strongly dependent on the ionosphere, the lowest usable frequency (LUF) is almost equally dependent on the ionosphere and the equipment employed. D layer absorption is the primary ionospheric input to the LUF. System S/N requirements may also impact the LUF depending on frequency congestion and environmental interference and noise. If the circuit uses a multi-hop mode, the conductivity of the surface at the ground bounce point will also influence the LUF by determining ground energy loss. A good conductor (e.g. sea water) will work to lower the LUF, while other types of terrain (e.g. forests) will elevate the LUF. System or engineering factors involved in LUF establishment include transmitter power, transmitter and receiver efficiency, antenna and feedline efficiency, and polarization matchup between transmitter and receiver.

The MUF, FOT, MRF, and LUF are the primary operational parameters of all HF systems. Radio operators occasionally interchange this terminology, yet the ideas remain unchanged. In reviewing observations or requests, it is critical to properly identify the parameter in question. An example of the potential confusion is the MUF/FOT forecast provided for Presidential Airways HF (Mystic Star). MUF/FOT data is tabulated as an URF/MRF (Ultimate and Maximum Range Frequencies) listing. The potential for confusion here is obvious; it is often considerably more insidious in an operational environment.

13.5.3.2 Diurnal HF Variations

Sunrise. Its effect on the ionosphere and HF systems accounts for one of the major ionospheric gradients and heralds significant diurnal variability. Ionospheric sunrise occurs as much as 3 hours before surface sunrise. The actual difference is a function of location, season, and ionospheric layer. F layer sunrise preceeds D layer sunrise, etc. These discontinuities cause confusion at best; at the worst, they may account for a complete shutdown of certain HF systems for several hours.

Under quiet ionospheric conditions, the sunrise transition at a given point in the ionosphere may require 1-2 hours. During this period, the D, E, and F1 layers are nearly reborn. The F2 layer becomes a distinct individual and rises slowly to maximum daytime electron densities. Except for paths oriented generally north-south, one-hop or longer paths will usually experience LUF increases before the MUF begins to rise. This is a consequence of the most eastern D region transit point of the path experiencing sunrise prior to control point sunrise. While the MUF is controlled by the F region at the control point (essentially a single point), the LUF is a result of the combined effects of two D region transits (for each control point). The more easterly D region transit, after sunrise, will control the path LUF. Depending on the path length and orientation, LUF increases may precede MUF rises by as much as an hour or more. The sunrise transition produces a very limited propagation "window" (range of usable frequencies) for the communicator to contend with. This propagation window is portrayed in a typical MUF/LUF plot shown in Figure 13.11.

The sunrise transition causes an even more extended and difficult transition for east-west oriented multi-hop paths. Such a system is shown undergoing sunrise in Figure 13.12. The westward sweeping terminator first sets off a LUF increase at the eastern most D region transit point. As a result, the circuit LUF will rise. Next, following sunrise on the eastern control point, the eastern MUF rises. The circuit MUF will not begin a simultaneous rise, because the western control point MUF will remain low. In fact, the western MUF will reach a relative minimum just before ionospheric sunrise. Despite the eastern rise, it is the western control point which, in this case, controls the circuit MUF. For a two-hop path, the additive effects of three sunlit D region transits will be available to elevate the LUF before sunrise occurs at the western control point and the circuit MUF begins to rise slowly. The longer the east-west component of a circuit, or the more hops it makes, the more severe and drawn-out will be its sunrise transition. If one of the path control points is located near the nighttime auroral oval, communications could even be lost for several hours during the transition due to the extreme MUF depressions present in subauroral trough. The trough will move slowly northward and dissipate as night becomes day in its vicinity. Nonetheless, MUFs in its vicinity will be slow to rise after sunrise.

Equatorial sunrise and sunset will be much sharper transitions than is the case in the middle latitudes. Since the sun consistently rises more nearly vertically in the equatorial regions (a consequence of the earth's axial tilt) the transition period is brief. This accounts for rapid LUF increases at sunrise and is probably responsible for the post-sunset instabilities which appear in the form of spread F.

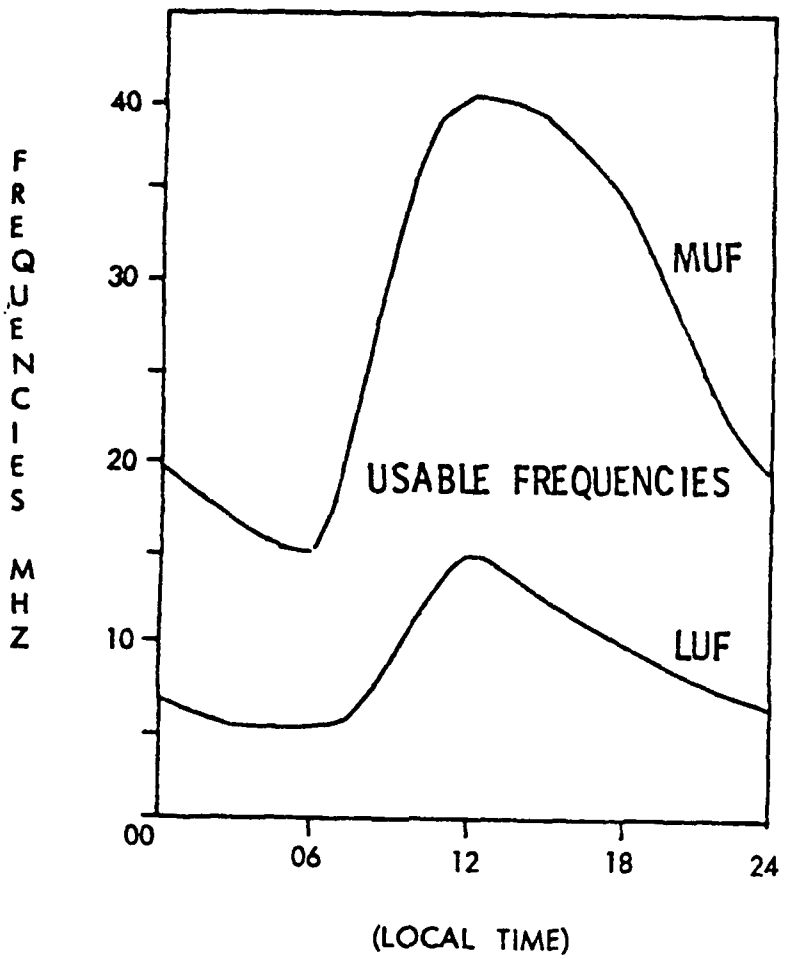



Figure 13.11 HF Propagation Windows Showing Typical Diurnal Variation in Winter Mid-Latitudes.

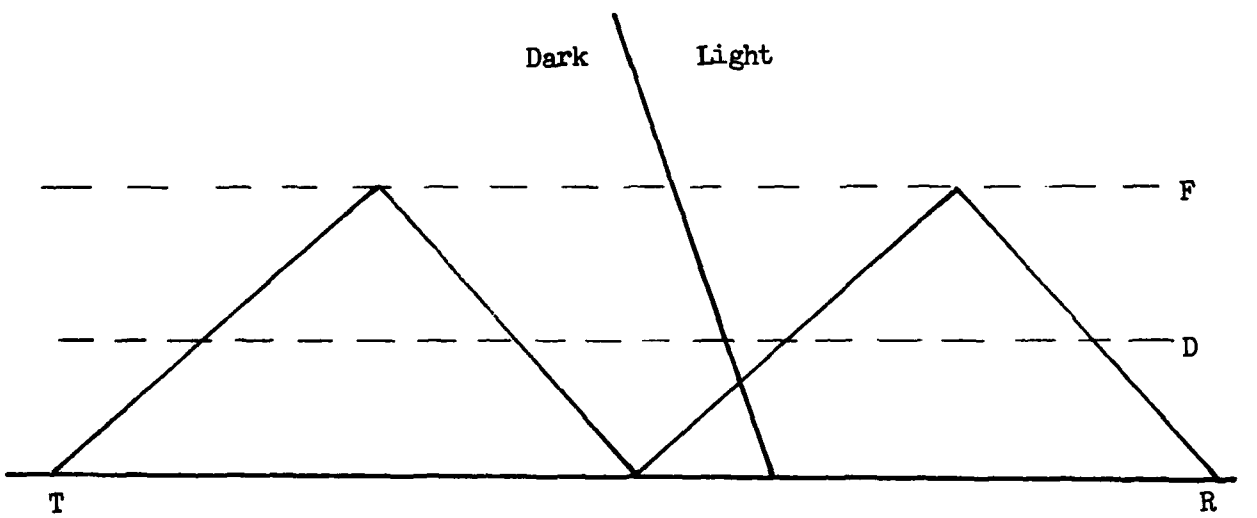



Figure 13.12 Multi-hop East-West Path Undergoing Sunrise from Right to Left.

A very different situation exists at the high latitudes. Above about 66 1/2° geographic latitude, the sun is above the horizon continuously for a portion of each year. Although slight diurnal "ripples" are observed, the diurnal curve shows very little transition at a given station. The problem occurs in communicating with lower latitude stations. North-south paths out of the polar regions must inevitably cross a terminator at some time. In fact, this terminator may remain generally fixed for long periods of time while the other end of the path is in night.

Depending on geometry, the control point (or one of the control points for multi-hop path) for a north-south path from the polar regions may fall in the nighttime oval or in the subauroral trough. The result can be extreme path and frequency variability. Since one D region transit (the high latitude one) will probably be continuously sunlit, the path LUF will remain elevated. As the control point slips into the trough, the MUF may fall precipitously. Moreover, path geometry may be so altered as to preclude normal (great circle) propagation. Alternately, blanketing sporadic E in the auroral oval may yield superior communications with an excellent S/N ratio. During the winter months, the reverse situation will occur. The polar terminus of circuit will be in nearly continuous darkness, and the terminator will lie nearly east-west across the path. When propagating into the nighttime hemisphere from a darkened polar cap, terminator problems are replaced by the subauroral trough to hold down the MUF while auroral zone activity may elevate the LUF (if one of the D region transits occurs near the auroral oval). Regardless of the season, the sunrise and sunset transitions will be sharper at the equator and more gradual at high latitudes. The auroral oval and subauroral trough complicate high latitude communications regardless of season.

### 13.5.3.3 Seasonal Effects

Just as seasonal change is apparent in the weather of the middlle and high latitudes, so too is it apparent in the ionosphere. The F layer is the most variable portion of the ionosphere, but it is not the only level to undergo seasonal change. Both the electron densities and the heights of all the ionospheric components show seasonal change. Moreover, the diurnal cycle of each layer changes with season. The results are readily apparent on HF systems.

LUFs are generally highest during the summer, but winter days occasionally yield abnormally high LUFs over certain regions. These regions may be as much as 100° in longitudinal extent and are typically confined to the middle latitude daytime sector. Typically, the highest LUFs are found within 20° of the dayside geomagnetic equator. These LUF enhancements generally persist for several days before gradually returning to more "normal" values. The origin of this sudden LUF enhancement is uncertain. It may result from solar proximity during northern hemispheric winter and the paucity of southern hemisphere observations, in which case it would more correctly be termed a December anomaly. Some researchers have found these LUF enhancements to follow closely (few hours) the onset of geomagnetic storms in the winter, day hemisphere. In this interpretation, the effect might be more correctly considered a storm-related phenomenon. Particle precipitation resulting from a storm would occur at both conjugate points simultaneously. Resulting LUF enhancements would be more apparent in the winter hemisphere, because the cooler atmosphere allows deeper penetration (by the precipitating particles) and greater ionization.

A more localized LUF enhancement results from stratospheric warming. Normally observed in January-February-March in the northern hemisphere, stratospheric warming produces temperatures tens of degrees above normal at and above 10 millibars (about 30 km). These changes occur above 30° - 60° geographic latitude and often begin over Siberia or the North Atlantic. (Southern hemispheric warmings also occur but are not as well studied.) They persist for several days in bands of 10°-15° longitudinal extent. During this period, D layer heights fall, and LUFs rise.

Since the strength of the E layer is a function of solar zenith distance, we might (correctly, for once) anticipate that the summer E region is stronger (exclusive of ionospheric and geomagnetic storms). A strong E layer means more low-level refraction of HF signals and some increase in absorption. This often results in degraded propagation and path anomalies (non-great circle propagation may become common).

F1 layer variations have a similar impact on HF operations. This layer is most intense during summertime and near sunspot maximum. At these times (particularly in conjunction with an ionospheric storm), $f_oF1$ may exceed $f_oF2$. The result will be F1 mode propagation. The F1 layer heights are significantly lower than those of the F2 layer. The resulting change in path geometry means that normal antenna alignments may not be workable. Since a maximum range one-hop F1 path will be shorter than a maximum range one-hop F2 path, the path may require an increased number of D region transits. This means a higher path LUF when compared to F2 propagation over the same path. The worst case will result when the F1 layer either (1) only periodically exceeds the F2 layer or (2) exceeds the F2 layer over only a portion of the path. The resulting mixed mode propagation (mixed in either space or time, or even both!) may cause multipath interference or even periodic circuit outage.

Seasonal variability is probably greatest in the F2 layer. This means that MUFs (and FOTs and MRFs) should show the greatest change. Comparing "propagation window" charts for an upper middle latitude circuit in winter and summer (Figure 13.13) confirms this. The increased diurnal variability in winter exacerbates the sunrise transition problem on east-west HF paths.

The information of Figure 13.13 can be extended by reference to the Institute for Telecommunications Services (ITS) (Roberts and Rosich, 1971) climatology. The salient feature of all these sources is a very high, single daytime MUF peak for HF systems in the winter hemisphere. Corresponding summer data suggests little variation in MUFs during the day. For the HF operator, this usually means few frequency changes are required to maintain communications during the summertime (assuming ionospheric quiet). By comparison, winter operations will require considerable frequency variation to maintain optimum circuit performance throughout the day.

Even winter variability is damped for equatorial and polar communicators. Polar diurnal MUF variations are slight (except for subauroral trough variations) regardless of season and are more related to the low latitude path terminus than to variations in the polar ionosphere. Low latitude communicators typically see a single MUF peak in the late afternoon (control point time).

It is critical to remember that HF effects depend on the location of the D region transits and the F layer control points. Ionospheric conditions at the transmitter and receiver sites are irrelevant (except, perhaps, for establishing background noise levels). This is not to say that the locations are unimportant. They determine the angle at which the signal strikes the ionosphere. The importance of this is revealed in the lack of symmetry on an HF circuit. Occasionally, propagation will only be possible in one direction on an HF path. The MUF, LUF, and geometry may vary depending on the direction of propagation on the circuit. This is at least partially a consequence of ionospheric anomalies. Multi-hop east-west circuits may see considerable variability in the winter, and even low latitude north-south paths may experience increased variability because of sun angle and proximity. Only a complete analysis will reveal the numerous potential sources of variability for a given circuit.

### 13.5.4 HF Anomalies

Compounding the difficulty in analysis of HF operations are the numerous ionospheric abnormalities. The South Atlantic and Southeast Asian positional anomalies have been mentioned elsewhere. Individual layer anomalies may be independent or in association with the geographical anomalies and include deviative absorption, spread F, ionospheric tilts, TIDs, and sporadic E. The discussion which follows is limited to the HF impact of these features.

#### 13.5.4.1 Geographic Anomalies

The offset of the geomagnetic field skews the ionosphere in a similar manner. The resulting anomalies are both near the magnetic equator and are similar in their strong horizontal gradients.

The Southeast Asian anomaly provides an abnormally strong magnetic field for each altitude. As a consequence, trapped particles may have higher energies here, and trapped densities are higher. This results in higher topside electron densities and higher than expected MUFs for systems with control points in this vicinity. The high $f_oF2$s (and resulting MUFs) probably result from the topside support being added to that of the F1 layer production. The enhanced MUFs are sometimes disguised by the subequatorial ridge structure. The ridge structure further confuses the horizontal gradients in this area. The end result is an enhanced but highly variable F2 region. Lack of path symmetry should be expected in this area at all times.

The South Atlantic anomaly roughly overlies Brazil and is a region of unusually low magnetic field strength for a given altitude. As trapped particles move into this vicinity, many are lost. In order to attain trapping field strength, the particles must move to lower than normal altitudes. Here, they encounter higher neutral density and are lost by collision. Heaviest precipitation occurs near the western edge of the anomaly (the earth rotates west to east) and produces ionospheric effects similar to auroral oval precipitation. LUF enhancements distinguish HF system effects in this vicinity.

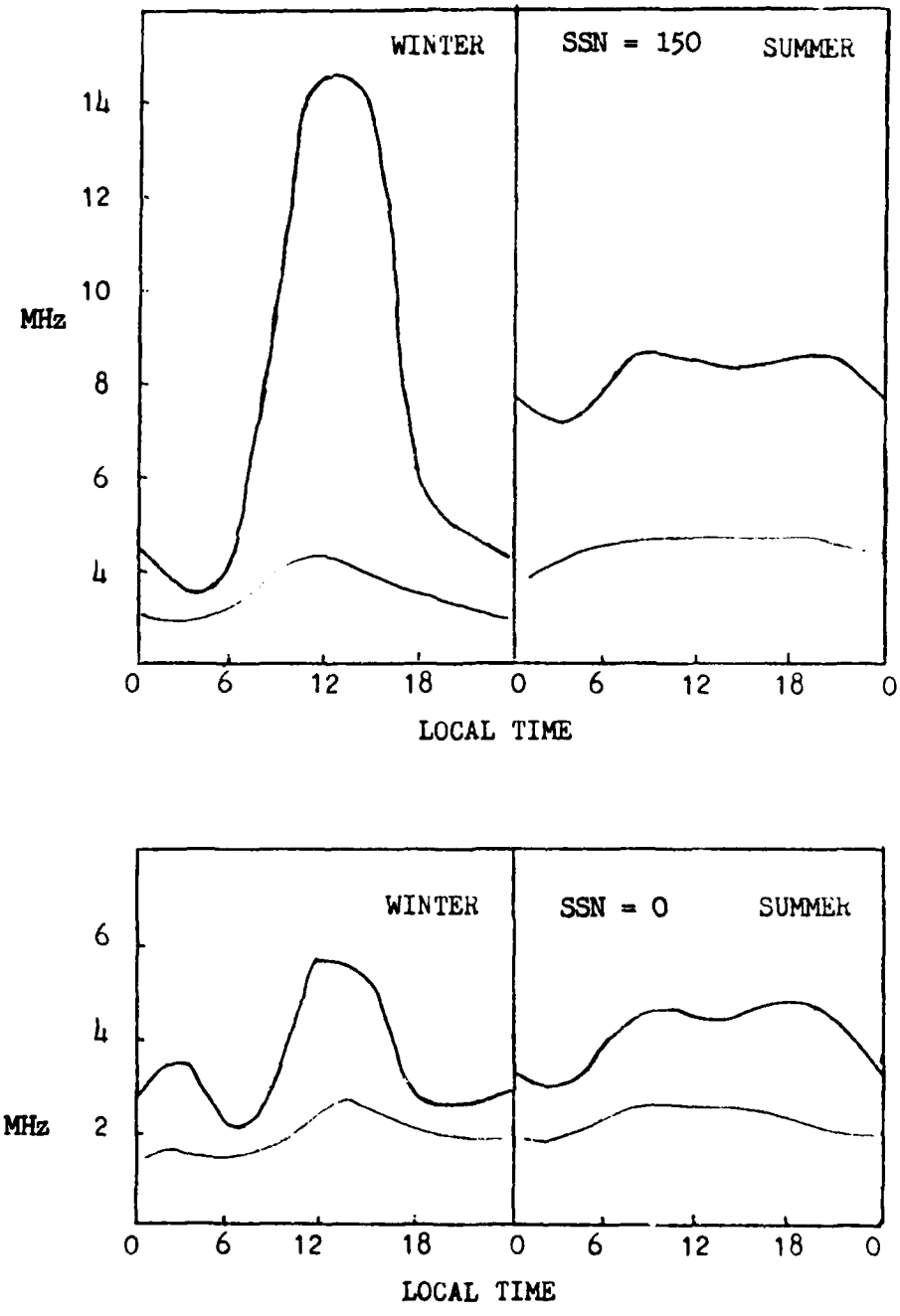



Figure 13.13 Comparison of Typical Winter and Summer "Window" Charts for an Upper-Middle Latitude HF System.

13.5.4.2 Structural Anomalies

Since both the Southeast Asian and South Atlantic geomagnetic anomalies involve vertical transport of charged particles, ionospheric abnormalities are common in these areas. Tilts, spread F, and deviative absorption are probably more common in these areas than they are in the middle latitudes. Of course, the auroral oval is probably most prone to structural disturbances, because precipitation, ionospheric currents, and sunlight are all considerably more

variable in time and location. These problems all impact primarily the F region. This means depth of F layer penetration provides some measure of the degree of impact on a given frequency. Higher frequencies should be most severely affected. The effects are also somewhat similar in that all three phenomena produce variations in both the vertical and horizontal electron density gradients at a given point. This destroys path symmetry and induces path geometry variations.

Multipath interference, fading, MUF failure, and non-great-circle propagation are all likely. Which effect is viewed really depends more on how the effect is viewed, since all the phenomena occur (or are possible) simultaneously. A radar might see non-great-circle propagation. At the same time, a communicator working near the FOT would probably experience intermittent fading. A high speed data link would experience increased bit/parity errors, and an HF system working at the MUF would suffer greater than normal MUF failure (50% would, of course, be normal, by definition of MUF). In order to determine the impact on a given system, it is necessary to determine how the system depends on the ionosphere. Tilts, spread F, and deviative absorption can also occur in the middle latitudes. Regardless of their location, the effects of these phenomena on HF operations are similar. At the middle latitudes, the impact may be more severe, because the event is not so common as elsewhere.

The Traveling Ionospheric Disturbance (TID) is also a structural anomaly. It is usually considered to be a broad wave of (in) electron density--usually in the F layer. Its east-west extent may exceed 3000 km, while the wavelength may be hundreds of kilometers or less. This yields a typical period of 2-20 minutes and a duration of less than 2 hours. A TID may result from an auroral substorm or a REP. In these cases, the TID moves equatorward and spreads. The sharp tilts and sudden density variations it produces are, in impact, similar to a SID, but they are much more limited in areal coverage.

Several stations may see the effect at different times and magnitudes, depending on local geometry. Electron density variations of 5% are common. This means that the effect on critical frequencies may be almost invisible--on the order of 2%-3%. The key to monitoring a TID--or identifying it--is looking for a change in the hourly trend of critical frequencies at a given site. The impact of associated tilts and height variations may be much more significant. Complete circuit outage is possible for short periods of time and may affect only selected frequencies. Often, it will not be SID-like (lowest frequencies affected most) in its effect. Such an outage results from the changing geometries involved. TIDs may not be observed as a consequence of the very course time resolution of current ionospheric observations.

While auroral activity is a common source of TIDs, other origins are also possible. Earthquakes, solar eclipses, flares, and gravity waves may all generate TIDs as a side effect of their primary influence. Such TIDs may travel with prevailing neutral winds and circle the globe once or more before dying out completely. Each TID is different and will affect each system differently. It is this extreme variability which limits our ability to specify TID characteristics with even the accuracy with which we discuss a flare-induced SID. Current knowledge of TIDs is summarized in Table 13.2

### 13.5.5 Sudden Ionospheric Disturbances (SIDs)

Sudden ionospheric disturbances differ markedly from the anomalies just discussed. SIDs are generally constrained to the sunlit hemisphere (though not always) as opposed to a fixed geographical locale. They are directly associated with energetic solar flares and are a consequence of the enhanced EM emission of these flares. Although all wavelengths show enhanced emission during a flare, the largest enhancements are typically at x-ray ($10^3$X) and radio ($10^3$-$10^4$X) wavelengths. These bursts deposit most of their energy in the D and E layers and end within, at most, a few hours. The primary impact of a SID is on the D region, though other layers may be affected. Like the emission which produces it, the SID shows a sudden increase followed by a slow decline. D region ionization is normally due to a balance between solar emissions and recombination. At the onset of a flare, this balance is briefly upset, and ionization climbs rapidly. The strongest SIDs usually result from large, rapidly rising x-ray bursts. Slowly rising events have much less impact on the ionosphere.

The radio burst associated with a flare may extend into the upper portion of the HF spectrum. If the sun is visible to an HF receiver antenna during such a burst, the receiver will experience a sudden increase in broadband radio noise. This will degrade the S/N ratio, perhaps below the point of usability. Since this effect works from the higher to the lower HF frequencies, it will affect frequencies near the MUF first.

The shortwave fadeout (SWF) is the most classical of all HF SID effects. A SWF effectively raises the LUF on a sunlit HF path by increasing D region absorption (via increased ionization). Since D region absorption is inversely proportional to wave frequency, SWFs spread upwards from the lower frequencies. A SWF may alone be sufficient to eliminate propagation on a particular path, or it may combine with a radio burst to limit the usable frequencies to a few in the center of the HF band. As if this weren't sufficient, the effects of both the SWF and radio burst are somewhat system dependent. Altering system/antenna configuration or increasing effective transmitter power may mitigate the SWF or burst effects. A SWF's effects depend on the portion of the path which is sunlit (a D region transit must be sunlit for the SWF to have any effect on the path) and the zenith distance of the sun as seen by the D region transit points (greater solar altitude means greater effect). The largest fades will occur for completely sunlit paths. SWF monitors attempt to select stations to maintain the total path in daylight. Ideally, the transmitter monitored will be strong and of constant output. In practice, this doesn't always occur. The absorption in dB resulting from a SWF is given in terms of the signal strength before the SID ($P_a$) and during ($P_b$) by:

$$\text{Absorption} = 10 \log 10 \, (P_a/P_b).$$

A sudden frequency deviation (SFD) may not be apparent on many HF systems because of the large bandwidths. Impact is most probable on high speed data links Doppler systems (e.g., a radar), and on narrow band systems. The typical SFD produces a frequency change of a few Hertz over a few minutes. It results from changing the index of refraction (hence, path length) of the ionosphere through which the signal passes. For lower HF frequencies, it may actually alter the reflection altitude of the control point.

Table 13.2 TRAVELING IONOSPHERIC DISTURBANCES

| TYPE OF DISTURBANCE | WAVELENGTH AND STRUCTURE | MOTION | PERIOD | FREQUENCY OF OCCURRENCE | SOURCE |
|---|---|---|---|---|---|
| Large-scale | 1000 km horizontal wavelength<br><br>Wave front width on order of 1000 km.<br><br>Phase fronts tilted nearly horizontal.<br><br>Retains shapes over thousands of kilometers. | 300 m/s north to south. | 30 min - 3 hr usually 1-3 cycles. | Infrequent, less than daily. | Events in the auroral zone.<br><br>Strong correlation with magnetic activity. |
| Medium-scale | 10's-100's km horizontal wavelength.<br><br>Wavefront width 100's to over 1000 km.<br><br>Phase fronts tilted 30°-60° from vertical.<br><br>Do not retain shapes well over distances 100 km; energy does propagate globally. | 100-250 m/s variable directions, with seasonal trends. | 10-100 min several cycles to trains. | Daily, more common in daytime. | Tropospheric phenomena.<br><br>Upper atmospheric and polar winter sources. |
| Small-scale | 10km horizontal wavelength.<br><br>Structure not well resolved. | 100-250 m/s (est) variable directions. | 10 min long trains to families as wavelength decreases. | Daily | Probably tropospheric; not well established. |

The Sudden Cosmic Noise Absorption (SCNA) is really a riometer event, since only these instruments routinely monitor background levels of cosmic radio noise at 30-50 MHz. An SCNA is just like a SWF: increased D region absorption. It's just observed by a different system.

A geomagnetic crochet, also known as a sudden flare effect (SFE) is the only SID which may be observed on instruments in the nighttime hemisphere. It isn't really an HF effect but does result from a flare-induced ionospheric variation. A very impulsive flare may produce a sufficient enhancement of the E layer free electron population to produce a momentary electrical current (like an electrojet). This current (like a geomagnetic storm ring current) will alter the geomagnetic field strength measured at magnetometer stations in the region where the current flows. For a sufficiently impulsive flare, the current may briefly extend into the nighttime hemisphere and produce a magnetometer response on sensors located there.

As with many other ionospheric anomalies, the primary difference among SIDs is not how they are formed but how they are observed.

### 13.5.6 Ionospheric Storms

Not all ionospheric disturbances end within a few hours of their onset. Some persist for several days and cover large portions of the globe. These large-scale disturbances are loosely termed ionospheric storms. Some are related directly to the energetic particles produced by a large solar flare, while others occur without apparent cause. Flare particle storms include PCAs and geomagnetic storms. These disturbances affect the geophysical environment in many ways.

#### 13.5.6.1 HF Effects of a PCA

A polar cap absorption event is a widespread, long-lived increase in non-deviative absorption confined to the polar ionosphere. Although the effects are similar in both poles, there is insufficient evidence to confirm or deny a one-to-one comparability, particularly in fine scale. High energy protons (greater than about 5 MeV) generated by a solar flare travel to the earth in 1-2 hours. These particles are guided by the earth's magnetic field lines into the polar caps. Here, they penetrate to altitudes as low as 50 km before giving up their energy in ionizing neutral atmospheric constituents. The resulting increase in electron density strengthens and lowers the polar D region and increases the absorption of HF signals passing through the polar D region. As with a SID, the effect is an increase in the LUF on polar sky wave paths. If proton fluxes possess sufficient energy density (high numbers and high energies) the resulting LUF will exceed the MUF. This shutdown of the HF window is known as a polar blackout.

PCA absorption typically begins in small patches and gradually expands to fill the polar cap. Absorption will usually be most intense in the sunlit polar cap (recall the ionosphere may be sunlit even when the surface is in darkness). A PCA is measured on 30 MHz riometers, so the absorption recorded will <u>not</u> be representative of the absorption encountered on lower HF frequencies, since PCA absorption is inversely proportional to frequency. PCA absorption also shows at least some relation to the energy density <u>gradient</u> of proton fluxes.

A "classical" PCA will produce the most intense absorption shortly after its onset. Absorption values will show a diurnal variation, with daytime values averaging 2X to 4X those at night. Higher daytime absorption results from solar radiation increasing D region ionization. Attenuation will reduce both the signal and noise levels on any given frequency, unless the noise is produced by a nearby source. The PCA will again become patchy as it dissipates. Limited sensor coverage severely restricts SESS ability to define the extent of PCA absorption. The patches may blanket a riometer but not be a problem for a communicator working another HF path. Similarly, a user may experience problems while sensors discern insignificant absorption.

13.5.6.2 Geomagnetic Storms and HF Effects

Shock waves and associated high density energetic (KeV) plasma streams produce geomagnetic storms. These storms may profoundly alter ionospheric electron densities over large segments of the middle and high latitudes. Since the alteration may be in heights as well as densities, the resulting impact on HF systems can show significant time variability. The discontinuous nature of geomagnetic disturbances exacerbates the variability of the ionospheric response. Ionospheric storms may occur in the absence of a geomagnetic storm. The origin of a purely ionospheric storm is uncertain.

A typical (geomagnetically-induced) ionospheric storm will begin almost simultaneously with a geomagnetic disturbance and often persists for sometime beyond the end of the geomagnetic disturbance. The ionospheric storm onset delay may range from 10 minutes in the auroral zone to 6-12 hours in the middle latitude evening sector following onset of the geomagnetic disturbance. The storm may persist for only a day at solar maximum to as long as a month near solar minimum due to the restorative power of solar background emission at sunspot maximum. It is important to note that the ionospheric effects of a geomagnetic disturbance are sometimes unrelated to the severity of the geomagnetic disturbance. In fact, rapid MUF ($f_oF2$) declines are occasionally observed in the absence of geomagnetic activity. The latter phenomenon seems most common after an extended period (1-2 weeks) of abnormally quiet ($A_p$ less than 8) geomagnetic conditions. A slight increase in geomagnetic activity (to unsettled conditions for 12-18 hours) is often sufficient to eliminate most of the depressions.

Although a geomagnetically incited ionospheric storm can affect the entire ionosphere, only bottomside effects are significant to HF operations. Storm characteristics are a function of local time, latitude, and season. Occasionally, significant differences are observed between the eastern and western hemispheres as well. Each storm will be unique, but most storms can be divided into two classes on the basis of their effects on the $f_oF2$ (MUF).

13.5.6.2a Negative Storms

Negative storms are probably the most common. The storms are characterized by a brief, (few hours) sharp rise in MUFs in the daylight mid-latitudes. The nighttime hemisphere is usually beset with nearly immediate MUF depressions. As the storm progresses, depressions will replace enhancements as sunset occurs, and enhancements may give way to depressions in

all but the afternoon sector of the middle latitudes. (Note, the discussion to this point has been limited to middle latitude effects.) In general, storm effects are greatest near the equinox. Equinoctual storms often initiate the transition from a winter to a summer ionospheric pattern. Winter disturbances are the most intense but tend to be constrained to the upper middle and high latitudes ($50^{o}$ - $60^{o}$ latitude). Summer disturbances are more moderate in their impact but may extend to within $20^{o}$ of the equator. Ionospheric effects spread equatorward at night and withdraw poleward in the sunlit hemisphere regardless of season.

As suggested, most negative storms have at least a brief positive phase at all latitudes. The magnitude and duration of the effect will vary from storm to storm, but Table 13.3 provides a general summary by disturbance magnitude, latitude, and time from storm onset. Figure 13.14 summarizes the time and magnitude of maximum $f_{o}F2$ depressions by latitude.

A change in $f_{o}F2$ is generally closely correlated with MUF drops or failures on oblique paths. Calculated MUFs (based on M factors determined by a VI sounder) are often suspect. Virtual height rises of as much as 600 km have been recorded with the onset phase of some storms, while actual heights are nearly constant ($\pm$20 km). This variation in virtual height is thought to result from increased signal retardation (due to increased electron density) at and below the reflection level. Careful analysis of available data is required to separate these effects.

Regardless of season, propagation effects of a negative storm are most severe at high latitudes, particularly near the auroral oval. Here, particle precipitation releases nearly 70% of the storm energy as joule heating. The precipitation also produces a large increase in E and D region electron densities, although E region heights seem to remain stable to within $\pm$0.5 km. The result is usually intense, blanketing sporadic E and absorption (D region) near the oval. Greatest intensity is normally found between midnight and local dawn. Depending on path geometry and frequency, LUF increases, multipath interference, and non-great circle propagation can be expected. The subauroral trough is more sharply delineated than on quiet days and is pushed to lower latitudes by the expansion of the oval and its equatorward shift.

Two additional phenomena are associated with auroral zone ionospheric storms. Auroral Zone Absorption (AZA) is loosely related to geomagnetic disturbances, though it may occur in their absence or as a part of an ionospheric storm. It is a D region phenomenon which can attain an intensity comparable to that of a strong PCA. Its temporal and spatial characteristics are very different. While PCA absorption changes slowly, AZAs vary randomly in minutes. AZA duration is usually measured in hours (vs days for a PCA). The AZA is typically confined to a latitudinally narrow band ($1^{o}$-$2^{o}$ wide) parallel to and just equatorward of the oval. Maximum AZA intensity occurs near 0800-0900L, but its effects extend 0400-1200L. During high levels of geomagnetic activity, AZAs will migrate equatorward ahead of the oval. AZA absorption may result directly from KeV electrons ionizing ionospheric constituents. It may also occur in response to bremsstrahlung x-rays produced by 1-40 KeV electron precipitation. (Slightly lower energy electrons account for the visual aurora).

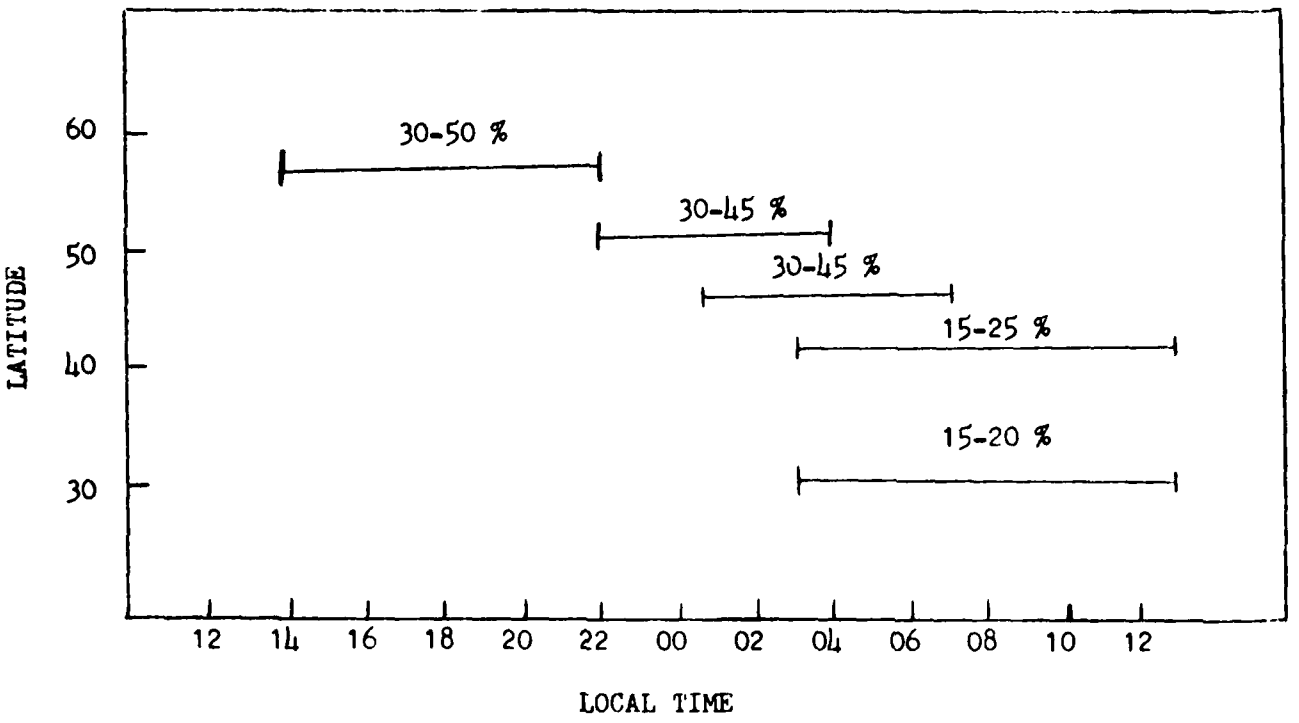



Figure 13.14 Local time of maximum $f_oF2$ depression resulting from a negative ionospheric storm (after Manley 1981).

REPs (Relativistic Electron Precipitation) are often part of an ongoing AZA. Produced by electrons of 100 KeV-1 MeV, REPs are highly localized in space and time. They are usually associated with storm recovery phase and last a few hours or less. Absorption results from bremsstrahlung x-rays which penetrate to 50-80 km (compared to 90 km for AZAs). Additional EM radiation is produced across the spectrum and produces increased radio noise. REPs are also thought to be a point of origin for TIDs which move equatorward from the auroral oval. Concentrations of lower energy electrons involved in the AZA may also be responsible for TIDs. The effects of AZAs and REPs differ primarily in areal and temporal extent though both are associated with the auroral oval.

Substorming may also produce a negative ($f_oF2$ depressions) tongue extending equatorward from the sunlit oval. This tongue most commonly occurs in the summer hemisphere and then corrotates with the earth. The night and winter sectors see less impact (less likelihood of tongue formation), because most joule heating occurs in the summer hemisphere cusp. As this tongue (tied to a geographic region) rotates into the nighttime hemisphere, it may confuse analysis of the overlying ionospheric storm.

Classical ionospheric effects of a geomagnetic storm do not extend poleward of the auroral oval. Nonetheless, some disturbances are accompanied by an increase in polar sporadic E. This sporadic E often attains blanketing intensity and produces effects similar to blanketing E at other latitudes. It

## STORM-TIME $f_oF2$ DEVIATIONS

### $f_oF2$ Changes

### Equinox

21 Mar - 20 Jun
21 Sep - 20 Dec

| Geomag Lat | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 | 60 | 66 | 72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Strong (Ap 50)** | | | | | | | | | | | | |
| 45-62 | 10 | -25 | -30 | -30 | -25 | -25 | -15 | -15 | -15 | -15 | -15 | -15 |
| 10-45 | +5 | +5 | +0 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | +0 |
| 10 | +0 | +5 | +5 | +10 | +10 | +10 | +10 | +10 | +10 | +10 | +10 | +10 |
| **Weak (Ap 50)** | | | | | | | | | | | | |
| 45-62 | +10 | +10 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 |
| 10-45 | +5 | +10 | +10 | +5 | +5 | +10 | +10 | +5 | +0 | +5 | +5 | +5 |
| 10 | +10 | +10 | +0 | +5 | +10 | +10 | +10 | +10 | +10 | +10 | +10 | +10 |

### Winter 21 Dec - 20 Mar

| | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 | 60 | 66 | 72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Strong (Ap 50)** | | | | | | | | | | | | |
| 45-72 | +0 | -15 | -20 | -30 | -25 | -25 | -25 | -20 | -20 | -15 | -15 | -15 |
| 10-45 | +0 | +10 | +10 | +10 | +10 | +10 | +10 | +10 | +10 | +10 | +5 | +5 |
| 10 | +5 | +5 | +5 | +10 | +10 | +10 | +10 | +10 | +0 | +0 | +0 | +0 |
| **Weak (Ap 50)** | | | | | | | | | | | | |
| 45-62 | +5 | +5 | -15 | -15 | -10 | -10 | -10 | -10 | -10 | -10 | -10 | -10 |
| 10-45 | +10 | +15 | +20 | +20 | +20 | +10 | +10 | +10 | +10 | +5 | +5 | +5 |
| 10 | +5 | +5 | +5 | +10 | +10 | +10 | +10 | +10 | +0 | +0 | +0 | +0 |

### Summer 21 Jun - 20 Sep

| | 6 | 12 | 18 | 24 | 30 | 36 | 42 | 48 | 54 | 60 | 66 | 72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Strong (Ap 50)** | | | | | | | | | | | | |
| 45-62 | +5 | -30 | -40 | -45 | -45 | -40 | -30 | -20 | -20 | -15 | -15 | -15 |
| 10-45 | +10 | -10 | -20 | -25 | -30 | -25 | -20 | -15 | -15 | -15 | -10 | -10 |
| 10 | +0 | +5 | +5 | +10 | +10 | +10 | +5 | +5 | +5 | +5 | +5 | +5 |
| **Weak (Ap 50)** | | | | | | | | | | | | |
| 45-62 | +5 | -5 | -20 | -20 | -20 | -10 | -10 | -5 | -5 | -5 | -5 | -5 |
| 10-45 | +5 | +5 | +0 | +0 | -10 | -20 | -10 | -5 | -5 | -5 | +0 | +0 |
| 10 | +5 | +5 | +10 | +5 | +5 | -20 | +0 | +10 | +10 | +0 | +0 | +0 |

Table 13.3 $f_oF2$ Variations by latitude, season, and storm magnitude based on time in hours after onset of the associated geomagnetic disturbance. Results are based on compiled statistics from a number of storms (from Manley 1981).

results from the poleward expansion of the auroral substorm during its expansive phase. The particle precipitation associated with this substorm produces increased E layer ionization which appears as sporadic E. Polar E occurs even more commonly during extended geomagnetic quiet. Its effects are identical to those of sporadic E at any other latitude.

Equatorward of the auroral oval, the primary storm effects are on the MUF. (Some increase in absorption and blanketing sporadic E is observed in the upper middle latitudes, and the Southeast Asian anomaly seems to show an increase in D region absorption near midday. Increased absorption is also likely near the South Atlantic anomaly.) MUF depressions mean slow MUF rises with sunrise and rapid MUF dropouts at sunset (a consequence of turning off F1 production, while storm induced transport and recombination persist unabated). Sunrise transition problems are exacerbated, and a mild sunset transition problem may develop. Figure 13.15 provides a generalized summary of storm effects on usable HF frequencies. Effects will be worse on multi-hop and east-west oriented paths. Conversely, low latitude propagation will usually improve as the Lorentz forces (V x B) resulting from the enhanced ring current drive additional electrons into the subequatorial ridges.

The overall impact of the negative ionospheric storm includes an intensification of existing ionospheric gradients and the creation of some new ones. This is probably accompanied by an increase in multipath and non-great-circle propagation, but since these problems also result from storm-induced sporadic E it is difficult to identify the actual source. Negative storm depressions may persist for weeks at solar minimum or be limited to a fraction of a day at solar maximum. Recovery time is directly tied to the background level of solar emissions and storm intensity.

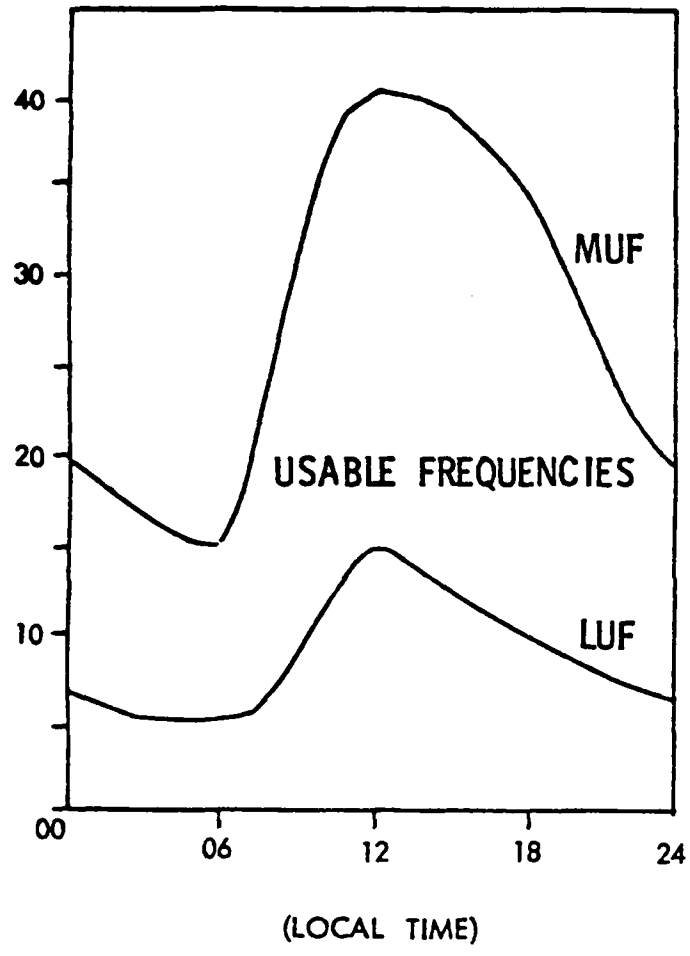
Solar emissions ensure that the ionosphere can never be completely eliminated. In other words, F1 layer production can only be turned off by sunset or an eclipse. This means that repetitive storms (several successive disturbances) of the same magnitude will probably not have equivalent ionospheric impact. Some minimum electron density is eventually attained, and additional geomagnetic activity will produce either enhancements above or oscillations about this value. No minimum climatology exists, but seasonal and solar cycle variations in the minimum are likely.

#### 13.5.6.2b *Positive Storms*

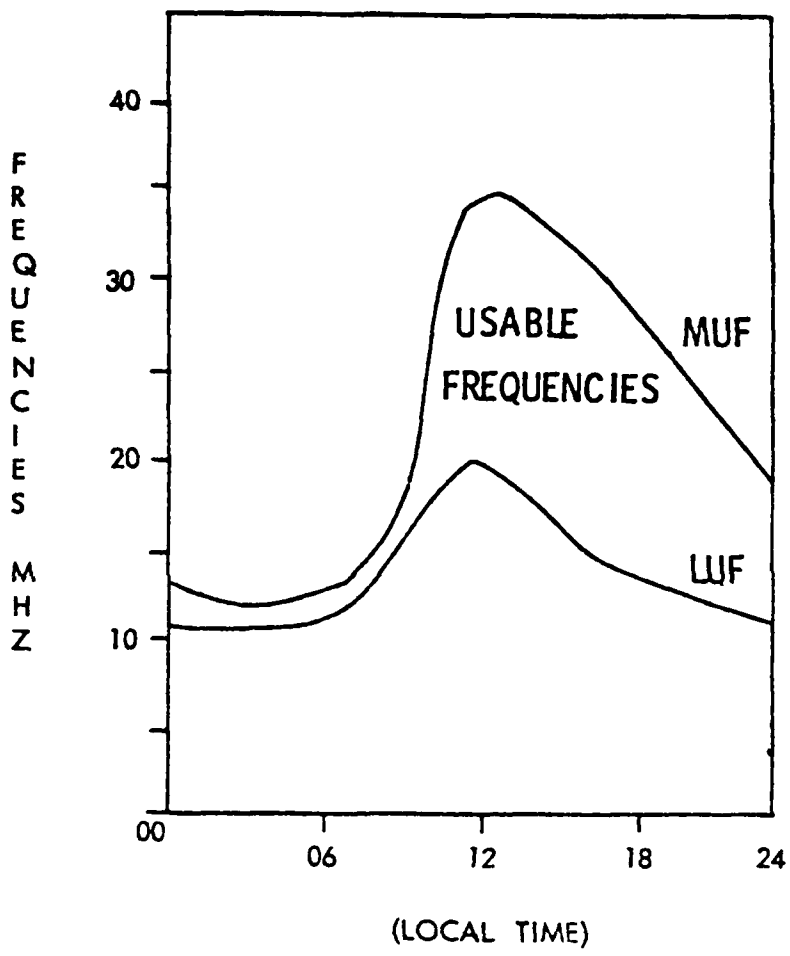
The second class of ionospheric storm is the positive storm. $F_0F2$ enhancements associated with these storms usually occur in the middle latitudes during the winter months. Positive storms are not common and probably result from complex transport processes. Ionospheric variability, the scarcity of observations, and the positive storm's rarity have thus far precluded development of a meaningful climatology.

### 13.5.7 HF Summary

The ionosphere is of major importance to HF operations. It simultaneously provides HF's long range capability while accounting for the bulk of operational variability. Path changes may easily mimic other types of changes in terms of their system effects. In many cases, the quiet ionosphere




QUIET IONOSPHERIC CONDITIONS




DISTURBED IONOSPHERIC CONDITIONS

Figure 13.15 Ionospheric Storm Effects on Usable HF Propagation Window.

accounts for as many difficulties as do disturbed conditions. Sudden ionospheric disturbances alter sunlit HF operations. Conversely, geomagnetic disturbances and polar cap absorption events can both affect nighttime operations as well. PCA effects are limited to the region poleward of the auroral oval. Geomagnetic disturbances are limited to those areas outside the polar caps. They may also trigger small, traveling disturbances.

## 13.6 Transionospherics - VHF, UHF, and SHF

Ionospheric electron densities are usually insufficient to reflect frequencies above about 30 MHz. These frequencies may be transionospheric and propagate by line-of-sight. Scattering provides limited over-the-horizon capability. These frequencies are generally used for line-of-sight communications (microwave, aircraft, etc.), radars, and satellite communications systems. System expense does not greatly exceed that of lower frequency bands, but VHF and higher bands provide considerably greater bandwidths.

Even though these bands may be transionospheric, they a not immune to ionospheric effects. Low signal strength and limite range beyond line-of-sight are consequences of inefficient scattering. id fading and associated scintillation plague line-of-sight operation i this frequency range. The path vagarities introduced by scattering result nterference of mutually independent channels and produce sufficient signa ss to limit effective sky wave range to about 2000 km.

### 13.6.1 Sky-Wave Propagation of VHF

Ionospheric effects generally fall off rapidly with increasing frequency above about 100 MHz. Consequently, only the VHF band has significant ionospheric dependence. Troposcatter, meteor burst, and line-of-sight are the most common VHF sky wave propagation modes. D and F region scattering and troposcatter can provide continuous propagation over 1000-2000 km at VHF frequencies. Sporadic E and the ionized meteor trails also permit VHF scattering, but are somewhat less reliable.

Of the possible modes, D region scattering is the primary means of VHF sky wave propagation. Scattering altitudes range from 70 km in the daytime to 85-90 km at night. The nighttime height rise is associated with a change in the primary ionization source from solar radiation to meteors and cosmic rays. The electron (and neutral) density "clumpiness" (which accounts for VHF scattering) is due to high altitude turbulence, neutral winds, and electrojets. Both these features and ambient electron density show strong diurnal variability. This means VHF signal strength should also show a strong diurnal variability. At middle latitudes, the maximum signal strength is usually achieved near midday. Minimum values occur in the late evening when the D region all but disappears in many locations. (That post-midnight values are not lower is probably a consequence of the increased meteor intensity between local midnight and noon.) A sharp midday peak in VHF signal strength during winter months gives way to a broad summer maximum. A broad midday maximum is common throughout the year at high latitudes and is centered earlier in the day than at lower latitudes. Day-to-day variability is greater at high latitudes than at middle latitudes.

Within 20° of the geomagnetic equator, F layer scattering replaces D layer scattering as a primary VHF propagation mode. F layer scattering occasionally provides ranges approaching 4000 km. It usually occurs at the bottom of the F region and is very aspect sensitive. Slight variations in angle of incidence with respect to the local magnetic field can produce significant signal strength variability. F layer scatter is often associated with spread F occurrence and shows strong Doppler shifts due to the eastward drift of the spread F cells. Unfortunately, spread F scatter also results in flutter fading. Fading rate increases with bandwidth and frequency and is a maximum between 1800L-2100L. Off-great circle paths are most strongly affected. Resulting path loss is tied to geomagnetic activity (proportional to $K_p$) but is generally unrelated to sporadic E or the occurrence of HF fading.

Sporadic E scattering is thought to be an important source of interference. It is occasionally useful for propagation over paths up to 2000 km, but is not particularly reliable.

### 13.6.2 Disturbance Effects on VHF Skywave

VHF sky wave responds to ionospheric disturbances very differently from HF systems. The differences are primarily due to the difference in control point altitude. SIDs produce two competing phenomena in the VHF spectrum. Increased D region ionization produces increased absorption and more efficient scattering. The lower end of the VHF band is primarily affected by absorption; while the high frequency end shows a significant increase in signal strength. The middle of the band shows little or no response to SIDs, the competing effects apparently cancelling each other.

Polar cap absorption events may produce initial signal enhancements in the nighttime cap. These initial enhancements typically give way to intense absorption in the daylight sector for moderate and strong PCAs. Absorption dominates for PCAs, because the increased ionization of a PCA occurs at a lower altitude (near 60 km) than does the VHF scattering.

The altitude of increased absorption also determines geomagnetic activity effects on VHF sky wave operations. Precipitating storm particles usually deposit most of their energy above 90 km (REPs are an exception, and may produce intense, localized VHF absorption). This means the geomagnetic activity will often provide increased signal strength in the VHF band (due to increased scattering efficiency) while effectively eliminating HF operations by LUF enhancement.

Radio aurora is a radio wave phenomena which affects both VHF sky wave and line-of-sight operations. It is the ionization (produced by precipitating particles) which reflects radio waves. It is not always found in one-to-one correspondence with visual aurora, nor would we expect it to be. Radar aurora, unlike optical aurora, is both ionization and aspect dependent. It has an average height of 110 km (close to the visual aurora) and varies from 75-135 km. Auroral returns are often detected within $\pm 20^0$ of perpendicular to geomagnetic field lines. Radar aurora impact is dependent on both the control (scattering) point and the transmitter location. By reflecting a portion of the transmitted energy back to the radar, radar aurora raises the

noise level. Locating a target against such a background necessitates a stronger target reflection than auroral return. Radar aurora also inserts Doppler shifts into its reflections, thereby mimicing a real target and further complicating detection. Auroral clutter may effectively jam some radars under certain conditions.

### 13.6.3 Satcom and Scintillation

Early proponents of satellite communications (satcom) systems suggested they would combine the range of HF with VHF bandwidth while eliminating the ionospheric variability associated with HF and VHF sky wave. UHF and SHF have provided additional steps in this direction by increased ionospheric "immunity." The primary ionospheric effect on transionospheric (satcom) systems is scintillation.

#### 13.6.3.1 Scintillation Origin and Specification

Scintillation is a rapid, usually random variation in signal amplitude, phase, or both. Frequencies above the critical frequency are those of concern, and phase scintillation may be more of a problem than amplitude variations. Scintillation is thought to result from abrupt variations in electron density along the signal path, but any sudden ionospheric shocks may also produce it. These changes produce rapid signal path variations and defocusing. While variations over the entire path are important, the most significant variations occur near the F2 peak between 225-400 km. A smaller change in electron density is sufficient to produce scintillation at lower VHF frequencies. Higher VHF scintillation usually requires much larger electron content discontinuities. Threshold TEC changes for scintillation at 40 MHz are about 6%, while a 30% change is necessary for 150 MHz scintillation. Content changes may result from any number of quiet or disturbance-related phenomena. Height changes do not generally produce scintillation. F region irregularities can, however, produce scintillation, and range spread F shows a close correlation with VHF scintillation. Sporadic E is also loosely related.

Scintillation is system and equipment dependent. Various indices have been defined (by Hawkins, 1974) to specify ambient conditions and include the scintillation index, S.I.

$$S.I. = \frac{P_{max} - P_{min}}{P_{max} + P_{min}} \quad X\ 100\%$$

where

$P_{min}$ = third lowest amplitude minimum in a 15 minute period, and

$P_{max}$ = third highest amplitude maximum in a 15 minute period. Scintillation data is typically cataloged in terms of the ionospheric penetration point coordinates. The third highest and lowest values are chosen as more representative of prevailing conditions than the extremes in either direction. Notice that if amplitude does not vary (no amplitude scintillation) $P_{max} = P_{min}$, and S.I. is zero.

13.6.3.2 Scintillation Climatology

Scintillation analysis is best done by geomagnetic latitude bands, with breaks at 20° and 60° from the equator. Much early climatology evolved relating scintillation to spread F in general and were based on areally-limited observations. More recent work seems to suggest that scintillation is most clearly related to range spread F. It shows considerable variability.

Current research suggests that the effects of scintillation are most pronounced in the equatorial (± 20°) latitude belt. Here, it seems closely associated with range spread F and attains maximum intensity between 2100L-0200L. The onset is abrupt, with rapid, deep fading giving way to slower fading on a given frequency. The phenomenon may persist for 20 minutes to 2 hours at a given location. As at other latitudes, space and time variability provide the most effective tools in maintaining communications. Frequency and polarization changes (unless it is a large frequency change) have little effect.

The effects of geomagnetic activity and season on diurnal scintillation are shown in Figure 13.16 (Rastogi, et. al., 1981). For equinoctual months, scintillation shows a sharp increase at sunset rising to a maximum near 2100L. This is probably associated with the sharp terminator discontinuity at the equinox. Both types of spread F respond to this variation. Scintillation occurrence then declines into sunrise. Notice that geomagnetic activity,

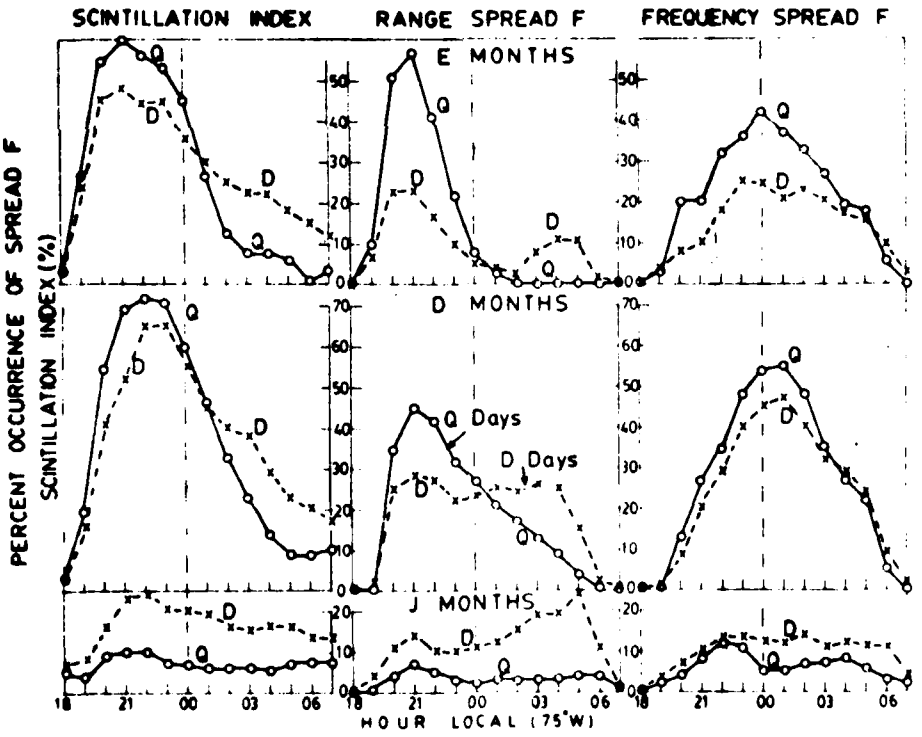



Figure 13.16 Nocturnal variations in S.I. for geomagnetically quiet (Q) and disturbed (D) days. E months are March, April, September, and October. D months are November-February, and J months include May-August (from Rastogi, et. al., 1981).

while reducing the occurrence of scintillation (and spread F) in the evening, increases its pre-sunrise intensity. Scintillation response to season, local time, and geomagnetic activity more closely mirrors range spread F changes than frequency spread F. Summer months (J months) generally result in minimum scintillation and a reversal of the impact of geomagnetic activity.

Scintillation patterns are poorly defined in the middle latitudes ($20^{o}$-$60^{o}$ geomagnetic). Like spread F, scintillation is not common in this belt and shows no strong association with any single phenomena. A slight relationship (positive) seems to exist with geomagnetic activity, particularly at upper middle latitudes. Essentially anything which produces a sudden, sharp change in the ambient ionospheric conditions possesses the potential for triggering spread F/scintillation. The lack of well-defined middle latitude patterns may not mean that such patterns do not exist. These latitudes have not been well monitored in the past. Most research has favored equatorial and high latitudes.

High latitude (above $60^{o}$) scintillation shows two separate centers of interest: the auroral oval and the polar cap. Particle precipitation and continual latitude variation account for significant temporal electron content variation at a given point. The polar ionosphere is heavily dependent on cosmic ray and particle bombardment for its maintenance. Spread F is a ubiquitous consequence, yet scintillations seems not to be as severe in the polar cap itself (except, perhaps during PCAs). The combination of these influences produces a structure of intense scintillation somewhat aligned with the oval. A scintillation "boundary" seems to exist about $5^{o}$ equatorward of the oval itself. Scintillation increases poleward of the boundary to a maximum in the oval and a smaller secondary maximum over the magnetic pole. The subauroral trough is not coincident with the equatorward boundary. Moreover, hemispheric symmetry may not exist for this scintillation zone. The southern hemisphere boundary may occur several degrees higher in latitude than in the northern hemisphere. The boundary seems nearest the geomagnetic pole in summer and most distant in hemispheric winter. Maximum scintillation seems to be associated with the dayside cusp region and near the geomagnetic pole. Intense scintillation is associated with most extreme ionospheric variability. Scintillation is also a maximum on paths parallel to geomagnetic field lines and a minimum for paths perpendicular to field lines due to the typical alignment of irregularites. This means that increased geomagnetic activity, polar summer, and subauroral spring markedly increase scintillation. In fact, scintillation within the polar cap nearly ceases after a week of extreme geomagnetic quiet. Minimum scintillation is usually found immediately equatorward of the equatorward scintillation boundary.

### 13.6.4 Forecasting Scintillation

Forecasting scintillation entails climatology and forecasting ionospheric anomalies. Most anomalies are associated with sharp ionospheric discontinuities. These include the sunset terminator, the auroral oval, and the geomagnetic anomalies, among others. Identifying areas and conditions conducive to sharp changes in ionization intensity or orientation is the most important step. Climatology is the most useful tool available. Table 13.4 summarizes the various potentialities. The sharpest gradients afford the greatest potential of scintillation.

Table 13.4 Solar-Terrestrial and Temporal Dependence of Scintillation.

| Parameter | Latitudinal Range (geomagnetic) ± 20o | 20o-60o | 60o - 90o |
|---|---|---|---|
| Scintillation Character | Greatest Extremes | Slight to Moderate | Moderate to Extreme |
| Diurnal | All latitudes show maximum activity at night. | | |
| Seasonal | Maximum near Equinox with slight longitudinal variations | Maximum in Spring | Maximum in Summer |
| Solar Cycle | Increases with SSN | May be associated with rapid changes in activity level | Increases with SSN |
| Magnetic Activity | Some longitudinal Dependence: Africa decreases with $K_p$. S America decreases with $K_p$ at equinox and increases at solstice | Slight increase with $K_p$ | Increase with $K_p$ |

## 13.7 Summary

Department of Defense operations make extensive use of the electromagnetic spectrum for communications, positioning, and detection. In some cases, the ionosphere facilitates these operations while, in others, it complicates operations. The EM spectrum is divided into bands loosely aligned with the type of interaction commonly expected between the radio wave and the ionosphere. Use of radio waves is predicated in part on their ionospheric interaction and in part on their available bandwidth and required equipment. ELF VLF-LF systems operate by a waveguide mode to provide truly global capability. Their utility is restricted by equipment expense and limited bandwidth. D region height variations provide the major ionospheric impact at these frequencies. MF systems are generally short-range during daylight hours because of D region intervention. At night, MF can cover considerable distance by E region refraction. HF is a truly long range frequency band. Operating by F region refraction, HF affords good bandwidth and inexpensive operation. Ionospheric variability is of primary importance to HF systems, since their operation is dependent upon it. VHF and higher frequency bands interact only slightly with the ionosphere--generally by scattering or scintillation. Conversely, they provide reliable circuits primarily over line-of-sight distances.

Understanding ionospheric radio wave propagation is really dependent on understanding ionospheric variability. Sharp ionospheric gradients, tilts, and transient inhomogeneities are the origins of most operational concerns. Locating these abnormalities is usually done by geomagnetic latitude, with "normal" or "classical" conditions confined to the middle latitudes. Geomagnetic field anomalies intrude on the straight forward latitudinal structure. No system is immune to these irregularities, but an understanding of their effects will permit selecting the optimum frequency band for a particular location and operation.

# CHAPTER 14

# IONOSPHERIC MODIFICATION AND MODELING

If you don't like the ionosphere you've got--change it. It sounds simple, and it's not impossible to do. In fact, demonstrated capability to alter the ionosphere (and portions of the magnetosphere) already exists. From a hydrogen bomb to water vapor, there are several ways of changing selected portions of the ionosphere.

Somewhat more difficult at least at present, is accurately modeling the current ionosphere and attempting to project its variation in response to various natural and manmade influences. Several models exist, but all are of limited applicability.

## 14.1 Tests and Trapped Radiation

A nuclear explosion is a dramatic and relatively simple way of producing large scale changes in the ionosphere and, to a lesser extent, the magnetosphere. The changes are transient, but "transient" may be as long as a decade.

A nuclear blast can have certain obvious effects on command and control systems--destruction. Less obvious , but no less real, are the electronic effects. Altering the electromagnetic environment can have as serious an impact on modern warfare as outright destruction of people and buildings. A one megaton high altitude blast can double the electron density of the ionosphere (you pick the level or layer). Moreover, it may incapacitate many electronic devices within a 900 mile radius of its detonation point. These two effects are known as the ionization phenomena and the radioflash (or electromagnetic pulse--EMP).

Although the theory for determining ionospheric effects of a nuclear blast has been around for a long time, tentative confirmation awaited the last few atmospheric nuclear tests of the early 1960's. These included the Argus and Starfish blasts--intended to study nuclear weapons effects on the trapped radiation belts. Unfortunately, available instrumentaion was poor, and placement of sensors inadequate. Nonetheless, an interesting picture resulted. These tests proved that artificial radiation belts could be created at will. The requirements were proper shaping and placement of the weapon.

The Argus tests revealed that particles, primarily MeV electrons, could be injected into trapping regions at altitudes of 80-400 km (Cladis, et. al., 1977). Belt lifetime ranged from 1 day to 2 years, depending on injection altitude (higher altitude means longer lifetime).

The Starfish blast off Johnston Island in 1963 revealed that even longer-lived belts were possible. This 1.4 megaton blast was fired at 400 km. It created a narrow, crescent shaped belt at about 0.2 $R_E$ above the surface. The outer edge of this belt apparently lay between 1.3 $R_E$ and 6 $R_E$ above the surface. This highly asymmetric blast (directed upwards) injected nearly 10% of the resulting ionization into this artificial belt. The belt persisted for several years before blending into the ambient trapped radiation.

These and similar tests suggested that at least three types of trapped particles would be generated by any high altitude blast. The most common products of such a shot are untrapped, energetic electrons. These high energy particles formed a 100-1000 kilometer diameter tube permitting a single pass through the earth's equatorial plane before the particles were absorbed at the magnetic conjugate point. Though existing for only a fraction of an hour, such particles could wreak havoc on low altitude spacecraft operating in their vicinity. A second, lower energy group of electrons and ions completed one or more bounces (back and forths between magnetic conjugate points) but failed to complete one drift cycle (one earth orbit). Existing for perhaps 15-20 minutes, these moderately high energy particles would be equally hazardous to spacecraft, but over a larger area. Finally, the low energy electrons, typified by the Starfish blast, might last for years as an enhancement of the ambient radiation belts.

These artificial radiation belts pose the greatest hazard to astronauts, manned space stations, and low altitude satellites. The effects of trapped radiation are more insidious though less severe than outright destruction. Spacecraft charging, normally a high altitude phenomenon, might well be brought into the low orbit regime. The accompanying damage to solar panels, on-board electronics, and physical control systems might well disable or totally blind the intelligence gathering and communications capability of a nation dependent on satellites for these missions.

## 14.2 Electromagnetic Pulse Effects

A nuclear blast at 250 miles over Omaha could eliminate, within a fraction of a second, most power, telecommunications, and computer capability in the continental United States. Little physical damage would be apparent, yet the economic and social impacts could be enormous. The cause--EMP radiation (Raloff, 1981).

Electromagnetic Pulse (EMP) radiation is one of two EM effects of an atmospheric nuclear blast (it may also be caused by non-nuclear events, but the nuclear generated effects are the most extreme). Current estimates suggest that perhaps a millionth of the energy of a typical nuclear blast is emitted in the form of EMP radiation. Yet this small fraction was sufficient to produce the simultaneous failure of 30 strings of street lights and set off hundreds of burglar alarms in Oahu, Hawaii in 1962. These effects were in response to a nuclear test conducted 800 miles away (Raloff, 1981).

EMP produces its effects in much the same way as lightning generates static on AM radios. It induces current or voltage surges in electrically conducting materials. This means that semi-conductor circuitry is particularly vulnerable, while vacuum tubes and electric motors are progressively less susceptible. EMP radiation is basically similar to other radio waves, except that it covers a much broader frequency spectrum, and the associated electric fields can be millions of times stronger. Of course, this also means that an antenna is required to "pick up" EMP radiation (assuming you want to receive it). Unfortunately, almost any unshielded conductor attached to an electronic device will function as an antenna. Signal amplitude varies widely over EMP's frequency spectrum (0-150 MHz, approximately), but most of the energy seems to peak below 100 MHz.

Higher altitude bursts will have, in general, a shorter rise time. Hence, the higher altitude burst puts more energy into the higher frequency regime. The earlier comparison between EMP and lightning might suggest the use of standard lightning protection against EMP. It won't work. A 100 kv/m lightning stroke produces an induction field of about 1 kv/m with a rise time of 1-5 microseconds. A large, high altitude blast results in rise times near 10 nanoseconds--nearly 100 times faster. The wide spectral distribution of EMP energy and its fast rise time effectively negate power shunting devices designed to cope with lightning.

Somewhat like lightning, EMP radiation probably begins life as a result of charge separation (which, incidentally, produces the second electromagnetic effect of a nuclear blast -- the ionization). Gamma rays (extremely high energy photons) are produced by both the nuclear blast and by neutron interactions with the bomb residue. These gamma rays ionize surrounding atmospheric constituents and impart considerable energy to the newly freed electrons. The electrons, being lighter than the ions, move rapidly away from the site of the blast. This rapid charge separation produces an electric field which may rise to its maximum strength in $10^{-8}$ seconds. A variable electric or magnetic field will produce an EM wave--EMP results. The field rise time determines the wave length/frequency of the radiation.

A second theory exists for the production of the radioflash (EMP, by another name). The projection of a dense cloud of high energy electrons upward should effectively "bubble" the magnetosphere. (The cloud's energy density equals or exceeds that of the local magnetospheric field.) This bubble initiates a shockwave which produces EM radiation as it propagates through the magnetosphere. These mechanisms may operate simultaneously or in conjunction. What is certain is that an asymmetric (upwards directed) blast maximizes the EMP output.

As asymmetry maximizes EMP output, so too does altitude maximize the area affected. For a near-surface burst, EMP effects are limited to a radius of 2-5 miles from the blast. At 19 miles altitude, EMP will affect a 9 miles ground radius. The ground radius becomes approximately 900 miles at an altitude of 100 miles, etc.

## 14.3 Atmospheric Ionization Phenomena

The altitude of the blast likewise determines the effect of the ionization produced. Ionization, energetic electrons in particular, is the second electromagnetic effect of an air blast. The ionization generates the EMP radiation (sort of) and, of course, provides the material for the artificial radiation belts. It also produces several direct effects on the ionosphere.

Perhaps 10%-75% of the energy of a high altitude blast goes into ionization. The effects are greatest in the D region due to the large amount of ionizable material. The blast produces two plasma clouds (in effect). Each is similar in size to the original blast and about one-half the original intensity. The second cloud forms at the magnetic conjugate point within about 13 minutes of the first (these are the untrapped, energetic electrons mentioned earlier). Consequently, similar effects will result in both the north and south hemispheres no matter where the blast originated (i.e., which

hemisphere). The decay time of these clouds--and their effects--is dependent on altitude and local time. It may range from a few minutes at night to several hours during the day.

The exact effects on ionospheric radio wave propagation will depend on blast altitude. Generally, however, they will spread across the entire radio spectrum and may include:

(1) The blast wave will produce an ionospheric shockwave. This shock wave will alter the ambient electron density at each point in the ionosphere as it passes. Such variations will produce scintillation on VHF signals and SFD's on HF signals.

(2) A lowered D-region will result in decreased range, SPA's and SES's on VLF systems (LORAN/OMEGA positional errors).

(3) Increased D-region ionization may reduce LF and MF systems (including AM emergency broadcast stations not wiped out by EMP) to line-of-sight operation.

(4) HF systems will experience severe fading (LUF enhancements) and considerable variation in path geometry. Intermittent MUF fading is also possible due to tilts in the F layer and induced sporadic E.

(5) VHF radars should experience improved forward scatter (due to stronger D region irregularities) but increased absorption. Most radars will experience range attenuation due to absorption, noise, and lowering of their ionospheric control point (if forward scatter). Of course, path geometry may also be altered due to horizontal ionization gradients.

(6) Few effects are likely on UHF and SHF systems.

In short, ionization effects on radio wave propagation are similar to a very large solar flare. The main differences being that the effects are limited in coverage (by blast altitude) and may occur in the nighttime hemisphere if the blast is in that hemisphere. Ionization results mainly from particle-induced collisions as opposed to EM radiation.

Many of the aforementioned effects are absent in a ground burst. Of course ground bursts do have the potential for generating very long-lived fades at all frequencies by eliminating the transmitter or receiver site. There are a few less obvious impacts also--most at shorter wavelengths. A low altitude burst will inject a large amount of material into the atmosphere. This material (like raindrops) can effectively block the shorter wavelengths. In this case, lower frequencies (longer wavelengths) are the mode of choice for optimum performance. A medium altitude burst can also produce highly localized stratospheric warming with correspondingly similar effects.

## 14.4 Non-Nuclear Variations

Highly customized ionospheric modification can be achieved without the use of nuclear weapons. Such efforts might be viewed as defensive in nature from a military point of view.

Rapid ionospheric recombination can be induced by injecting selected materials into given levels of the ionosphere. The resulting loss of ionization may persist for several hours and will remain somewhat static in location. Water vapor or molecular hydrogen deposited in the F layer by the launch of Skylab resulted in a several hundred mile diameter hole in the F region off Cape Canaveral. It persisted for several hours. Similar effects might achieve a drop in D region ionization making HF communications possible below the normal LUF. Such a technique might be used to effectively eliminate HF "party lines". Using a frequency below the LUF and letting it into and out of the ionosphere at selected points via holes would achieve this. Such signals would be trapped between the D and E or F layers except at the holes. By eliminating passes through the D-region, signal attenuation could be reduced. This means the same signal-to-noise ratio at the receiver for less power input at the transmitter. Such "whispering gallery" propagation also inhibits listening in by other ground or space-borne receivers--an ideal arrangement for a military HF link. The "only" problems involve calculation of path geometry (to accurately place the holes) and the limited size and duration of the holes versus their cost. Moreover, neutral winds might shift or fill the hole at an inopportune moment.

A somewhat less costly and less precise scheme for hole-making is ionospheric heating. EM waves of frequency equal to the plasma frequency will heat the plasma. Since the plasma frequency increases (in general) vertically, one picks the level to heat by picking the frequency of the RF (radio frequency) transmitter/heater. RF heating works somewhat differently than the injection methods discussed above.

Ionospheric heating doesn't eliminate free electrons; it moves them. Adding energy to the electrons at a given altitude causes those electrons to drift away from the site of the heating. This drift creates a field-aligned ionospheric hole which persists briefly after the heater is turned off.

Each of these non-nuclear mechanisms has certain advantages and disadvantages. Certainly either is much more limited in scope and side effects than the nuclear methods. Moreover, neither creates noticeable effects at the conjugate point. All are costly in one way or another and certainly have both military and civilian application. Much research remains to be done in these areas and in the insight they provide into the morphology and functioning of the ionosphere, trapped radiation belts, and magnetosphere.

## 14.5 Ionospheric Modeling

Much of what we understand of the functioning of the ionosphere, radiation belts, and magnetosphere is incorporated into operational models. These models combine insight, observation, and science to form the basic analysis and forecast tools of the AFGWC/SESS organization. To date, modeling effort has emphasized the ionosphere because of its strategic importance to DOD.

Attempts at designing computer models of the ionosphere fall into two separate classes. Research modeling can be generally defined as the design of programs which will accurately simulate the dynamics of changing ionospheric features. Often, these models concentrate on a particular region or feature of the ionosphere in order to provide details of the physics and chemistry of that region or feature. This class of models uses ionospheric observations as

a starting point and simulates the temporal changes based on driving energy sources (solar flux and auroral input) and atmospheric chemistry. As one would expect, this approach can yield highly accurate specifications and short term predictions at the expense of very large amounts of computer time. The other class of models can be defined as morphological. These models specify the morphology, or general characteristics, of the ionosphere. In fact, many models specify the occurrence of only one characteristic; for example, the scintillation model running at AFGWC produces no products except scintillation estimates. The morphological models use inputs of solar flux, magnetic activity, etc., as indices to reference and modify existing ionospheric specifications stored as data bases. While these models are less precise and not universal in application, they run reasonably quickly on computers. The constraints imposed by available computer resources dictate that SESS models be of the morphological class. A brief description of the four major ionospheric models used by AFGWC follows. Each of the models is actually a specification model, but is used as a forecast model in the sense that the model inputs are forecast.

### 14.5.1 ITS-78

The basic ionospheric climatology system in use by SESS is ITS-78. Designed for long-term frequency management and circuit planning, the Institute for Telecommunications Services' climatology is in wide use. It is based on critical frequency and virtual height fields observed in 1958 and 1964. The available data was organized by latitude, longitude, and universal time and is loosely grouped by level of solar activity in the form of the smoothed Zurich sunspot number. The data is <u>not</u> separated by level of geomagnetic activity. Consequently, such important features as the auroral oval and the subauroral trough are greatly smoothed.

AFGWC modifications have extended the range of sunspot numbers for which the ITS data is calculated. The limited data on which ITS is based and the extensive smoothing involved in its creation limit its real-world applicability. Indexing the ionosphere to a sunspot number further confuses the issue, since the sunspot number has no direct relationship to ionospheric conditions. Nonetheless, ITS-78 fields are regularly used to specify ionospheric conditions at many locations not routinely monitored by ionosondes. The sunspot number provides the means of identifying a particular day's ionosphere or differentiating one day from another.

The problems of recreating a given past day's ionosphere from climatology resulted in the program UKFILE. Available ionospheric data for a given day (actually the day in question averaged with the preceding 4 days) is compared with the ITS data. The result of this comparison is a sunspot number (termed the <u>effective sunspot number</u> because of its relation to ionospheric conditions) identifying the ITS fields most nearly matching a given day's data. A record of these effective sunspot numbers then permits creation of a best fit ionosphere for the day in question and direct comparison with other days to derive trends. Vertical electron density profiles could then be extracted from the ITS data (or from live data, of course) using the RBTEC (Ramsey-Bussey) model. Major limitations of the UKFILE/RBTEC combination include:

(1) Only F region critical frequency and height information are used in this system. Total electron content (TEC) data cannot be incorporated into the analysis;

(2) There is no resolution for level of geomagnetic activity; and

(3) Sharp gradients, particularly near the auroral zone, are not duplicated.

Although ITS-78 climatology was only intended to support HF communications over well-defined paths, AFGWC has greatly expanded its use, if not its applicability. It is often employed as a basic input for both the Polar and 4D models.

14.5.2 Polar Ionospheric Model

The Polar Ionospheric Model was developed at AFGWC to support the experimental over-the-horizon backscatter (OTHB) radar system tested in Maine during 1980-81. The model incorporates the results of many previous attempts to model the auroral region. The primary inputs to the model are effective sunspot number and an auroral activity index ($Q_e$). The ITS climatolgy serves as the basis for the F2 fields which are then modified to represent auroral and electron density trough features not present in the ITS fields. The model uses $Q_e$ to determine the position and magnitude of these features. The design of the model is modular in order to minimize any future upgrading effort. The three major modules are FLAG, FLDSIX, and SUPPRO.

FLAG specifies whether the trough, auroral E, sporadic E, F layer irregularity zone, etc. occurs at each grid point. This information is used by later modules to determine values of six layer parameters and to build verticle electron density profiles. The auroral region is determined from a Starkov oval driven by the effective Q (derived from satellite imagery of the aurora). It is defined as the Q index which corresponds to the observed equatorward boundary. When imagery is not available, $Q_E$ is inferred from $a_p$.

FLDSIX contains most of the science in the Polar Model. It produces critical frequency and $h_{max}$ values for each of the three layers at each grid point. The initial values for the F2 layer are derived from the ITS climatology. They are modified to show auroral and trough features as follows. The auroral region densities are enhanced over the entire auroral zone with the enhancement equatorward of the statistical center being twice that poleward of the center. The maximum heights in the auroral region and the polar cap are lowered 30 kilometers below ITS values. The subauroral trough is created by multiplying the ITS values by a factor which decreases the density to a minimum at 0300L at a latitude a few degrees equatorward of the equatorward auroral boundary. The multiplication factor is dependent on Q, local time, and time of year. The density trough does not exist during the hours from 0600L to 1800L nor when the sun is above the horizon. The height of maximum density is set as 450 kilometers at the point corresponding to the minimum critical frequency. The height decreases to ITS values at 2100 and 0600 local geomagnetic time and 1.5 degrees poleward of the trough wall (assumed to be colocated with the equatorward auroral boundary). The

equatorward boundary of the trough is defined as the point at which the height values fall to within ten percent of the ITS values for a given geomagnetic latitude.

The F1 layer is assumed to be entirely of solar origin. The values for this layer are calculated internally using the effective sunspot number. The E layer is assumed to be driven both by solar and auroral inputs. The values for each contribution are calculated separately and then combined to determine the final E layer parameters.

SUPPRO accomplishes the final step in the modeling process by building electron density profiles for each grid point. The six parameters calculated by FLDSIX are used together with auroral E and trough flags set by FLAG and factors describing the topside scale height to build the profiles. These profiles extend from 90 to 500 kilometers with a verticle resolution of 10 kilometers.

The primary limitations on model accuracy result from the basic ITS constraints. $H_{max}$ in the sub-auroral trough region may be too high. Moreover, the model suggests that $h_{max}$ increases as electron density increases. Physical evidence suggests that the reverse is probably true in the trough region. Finally, the model shows no local time dependence for auroral E. Despite these limitations, the model seems to provide a significant improvement over other available models near the auroral zones. Combined with modification capability (IONMOD) and restricted to the high latitudes, the Polar model seems to do a good job of modeling the real-world ionosphere. Outside the high latitudes, the model is no better or worse than the ITS climatology on which it is based.

### 14.5.3 Four-Dimensional Ionospheric Model (4D)

The Four-Dimension Ionospheric Modeling system is the primary AFGWC ionospheric model. In use since July 1977, the 4D was developed to combine all types of input ionospheric data in a global ionospheric analysis. It is a mathematical model providing simultaneous modeling of latitude, longitude, heights, and time variations. Horizontal and time variations are manipulated in manner similar to that of UKFILE, except that spherical harmonics provide greater sophistication in reproducing steep gradients accurately. Height analysis incorporates a modified RBTEC analysis and represents vertical electron density structure using a set of eigenvectors. The combination permits inserting data at any point in time or space.

The 4D performs its analysis in three mathematically distinct steps. First, the electron density profiles are constructed for points at which ionospheric observations are available. Second, the coefficients describing the electron density profiles are connected in latitude and longitude using a spherical harmonic technique which employs Associated Legendre Polynomials. Finally, the time analysis is accomplished using trigonometric functions. The result is a set of 16,320 coefficients from which any ionospheric parameter of interest can be generated for any location, height, and time of day.

The part of the analysis which requires the greatest effort is the construction of an electron density profile (EDP) combining all the observed data types for a given location. The data types normally used are $f_oF2$, TEC, and electron density observed by a satellite-borne plasma probe near 800 km. This data and the climatological $h_{max}$ are input to a modified version of RBTEC. The model performs several iterations to adjust the topside scale height so that the TEC summed from the calculated profile agrees with the observed TEC at that location. The final step in the construction of the EDP is to represent the profile by a series summation of four orthonormal functions designed to faithfully reproduce middle latitude profiles. Another round of iterations determines the four weighting coefficients which, when applied to the four functions, result in the best fit to the previously calculated EDP. These four coefficients can then be used to reconstruct the EDP whenever desired. Once the weighting coefficients are determined for all the points, the coefficients themselves are connected in time and space as described earlier.

The actual physics of the 4D are contained in the data preprocessors. The wide diversity of possible input data necessitates some form of preprocessing of raw data so that the mathematical system can work with similar input formats. Three types of preprocessors exist in one form or another: (1) analysis of conventional ionosonde data; (2) analysis of total electron content (TEC) data; and (3) analysis of satellite data. Of course, climatological data can be substituted for live data. Among other functions, these preprocessors spread available data over a large geographic area.

Data sparsity creates considerable problems with this model. In data-sparse areas, the model is currently forced to use ITS-78 data as an input. This results in a significantly degraded analysis for reasons identified earlier. This problem also accounts for the 4D's poor performance in modeling the steep gradients present in the high latitude ionosphere - ITS data doesn't show these gradients accurately, and live data is extremely sparse in this region. Then, too, different eigenfunctions apply to the polar ionosphere than is the case in the middle latitudes. This means that coefficient fields would change discontinuously between adjacent points. Since 4D attempts to smooth coefficient fields, use of different functions is not feasible. A combination of the 4D and Polar models might alleviate some of the problems now extant in each. It will not, unfortunately, resolve the sparsity of data and the consequent reliance on ITS climatology.

The 4D, like Polar, is compatible with the IONMOD series of data bogus/modification software. This permits the 4D's use as a forecast tool. Nonetheless, the 4D contains no dynamic considerations. It, like Polar and ITS, is a static specification model. This approach is not from preference. It results from an incomplete understanding of the physical dynamics involved. Until the science can be significantly improved, the model's (4D or Polar) forecast success is tied to the forecaster's knowledge and skill.

### 14.5.4 The Ionospheric Scintillation Model

The scintillation model is the newest of the SESS models. It was developed using observed scintillation data from the Wideband satellite.

The data was collected at auroral, middle, and equatorial latitudes with primary emphasis on the auroral region. Inputs to the model are sunspot number and a magnetic activity index. Outputs are amplitude and phase scintillation indices ($S_4$ and ).

Internally, the model uses a thin screen transmission technique to describe the ionospheric irregularities causing scintillation. The shape of the irregularities is driven by the magnetic activity index. Other factors considered include the velocities and locations of the transmitter, the receiver, and the ionospheric irregularities. Of particular importance is the geometrical relationship of the signal path to the irregularities. In the auroral zone, the irregularities tend to stretch out along field lines. Consequently, transmissions along the field lines are more seriously affected than transmissions normal to the field lines. Since this model requires the precise locations of the transmitter and receiver, it is most useful in a post analysis mode. Another version of this model produces grids of the physical parameters describing the irregularities. This product is useful to customers who can supply their own coordinates in a forecast mode. The tacit assumption is, of course, that the magnetic activity index and effective sunspot number used as input can be accurately predicted, and that they accurately describe the state of the ionosphere in the vicinity of the path being studied.

The scintillation model faithfully reproduces the Wideband scintillation data used to develop it. It has also been tested against data acquired at Goose Bay and found to perform adequately. Research is ongoing to continue improvements and to find methods of incorporating near real-time satellite data into the model specification.

## 14.6 Summary

A nuclear air burst produces gamma rays and energetic neutrons. The bomb debris yields additional gamma rays and beta particles. The resulting ionization produces both EMP radiation and free electrons. Some of the free electrons may form artificial radiation belts, while others will be involved in ionospheric modification. X-rays and EUV from the thermal component of the blast will also account for localized SIDs and variable ionospheric refraction. The magnitude and range of these EM effects depends on blast altitude (maximum at high altitude) and asymmetry (maximum when directed upwards).

Ionospheric heating and artificially induced recombination provide a possible means for establishing secure HF communications. They also permit analysis of the ionized regions of the earth's atmosphere and of the magnetosphere, and prevent the potential for making highly localized ionospheric modifications.

Ionospheric observations are combined with climatology and available science to produce several ionospheric specification and analysis models. These include ITS-78, Polar, 4D, and the scintillation model. Except for ITS-78, the major limitations of these models are a lack of dynamics and the sparsity of real time observations. ITS-78 partically circumvents these difficulties by relying solely on climatology, but even this lacks universal applicability. All of these models are to some extent customer or region unique. Much work is needed to develop an accurate, whole-ionosphere model.

# CHAPTER 15

# SPACECRAFT OPERATIONS

The growing importance of space to defense operations has resulted in several new geophysical problems. The trapped radiation belts present significant dangers to both spacecraft and astronauts. Changes in the trapped radiation environment are typically associated with solar activity. Long term variations in solar activity are acknowledged as a cause in the changing structure of the earth's upper atmosphere. Such changes may also result from geomagnetic storms, with a consequent increase in drag on low-orbit satellites. Spacecraft charging, drag variations, and radiation dangers to manned spacecraft are increasingly important features of the geophysical environment.

## 15.1 Satellite Charging

Explorer I discovered the trapped radiation belts and set off a furor of concern over the potential effects of radiation on spacecraft and astronauts. For many years, these concerns seemed generally unfounded. The advent of more sensitive measuring devices and the switch to very low voltage (5 volts or less versus the earlier 28 volts) integrated circuit technology has revived this concern. Over the past few years, it has become ever more apparent that the charged particle environment may affect spacecraft. The potential magnitude of these effects was dramatically revealed by the flight of the Jupiter Pioneer spacecraft. Its encounter with the energetic particles of the Jovian radiation belts nearly destroyed many on-board systems. That the problem is not limited to interplanetary probes was revealed by the ATS-6 vehicle. Static surface potentials as high as 20,000 volts were recorded on board this satellite.

### 15.1.1 The Problem

Technically, spacecraft charging is a variation in the electrostatic potential of a spacecraft surface with respect to the surrounding plasma. The build up of large static charges may confuse or blind certain sensors, and the resulting discharge may result in structural damage. Smaller discharges have been related to a variety of problems. Included are the following:

(1) Spurious electronic switching activity (such as turning off a recorder or activating a radio);

(2) Breakdown of vehicle thermal coatings;

(3) Amplifier and solar cell degradation; and

(4) Degradation of optical sensors.

Although any vehicle operating above a few hundred kilometers (perigee) may be susceptible, the highest probability seems to be with geosynchronous vehicles. (Geosynchronous satellites have an orbital period of 24 hours,

so they pass over a given spot on the earth at the same time each day. Geostationary satellites are a subset. They orbit in the earth's equatorial plane with a 24 hour period. These vehicles appear to remain fixed over a given point on the earth's equator.) The high occurrence of charging on geosynchronous spacecraft may be an artifact of our reporting system, since a large number of high altitude satellites are in geosynchronous orbit. However, theory does suggest that orbits above about 4 $R_E$ should experience more problems.

The magnitude of the problem is vehicle as well as orbit dependent. A spherical satellite with a homogeneous, conducting surface would probably not exrerience significant charging-related problems. The utility of such a design is, of course, extremely limited. Nonetheless, vehicle design is an important consideration.

Further complicating the charge-discharge problem are cosmic rays. These extraordinarily high energy particles produce -- we think -- effects which mimic those generated by charging. Their energies permit cosmic rays to penetrate the spacecraft. Cosmic rays incident on a spacecraft computer can alter memory units and produce spurious commands. Low voltage solid state circuitry seems particularly prone to this sort of cosmic tampering. Meanwhile, a vehicle has been flown (SCATHA, Spacecraft Charging at the High Altitudes) solely to test various hypotheses regarding the charge/discharge problem.

### 15.1.2 Spacecraft Charging

Two different mechanisms are thought to combine with vehicle design to generate spacecraft charging. Photoelectric effect and plasma bombardment are common terms for these culprits.

Eclipse amplifies the charging impact of the photoelectric effect of sunlight on a spacecraft. Bombardment of the vehicle skin by photons knocks loose electrons. As these electrons are freed from the spacecraft (photoemission), the skin develops a relative positive charge. The electrons may form a negative plasma cloud or sheath near the vehicle skin. If the entire surface of the spacecraft were a homogeneous conductor, this charge build-up would generate a current flow to spread the charge evenly over the vehicle. Since most spacecraft exteriors have solar panels, probes, lenses, etc., there is a marked difference in conductivity across the surface. The result is differential charging of the sunlit surface with respect to the unlighted portions of the vehicle. Note that depressions or holes in the vehicle may be constantly shaded. This means that even spin-stabilized satellites are subject to photoelectric charging. Obviously, spacecraft structure exacerbates the charging problem. It may also help eliminate the problem by including electron guns at certain points. A judicious squirt of electrons might be just the thing to preclude a static charge buildup. The primary effect of photoelectric charging is to counteract plasma bombardment effects.

The success of plasma bombardment in charging a spacecraft is structure dependent. A vehicle immersed in a hot (energetic) plasma is constantly bombarded by charged particles. If the plasma is isothermal, the electrons will generally have much higher velocites than the heavier protons or ions.

(Recall, the energy of a particle = $1/2mv^2$.) Electrons with energies above a few KeV are capable of penetrating one micron or more into a dielectric. Consequently, they "stick" to the spacecraft skin, causing a negative charge build up. Holes or cavities in the front end of a vehicle (relative to its direction of flight) may actually scoop up energetic particles and accelerate this charging process. The higher energy electrons are probably somewhat more efficient in causing spacecraft problems, but this is not yet certain.

Plasma bombardment can occur in several different ways. Vehicles at geosynchronous altitudes are susceptible to plasma injection events associated with geomagnetic disturbances and substorms. These events occur several times a day even on quiet days, and may produce a ten-fold enhancement of ion densities and a thousand-fold jump in electron density at geosynchronous orbit. Particle injection is predominately in the nighttime sector and results from an inward motion of the plasma sheet. The location of the plasma sheet and ring current is shown relative to geostationary orbit in Figure 15.1.

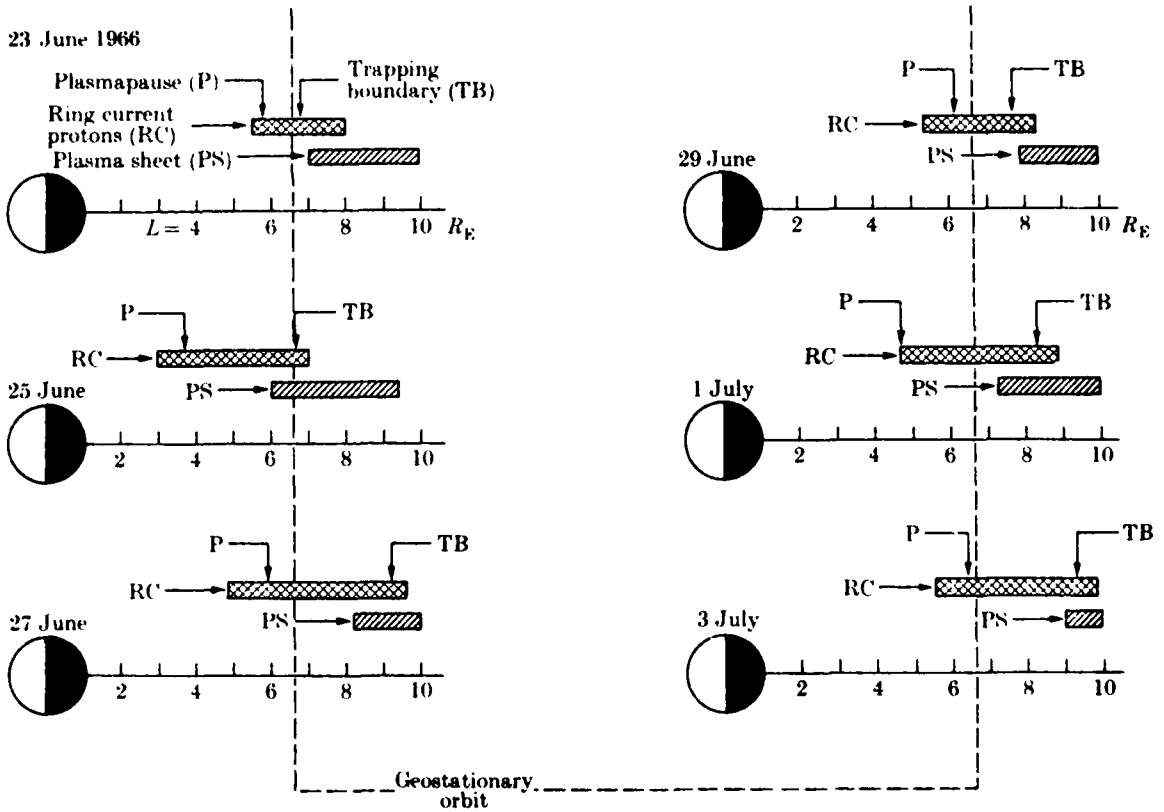



Figure 15.1 Changing Position of Magnetospheric Plasmas (Akasofu and Chapman, 1972).

Note that relative positions can change in a matter of an hour or less as well as over days. The injected electrons drift eastward into the midnight-dawn sector, and the protons move westward into the evening sector.

Such injection events have been observed in KeV electron fluxes measured aboard geostationary satellites. This data, termed EP (energetic particle) data, shows a unique, injection signature (Figure 15.2). When plasma fluxes are stable (outside injection events), EP spacecraft see trapped particle fluxes moving up and down along geomagnetic field lines. Since the vehicle's equator looks roughly up/down the field lines (its equator is perpendicular to that of the earth), this field-aligned distribution of electron pitch angles is termed "cigar-shaped." The beginning of an injection event near the local midnight meridian is marked by a gradual decline (over 20-40 minutes) in particle fluxes. This is interpreted (Kokubun and McPherron, 1981) as a conversion of the geosynchronous field lines to a tail-like structure. The trapping boundaries move earthward, and since particles are not trapped in a tail-like structure, particle flux drops. Dumping into the tail or auroral zone results. This decline is followed by an expansive phase beginning with an abrupt change to a " pancake" distribution. Particle flux increases rapidly (few minutes), and distributions return to a "cigar-like" arrangement. This phase probably represents the refilling of the trapping regions with tail plasma and the closing of field lines. Such a disturbance is probably apparent, in reduced intensity, at lower altitudes, but confirming observations are not yet available. The result is a sudden immersion of high altitude spacecraft in hot, tail plasma. The greatest flux variations are to be expected slightly above or below the geomagnetic equatorial plane. Likewise, the signature is most apparent near local midnight (spacecraft time) and is nearly invisible to vehicles operating on daytime meridians.

Plasma bombardment may also, under certain conditions, affect low altitude spacecraft. The many field aligned currents flowing inside the magnetosphere provide one mechanism. Generally, these currents are thought to neutralize each other. For spacecraft orbits which pass through the auroral ovals at abnormal angles or with high eccentricity, built-up charge many not be neutralized before an anomaly is produced. Charging may also result from anomalies in the trapped radiation belts resulting from the offset of the geomagnetic axis towards the Southeast Asian area. This offset causes greater than normal trapping over the Southeast Asian area and leakage over the South Atlantic. The earth's rotation results in abnormally high concentrations of energetic electrons and protons being dumped near the western edge of the Atlantic anomaly. Figure 15.3 shows the location of this anomaly in terms of geomagnetic field strength. Low altitude problems may, in some cases, result from induced electric fields (inside the spacecraft) as opposed to surface charge build up, but observations are again lacking.

Geosynchronous vehicles are thought to be most susceptible to charging for two reasons. First, they are close to the magnetopause. In fact, on rare occasions they are outside the magnetosphere on the day side of the earth. At this altitude, they are also capable of immersion in the night side plasma sheet, while vehicles below about 5 $R_E$ probably are not. Second, the ambient plasma density at 6.6 $R_E$ is low. This means that, unlike low orbit vehicles, the ambient atmosphere is incapable of "bleeding off" or neutralizing small charges before a discharge can occur. Low orbit charging requires fairly unique conditions (as described above) but is not, unfortunately, impossible.

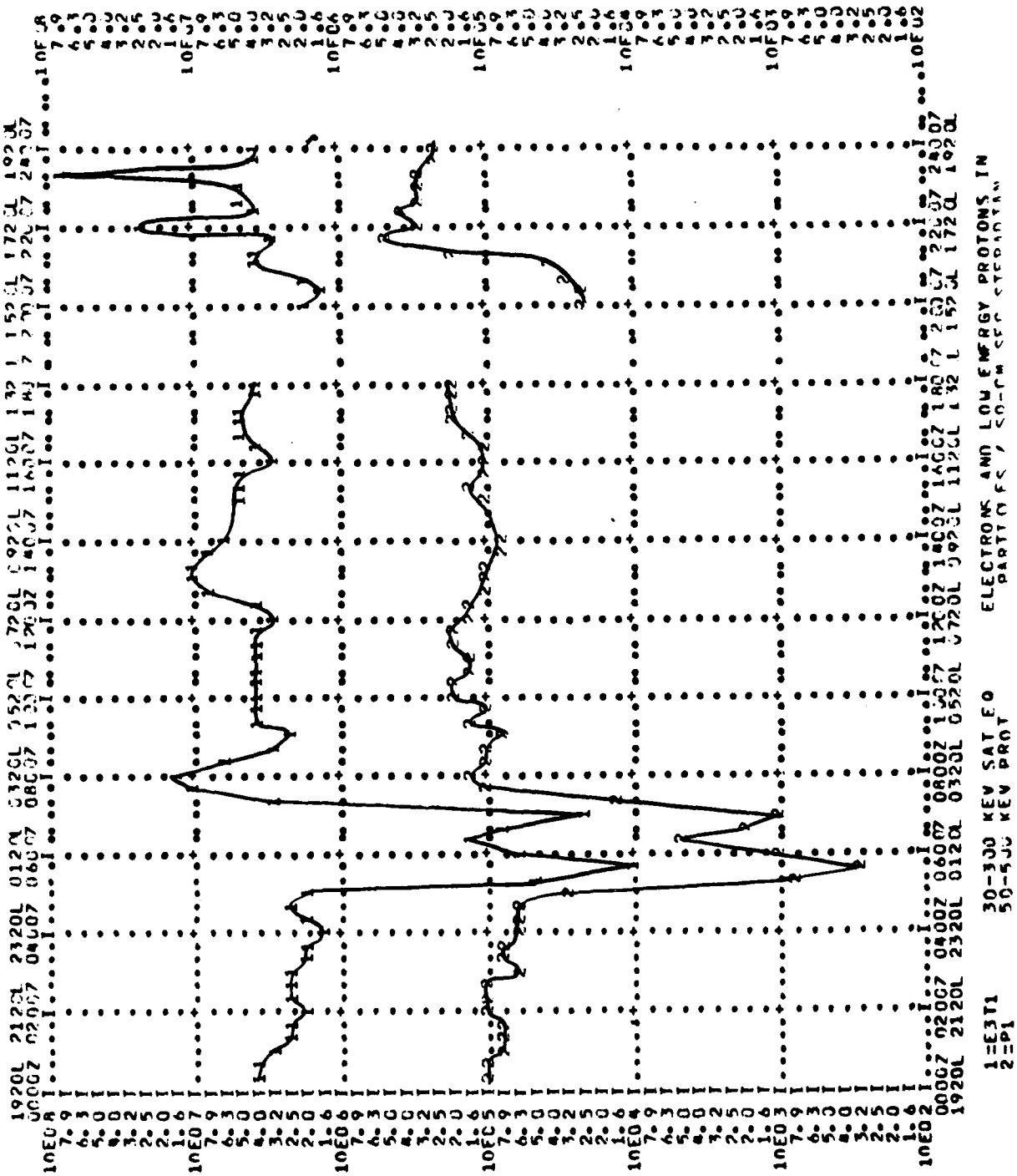



Figure 15.2 Plasma Injection Signature from EP Data.

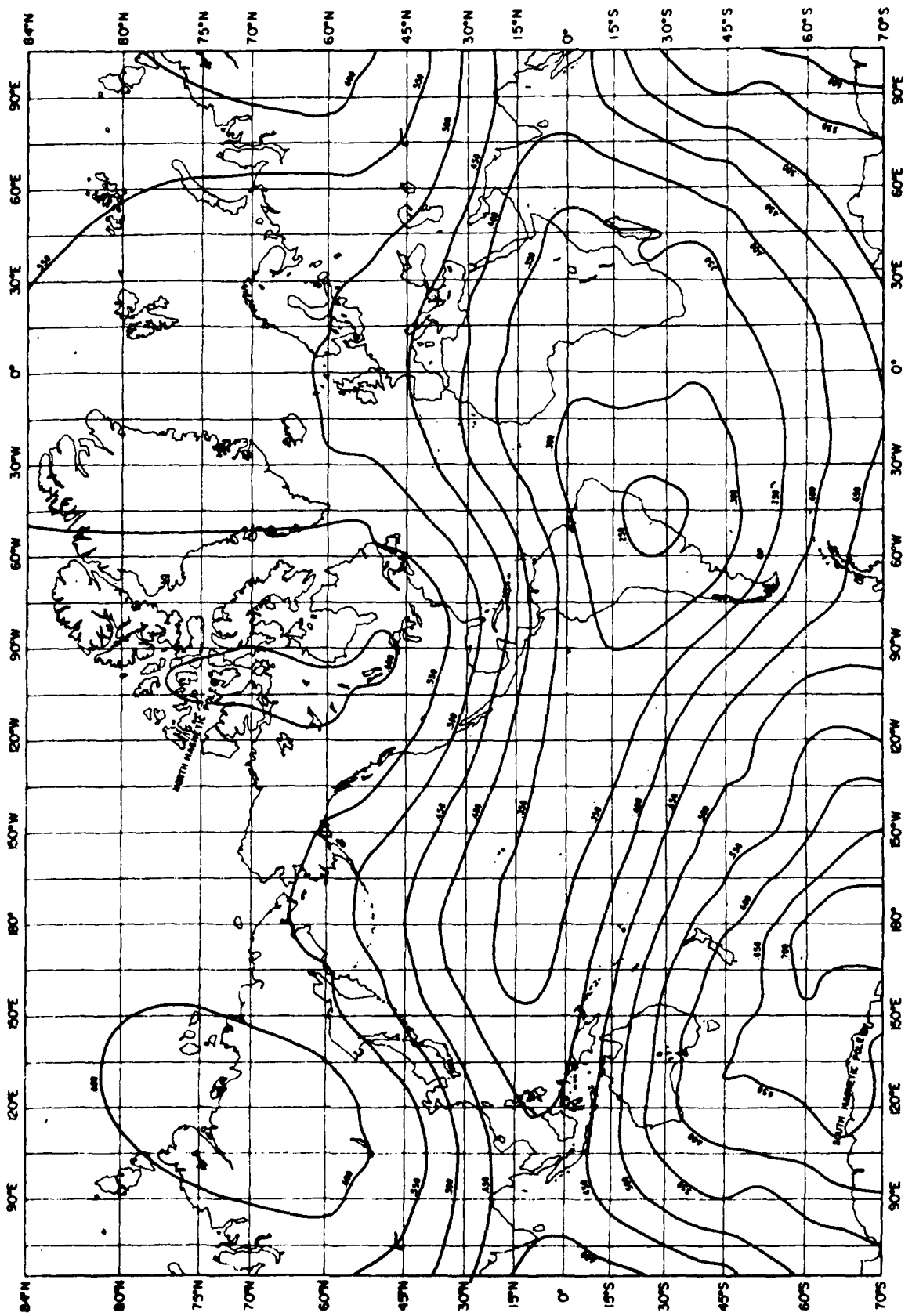



Figure 15.3 Anomaly in Geomagnetic Field Strength (after Valley 1965).

It would seem that the photoelectric effect and plasma bombardment might at least partially offset each other. To some extent, they seem to do so. The difficulty is again in vehicle design. What happens on the surface of a spacecraft is a sort of balancing act. All currents (positive and negative) to and from the surface must balance. In order to obtain this balance, the surface potential (voltage) must vary. Some parts of the craft will, therefore, generate higher potentials than others. The basic idea is revealed in a current density equation:

$$J_e - [J_i + J_{se} + J_{si} + J_{bse} + J_{ph}] = 0,$$

where $J_e$ = ambient electron current;
$J_i$ = ambient ion current;
$J_{se}$ = secondary electron current due to electron bombardment ($J_e$);
$J_{si}$ = secondary electon current due to ion bombardment ($J_i$);
$J_{bse}$ = backscattered electron current (from $J_e$); and
$J_{ph}$ = photoelectron current.

Of course, all of these currents are functions of spacecraft potential (voltage).

### 15.1.3 Spacecraft Discharging

Spacecraft charging sets the stage for subsequent discharges. Since discharges are capable of producing greater lasting damage, it is important to identify those conditions conducive to discharging. Any time charging is occurring, conditions are favorable for discharge. Experience indicates sudden changes in the electrical environment of the spacecraft may trigger static discharges. The simplest triggers include orbital maneuvers, the onset of downlink telemetry, or other electronic activity onboard the spacecraft. Eclipse or movement out of eclipse of a geosynchronous vehicle (happens near equinox) or the movement into/out of sunlight for a lower orbit vehicle may also trigger a discharge. Encountering an intense current or the boundary of the magnetosphere may also trigger spacecraft discharges.

The asymmetry of the plasmasphere during quiet and disturbed periods provides an additional, highly variable trigger. As Figure 15.4 shows, the plasmasphere occasionally reaches geosynchronous orbit and beyond. Moreover, the local time of intersection of the plasmasphere and a geosynchronous satellite will obviously vary. The odd shape of the plasmasphere results from several factors. There is a lag between the onset of solar heating (at plasmasphere sunrise) and the vertical expansion of the plasmasphere. This combines with frozen field theory (remember, to expand, the plasma must carry the magnetic field lines outward with it, and the field's inertia will slow this expansion) to delay the expansion into the afternoon sector. Then, too, the IMF compresses the magnetosphere a bit asymmetrically near the 0900-1000L meridian. (Don't confuse plasma flow angle with IMF angles. The IMF angle varies significantly due to the position of the current sheet, etc. Finally, the angular momentum imparted to the expanding plasmasphere (remember, it corotates with the earth) will tend to "sling" it into the afternoon sector.

Initial studies have suggested some statistics for analysing (or forecasting) discharge times. During low magnetic activity, discharges seem more common between 0400-0600L (spacecraft time). This may be due to quiet time injection events and the preferred direction of drift for injected electrons (eastward). Noon local is a favored discharge time during high magnetic activity, perhaps due to the increased likelihood of encountering the magnetospheric boundary (or associated current systems). Early evening (1900L) seems to yield the minimum probability of discharges. Westerly proton drift from injections may help neutralize the vehicle, and it will (normally) be comfortably away from the plasmasphere and magnetosphere by this time. The equinox brings eclipses for geosynchronous vehicles, and, simultaneous increases in discharge probability.

It must be emphasized that much of the work done to date (including the above statistics) has been with geosynchronous satellites. SCATHA is the first vehicle specifically equipped to study the effects of design, active modification (electron guns), and internal electronic activity on discharges. A complete analysis of the results of this mission will, hopefully, clarify what is so far a rather murky area of spacecraft operations.

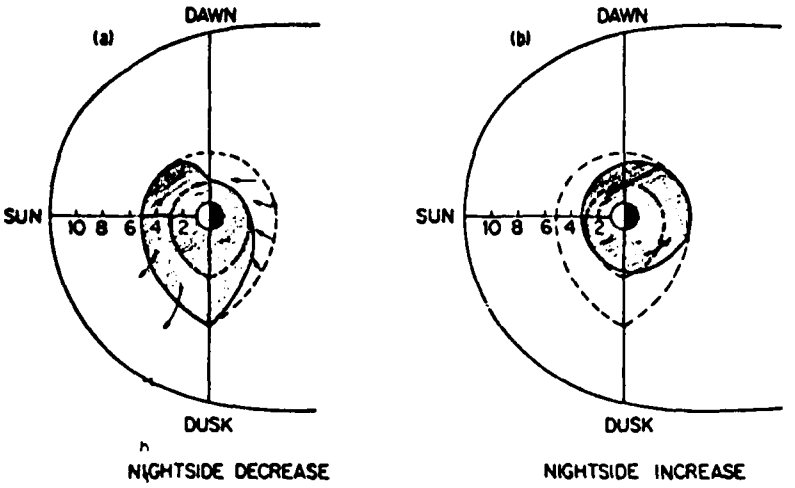
### 15.1.4 Forecasting

While we can, with some success, forecast the environmental "conduciveness" for spacecraft charging, we are not yet at that point for discharges. This is due, in large part, to the differing causes of the two phenomena. Charging is heavily dependent on the environment for all vehicles. Discharging is perhaps more dependent on vehicle design and depends on the environment in unknown percentages and ways. Nonetheless, a rapid post analysis of a spacecraft anomaly may be very useful to an operator. If a problem can be traced to spacecraft charging, it may preclude the expense (in terms of money and data lost) of vehicle shutdown and subsequent testing. A better understanding of vehicle design requirements and the environment in which the vehicle must operate may help engineers design more discharge-resistant vehicles.

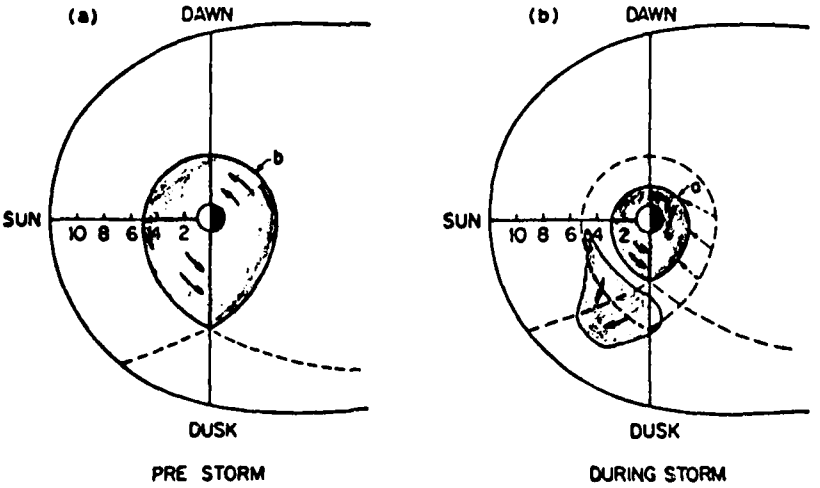
## 15.2 Spacecraft Drag

Spacecraft operating below a few thousand kilometers encounter a significant number of atmospheric particles during each orbit of the earth. Theoretical atmospheric models permit us to calculate the expected drag and its effects on vehicle orbit. Changes in atmospheric density at vehicle altitude can rapidly and significantly alter the vehicle's orbit. If these are unexpected changes, the vehicle may be temporarily "lost" by ground tracking stations using narrow-beam tracking radars. This happened several times during the Space Shuttle flight in Spring 1981. In order to eliminate the unexpected nature of such changes, we must identify their origin/cause and attempt to forecast it.

Any mechanism capable of heating the earth's atmosphere will produce density changes at altitudes above the level heated. Such heating may result from stratospheric warming, meteor showers, geomagnetic storms, and changes in solar EUV (extreme ultraviolet) emission.




A sketch showing the changes in plasmaspheric size and shape during changing magnetic activity. (*a*) shows a decrease in the nightside plasmapause radius with increasing activity; (*b*) shows an increase in nightside radius with decreasing activity




A sketch of the effect of a magnetic storm on the size and shape of the plasmasphere. During the quiet period before the storm, the plasmasphere is full of plasma. Following the onset of enhanced convection during the storm, the nightside of the plasmasphere is convected to lower $L$ shells; the outer edge of the dayside and bulge regions is peeled off and convected sunward

**Figure 15.4 Variation in Position of Plasmasphere with Time and Magnetic Activity (after Chappell, et. al., 1971).**

### 15.2.1 Heating Mechanisms

Stratospheric warming probably has the most limited impact of all heating mechanisms. It normally occurs over middle and high latitudes in the winter hemisphere and persists for a few days. Impact is probable only for orbits which pass through the affected area. The heating occurs low in the atmosphere, so effects are probably restricted to orbits below perhaps 200 km.

Meteor showers are also limited in duration to a few days and deposit most of their energy near 100 km in the midnight to local noon hemisphere. Their effects do, however, include most of the earth. Meteor showers are very regular, and the more intense showers are cataloged in astronomical tables for easy reference.

The next largest impact heating mechanism is the geomagnetic disturbance. It is several orders of magnitude more significant than stratospheric warming or meteor showers, but is somewhat limited in areal extent. During a geomagnetic storm, large numbers of particles are dumped into the high latitude atmosphere. Even substorms can result in significant, though highly localized, dumping. These particles generate considerable heating by collisions near 100 km and alter densities to 1000 km altitude or higher. Heating effects (i.e. drag) are first observed a few hours after the disturbance begins and may persist for 12-24 hours following a large disturbance. The strong field-aligned currents, the enhanced electrojets, and the ring current also contribute to atmospheric heating. Most of this heating will be near the auroral zone, though some may also occur near the South Atlantic anomaly. Consequently, polar orbit spacecraft will experience the greatest effects from geomagnetic storm heating. The effects of geomagnetic activity on density can be implied from changes in upper atmosphere temperature. Figure 15.5 demonstrates the variation in these effects with season, altitude, and latitude.

The greatest long-term impact on atmospheric density probably results from solar EUV fluctuations. Since EUV measurements are not available in real-time, daily measurements of 10.7 cm radio emission are taken to be representative of solar EUV emission. Generally, this is a good assumption even though the two types of radiation originate at different levels in the solar atmosphere. Solar EUV heating affects the atmosphere above about 100 km. Solar emissions vary daily, with a 27 period, and over the solar cycle. EUV and 10.7cm daily and monthly variations seem to be related to changes in plage area on the visible disk. Long term changes seem to originate from interior variations and are not so readily discernable optically. Generally, day-to-day density variations are masked by long-term trends. Occasionally, they are reinforced by geomagnetic activity, and the time delay between increased flux and density variations is decreased. Figure 15.6 provides a comparison of solar emission and density variations at 350 km. Note the exospheric temperature variations as well with respect to Figure 15.5. Unfortunately, geomagnetic activity effects are lost in the averaging inherent in Figure 15.6.

15.2.2 Spacecraft Impact

All variations in density result in variations in atmospheric drag on low altitude (below 1000 km) satellites. Any orbital body in an atomsphere experiences a drag which causes changes in its orbit. The equation which governs this drag is: $A_d = \frac{C_d A V^2 D}{2M}$,

where $A_d$ is the deceleration of the spacecraft;
$C_d$ is a drag coefficient (assume it is constant);
A is the cross-sectional area of the spacecraft;
M is the spacecraft mass;
V is the velocity of the spacecraft; and
D is atmospheric density.

If we assume that the spacecraft is spherical (so $A_d$ is a constant) and is not changing weight, e.g. using fuel or burning up (so M is constant), then

Drag = constant x velocity$^2$x atmospheric density.

By Kepler's Areal Law, we know that velocity is greatest at the lowest point of a spacecraft orbit. Density decreases rapidly with altitude, so drag, which depends on velocity and density, is greatest at the low portion of the orbit. A spacecraft in earth orbit will encounter much higher drag at perigee than at apogee. In fact, for a highly elliptical orbit, drag at apogee is less than at perigee by a factor of a thousand, a million, or more. Consequently, we normally ignore drag at apogee and concentrate on its effect near perigee.

We assume that the velocity at perigee is directly related to the height of apogee, and the height of perigee is directly related to the velocity at apogee. To a first approximation, a decrease in perigee height will produce a decrease in apogee velocity (and vice versa), and a decrease in perigee velocity will produce a decrease in apogee height (and vice versa).

Drag slows a spacecraft. Since drag is greatest at perigee, the decrease in speed is most significant there. In fact, to first order approximation, the decrease of velocity at apogee may be ignored unless the apogee is nearly the same altitude as the perigee. The decrease in perigee velocity results in a decrease in apogee height; while the lack of significant drag at apogee results in no significant change in perigee height. This is shown in Figure 15.7. The orbit will slowly become less elliptical and more circular. When the perigee and apogee become sufficiently close in height, drag becomes significant at all points along the orbit, and the orbit quickly decays. The spacecraft impacts the earth or burns up in the atmosphere. Note the perigee is not fixed, but actually decreases slightly as the orbit decays.

Kepler's third law relates the average height (actually, the semi-major axis) of the orbit to the period of the orbit. A decrease in mean height gives a decrease in period. This decrease in period means the spacecraft is observed to complete each successive orbit in less time (i.e. it accelerates).

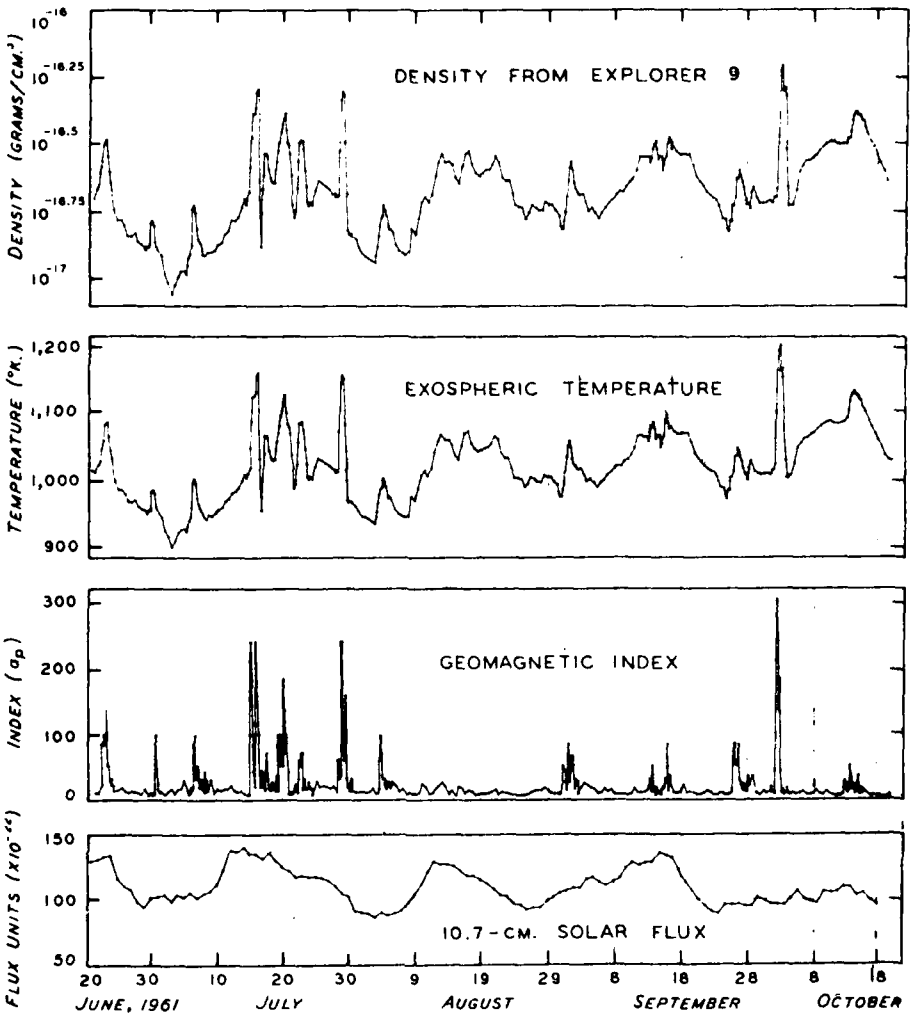



Figure 15.5 Impact of Geomagnetic Activity on Exospheric Temperature (from Jacchia, 1975).

If a spacecraft (neglecting drag) should be overhead at noon, by adding drag it may be overhead at 1130 (but at a lower altitude). A further increase in density, such as due to a geomagnetic storm, may cause it to be overhead at 1115 (all times are greatly exaggerated for this example). A spacetrack radar like the one at Eglin may detect an object where none "should" be. An analyst at the NORAD Space Defense Center, assuming no density changes, may decide there is a new spacecraft of unknown origin in orbit, or that a known vehicle has changed orbit. Density need not steadily increase to cause problems. A single increase will alter not only the next orbit, but also subsequent orbits, because the orbit-to-orbit variation will also be different than before the density change. Likewise, an unforecast decrease in density can cause probelms. Thus, NORAD and Sunnyvale use the SESS geomagnetic index

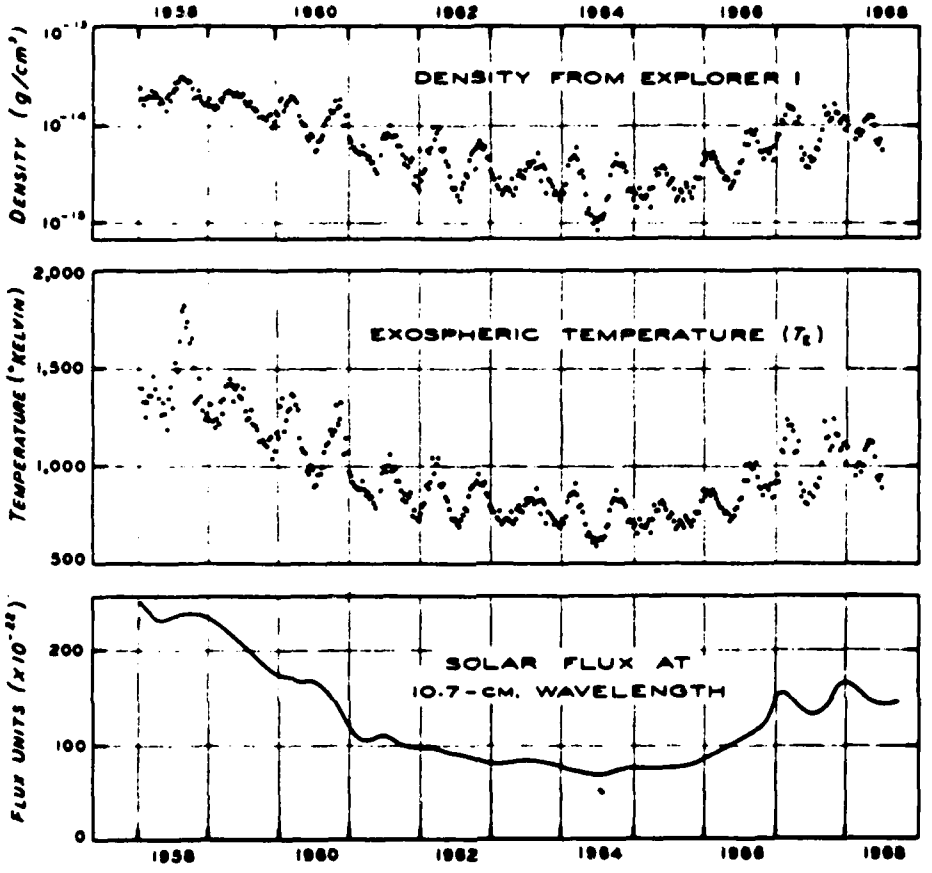



Figure 15.6 Impact of Changing Solar Emissions on Atmospheric Density at 350 km (from Jacchia, 1975).

and 10.7 cm radioflux observations and forecasts to change their atmospheric density model. This, in turn, is used to correct the expected locations of satellites for some future time and to plan orbital maneuvers.

Specific comparisons for $\varepsilon$ satellite at 185 km can be made in terms of resulting in-track displacements (how far ahead or behind it is compared to its calculated position). Over 12 hours, a $K_p$=3 results in a 5 nm displacement, while a $K_p$=8 produces a 50 nm displacement. Note that effects of geomagnetic activity will be most severe on polar orbit vehicles. Moreover, density variations resulting from geomagnetic activity may persist for 8-24 hours after the end of the disturbance as measured on ground-based magnetometers. The effects are likewise more severe for higher resolution/less areal coverage tracking systems. Since the level of

geomagnetic activity varies rapidly, atmospheric density models must be continually updated during disturbed periods. SESS provides magnetometer analysis data every 90 minutes, since many low altitude vehicles have 90 minute periods.

Over 3 days, a 10.7 cm flux level of 70 SFU (solar minimum) results in a 600 nm in-track displacement at 185 km (compared to no density/drag effects). A 220 SFU (solar maximum) flux level produces an 1800 nm displacement in 3 days for a vehicle at 185 km. This, of course, was emphasized by the early demise of Skylab.

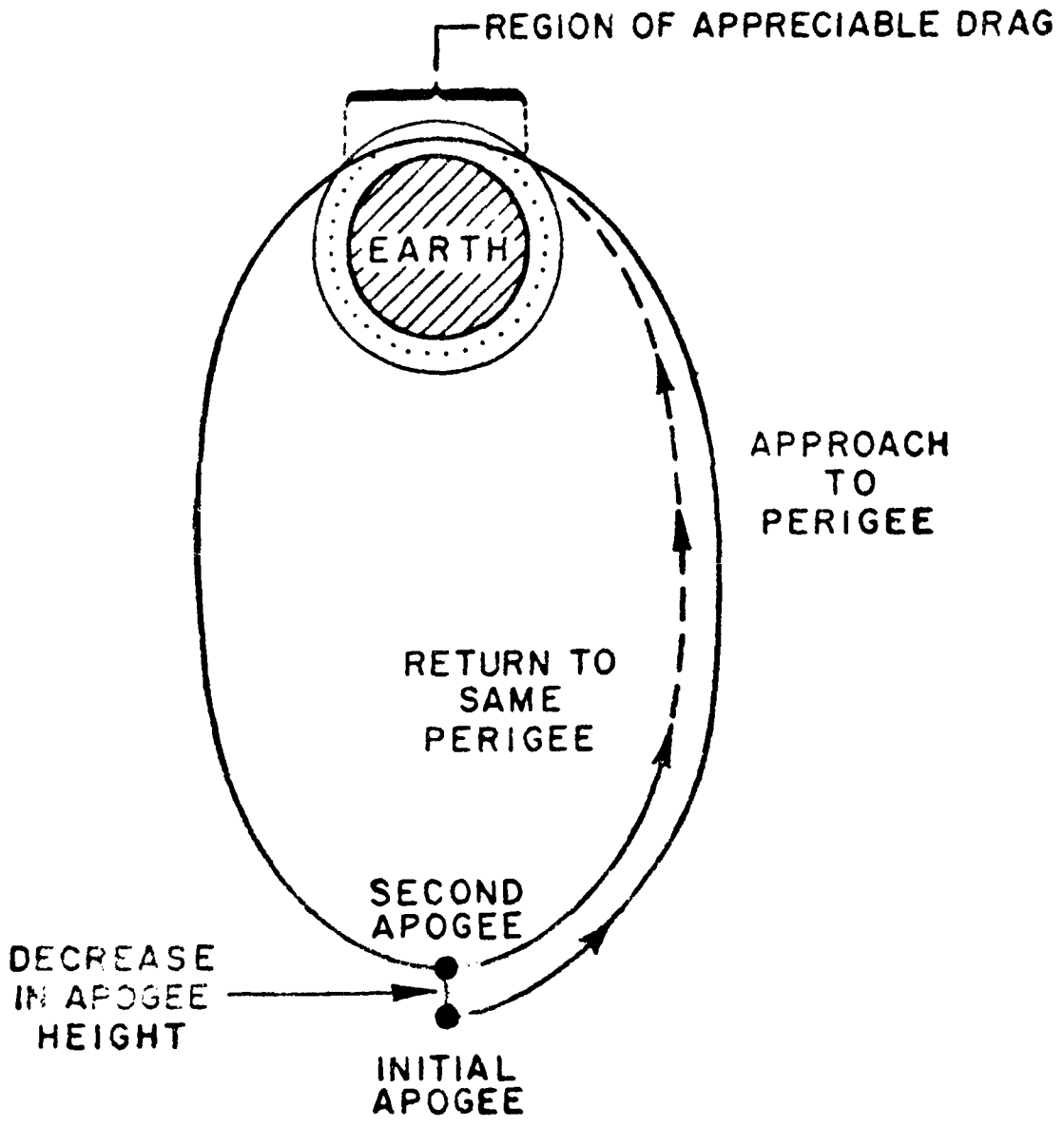



Figure 15.7 Orbital Effects of Satellite Drag.

## 15.3 Radiation Effects on Manned Systems

Although the study of space radiation began early in this century, most of our knowledge was acquired during the last two decades. The advent of manned space flight provided an immense number of observations and analytical treatment of space radiation hazards. While research efforts have provided a basic understanding of these hazards, a great deal remains unknown.

### 15.3.1 Space Radiation

We are constantly bombarded by many types of ionizing radiation. Fortunately, our atmosphere blocks much of the harmful component. In space, this protective shield is not available. The ultraviolet, visible, and infrared radiation found in space could pose a problem for unprotected crew members. Another, more lethal, type of ionizing radiation results from the activity of subatomic particles of various masses and charge as they interact with materials.

The main hazard to life in space is found in the ionizing radiation resulting from exposure to high energy particles. If a particle possesses sufficient kinetic energy it can pass through protective equipment and impact a crew member's body. Particles of the highest energy may pass through an individual with no serious effects. Particles which are stopped by human tissue pose the most danger. As these particles rapidly deccelerate, their energy is converted into a pulse of electromagnetic (EM) radiation. This radiation can ionize atoms within the crew member's body. In addition, the impact of the energetic particle can excite neighboring atoms. As these atoms fall back to lower energy states, they too can produce ionizing radiation.

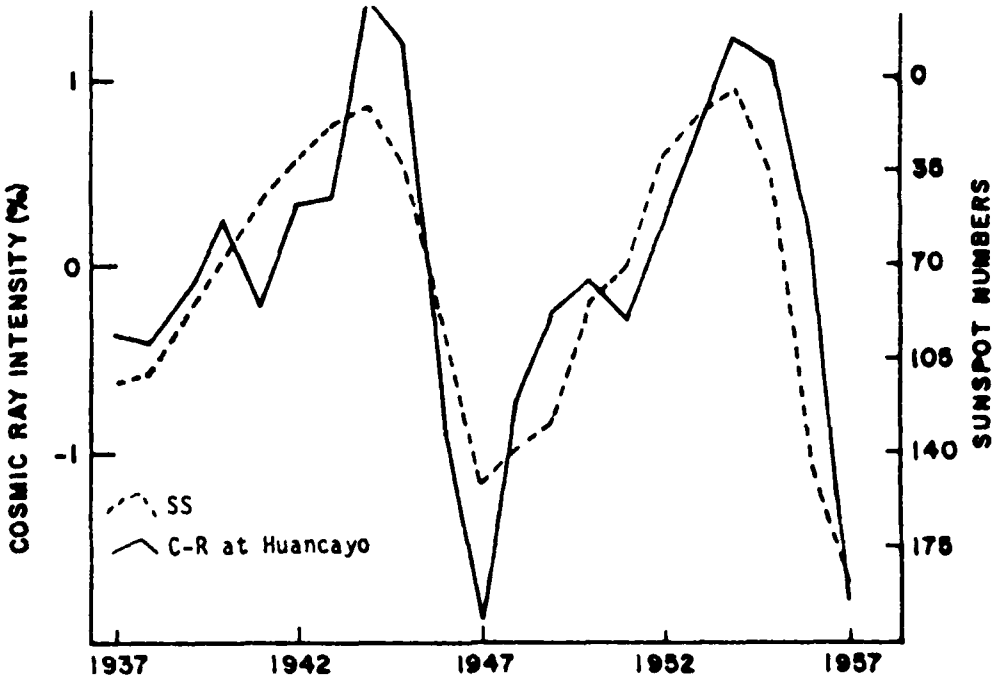



Figure 15.8 Cyclic Variation in Galactic Cosmic Radiation (Forbush and Venkatesen, 1960).

### 15.3.2 Space Radiation Environment

Natural radiation in near-earth space (up to geosynchronous orbit) has three primary components: galactic cosmic radiation (GCR), radiation produced by trapped particles (Van Allen Belts), and radiation from solar flare particles. All three components are influenced by solar activity and the earth's magnetic field. Their relative contributions to radiation hazards are most easily understood when considered separately.

#### 15.3.2.1 Galactic Cosmic Rays

GCR are high energy (greater than .1 BeV) protons, electrons, or other heavy, energetic particles. Emitted by distant stars and even more distant galaxies, they diffuse through space and arrive at earth from all directions. Spatial variations in GCR flux, and therefore GCR related radiation, are produced by variations in source location, the earth's magnetic field, and by atmospheric shielding. Flux increases with increasing altitude. Particle flux is also larger over the polar regions where "open" geomagnetic field lines allow easier access. The most important temporal variation in flux is associated with the 11 year solar cycle (Figure 15.8). During solar maximum, when the interplanetary magnetic field strength is greatest, cosmic ray particles are scattered away from the earth. This produces a GCR flux minimum. Conversely, GCR flux is largest during solar minimum. The 11 year variation produces a 2X variation in the cosmic ray dose at geosynchronous orbit. Low altitude, low inclination orbits would experience almost no dose variations due to the strong shielding produced by the combined effects of the atmosphere and geomagnetic field.

Due to their extremely high energy, GCR are very penetrating, and spacecraft shielding is not very effective in reducing the radiation dose. Fortunately, GCR flux is comparatively low, so it doesn't pose a serious threat to humans (several particles have probably passed through your body since you started reading this section). In all orbits, approximately 5-10% of the total effective radiation dose is due to GCR. This small amount is sometimes referred to as background radiation.

#### 15.3.2.2 Trapped Radiation Belt Structure

The Van Allen radiation belt is a doughnut shaped region which surrounds the earth. It consists of geomagnetically trapped energetic (KeV to MeV) particles. These energetic particles are capable of producing ionizing radiation when they impact shielding or body tissue. The most hazardous regions within the belts are marked by maximum densities of the most energetic particles.

The belt has an inner region rich in energetic protons and an outer region populated primarily by energetic electrons. Proton flux is most intense at about 2,200 miles above the earth's surface, while electron flux in the outer region peaks near 9,900 miles altitude. The low density area separating these two regions is often called the "slot" and represents the least hazardous region for manned spacecraft operations.

The inner Van Allen belt first appears at an altitude of about 250 to 750 miles, depending on latitude. It extends outward to about 6,200 miles, where it begins to overlap the outer belt. Energetic protons trapped in the inner belt are the major source of radiation for low, earth-orbiting spacecraft. The amount of radiation varies with latitude and longitude (the inner belt extends to about 45° latitude). Maximum dosage can be expected over the South Atlantic anomaly, where the geomagnetic field is weakest. This allows large fluxes of precipitating particles.

The inner belt proton population is also susceptible to solar-induced variations. Population density varies out of phase with the 11 year solar cycle, so that the inner belt is most inflated during solar minimum. This variation in particle population produces a factor of two variation in radiation dose rate during the solar cycle for low orbiting spacecraft. Dose uncertainly is of the same order due to the steep gradient in particle flux with altitude.

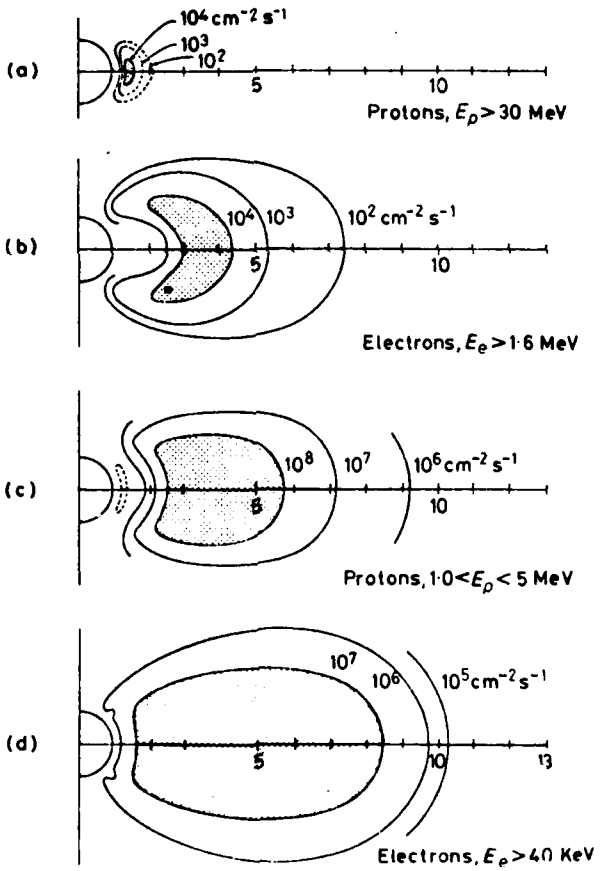



**Figure 15.9 Trapped Radiation Belt Structure showing trapped proton and electron distribution by energy (Hess, 1968).**

The outer Van Allen belt begins near 6,200 miles and extends to 37,000 - 52,000 miles depending on solar activity. The outer or electron belt contains both electrons and protons. However, the electrons are much more energetic (KeV-MeV) and are responsible for most of the radiation dose within this region. The outer belt is asymmetric, with the night side being elongated and the dayside flattened.

Variations on several time scales in electron energy, location, or both result in variations in radiation dose at a given location. Generally, particle energy and outer boundary location vary with the 11 year cycle. During solar maximum, the outer boundary of the electron belt is closer to the earth and contains higher energy particles. At solar minimum, the outer boundary is near 52,000 miles and contains less energetic electrons. Outer belt electron densities undergo order of magnitude changes over time scales of weeks. These short term variations can produce significant radiation dose variations and are related to the level of geophysical activity. During very active periods, the outer belt is inflated with high energy electrons which increase the radiation hazard substantially. Diurnal variations in radiation dose inside a spacecraft in high altitude circular orbits can occur when the trajectory crosses the asymmetric outer electron belt.

The trapped radiation belts present a serious threat to the space traveler. The locations of the most hazardous regions are sufficiently well known so that flight trajectories may be planned to limit time spent in these areas. For missions confined to these dangerous regions, close monitoring of the radiation dose and acceptable dose accumulation must be specified.

15.3.2.3 Solar Flare Particles

Solar protons, also referred to as solar cosmic rays (SCR), represent the third and most variable component of natural space radiation. Solar cosmic rays are composed of protons, electrons, or other heavy nuclei accelerated to energies between $10^7$ and $10^9$ eV during large solar flares. These particles can be responsible for a thousandfold increase in the radiation dose over short periods of time. Similar to the other energetic particles, SCR produce ionizing radiation when they interact with atoms (shielding or body tissue).

Solar particle events show a correlation with the 11 year solar cycle. The largest events normally occur in the months following sunspot maximum. Individual events vary considerably in particle constituents, energy spectra, and particle flux. Usually, a few very large flares dominate the total particle emission for the entire solar cycle. Due to the "individuality" among events, the radiation dose accumulated due to solar protons may vary from negligible to well above lethal.

The altitude to which a solar proton may penetrate is related to the particle's energy and injection latitude. This results in significant variations in radiation dose with altitude and latitude of the spacecraft. For a fixed altitude, spacecraft can experience different levels of radiation depending on orbit trajectory. Equatorial orbiting spacecraft will experience lower proton fluence (and therefore lower radiation dose) than a polar orbiting satellite at similar altitudes. In general, solar proton radiation is a significant hazard for orbits passing above $50^o$ latitude at altitudes above a few earth radii (1 earth radius = 6378 km = 3960 miles). Solar cosmic

rays emitted during a large solar flare present the greatest uncertainty and the greatest threat to manned spacecraft in regions beyond the protection of the earth's atmosphere.

### 15.3.4 Radiation Hazards

The hazards posed by trapped particles and high energy flare particles are difficult to assess in a quantitative manner. The deleterious effects of radiation depend on total dose, dose rate, particle identity, and energy, or a combination of all these variables.

#### 15.3.4.1 Dosimetry

When high energy particles encounter atoms or molecules within the human body, an atomic interaction (ionization) may occur. A direct interaction occurs when the particle is suddenly stopped by collisions resulting in a release of energy which may remove electrons from nearby atoms or molecules. Ions result. Indirect encounters occur when the high energy particle, usually an electron, is deflected by another charged particle. The deflection causes a release of energy (radiation) which also may produce ionization. The close encounter process is commonly referred to as bremsstralung. In either interaction, the effects of the ionizing radiation are proportional to the amount of energy absorbed by the surrounding material. To quantify this absorbed radiation, a unit of measurement called a RAD was defined. A RAD is the amount of ionizing radiation corresponding to 0.01 joule absorbed by one kilogram of material (equal to 100 ergs per gram). Note that a dose of 10 RADs from high energy protons is the same as 10 RADs from x-rays. The RAD represents an amount of absorbed radiation energy and not what produced it. Another unit, the GREY, is defined as 100 RADs.

Radiation physics indicates that 1 RAD received from x-rays produces far less bodily damage than 1 RAD received from high energy protons even though both deposit equal amounts of energy. Another unit was developed to express the effects of radiation on biological tissue. To define this unit, each type of ionizing radiation was given a relative biological effectiveness (RBE) compared to a beam of 200 KeV x-rays. Table 15.1 lists the RBE for several types of radiation. From the table, we see that protons are twice as damaging as 200KeV x-rays. A one RAD proton dose will be twice as damaging as a one RAD dose from x-rays. A REM relates biological damage to type of radiation. The biological equivalent dose (REM) = dose (RAD) x RBE. For example: a one RAD dose of 200 KeV x-rays gives a biological equivalent dose of 1 REM, but a one RAD dose from protons gives a biological equivalent dose of 2 REM. The larger REM value for protons accounts for the increased biological damage. Accumulated data suggests that electrons, protons, neutrons, and alpha particles are the most damaging (largest RBE value) due to their ability to penetrate deeply into human tissue and release or produce a large number of ions. Another unit commonly used in discussing REM dosage is the SIEVERT. A SIEVERT equals 100 REM.

| Radiation | RBE |
|---|---|
| 5-Mev gamma rays | 0.5 |
| 1-Mev gamma rays | 0.7 |
| 200-Kev gamma rays | 1.0 |
| Electrons | 1.0 |
| Protons | 2.0 |
| Neutrons | 2-10 |
| Alpha particles | 10-20 |

Table 15.1 Ralative Biological Effectiveness of Various Radiation Sources (Bueche, 1981).

On the average, we experience about 40 millirems (MR) each year from radio activity in soil, rock, and wood around us. This figure varies from place to place. On the East Coast, it is around 20 MR; while near the Rocky Mountains, the value is closer to 90 MR. Cosmic rays passing through your body provide an addition 40 MR annual dose (160 MR if you live high up in the Rocky Mountains). Inescapable sources in food and water provide an additional 20-50MR which brings the total yearly dosage for an earth-bound person to about 170 MR (add 4 MR for each New York to Paris airline flight).

In contrast to this, the space traveler will receive considerably more radiation. The dose received varies with mission duration, orbital profile, and shielding. Table 15.2 highlights some of the radiation doses experienced by Gemini astronauts during the mid-1975 period.

| Mission | Launch Data | Apogee (NM) | Perigee (NM) | Inclination (Revs) | (Deg) | Avg Dose (Milliard) |
|---|---|---|---|---|---|---|
| V | Aug 21, 1965 | 189 | 87 | 120 | 32.5 | 176 |
| VIII | Mar 16, 1966 | 161 | 86 | 7 | 29 | 10 |
| IX | Jun 3, 1966 | 168 | 86 | 45 | 29 | 19 |
| X | Jul 18, 1966 | 412 | 161 | 8 | 29 | 726 |

Table 15.2 Gemini Orbital Parameters and Average Radiation Doses in Millirads (after Atwell, 1980).

Shielding stops or alters the trajectory of high energy particles before they encounter the more sensitive human tissue. In general, the denser a material the more effective it is as shielding. Unfortunately, with booster limitations on weight launched, denser, heavier shielding means reducing payload elsewhere. Aluminum is used extensively, since it combines both high density (it's a metal) and lightness. A typical space suit contains only 0.2 gram/$cm^2$ of aluminum but effectively stops up to 10 MeV protons. Higher energy particles pass through the suit and can deposit most of their energy in the individual. Spacecraft typically have several grams per $cm^2$ of aluminum shielding and can stop even higher energy particles. Table 15.3 reflects the doses expected for various orbits and shielding thickness.

| | 0.1gm/cm$^2$ Shielding | 5.0 gm/cm$^2$ Shielding | | |
|---|---|---|---|---|
| ORBIT | TRAPPED + COSMIC RAY | TRAPPED + COSMIC RAY | SOLAR FLARES | TOTAL |
| 400 km/30$^0$ | 400 | 32 | - | 32 |
| 400 km/90$^o$ | 15000 | 28 | 80 | 108 |
| GEOSYNCH | 3 X 10$^6$ | 300 | 250 | 550 |

Table 15.3 Predicted Radiation Dose (REM) for 1 Year Exposure Assuming an RBE of 1 (after Watts, et. al., 1976).

Note that low orbits, especially those confined to the equatorial plane, are substantially less hazardous than geostationary orbits. The former make use of the earth's natural shielding, while the latter expose an individual to ambient energetic particles in a region where natural shielding is of limited value. Shielding is particularly important at geosynchronous orbits. In this region, radiation penetrating less than 3 grams/cm$^2$ of aluminum is due to primary energetic electrons. Behind at least 3 grams/cm$^2$ of aluminum, radiation is due to Bremsstralung. Workers protected with only a space suit during extra-vehicular activity (EVA) could receive about 0.43 REM per day in this region. This is sufficient to damage the eyes and other vital organs.

Models employed in predicting doses are usually based on previously measured particle fluxes for a given environment. Unfortunately, methods used to convert particle flux into a biologically equivalent dose are often limited by assumptions about particle types and energy range, flux density, and shielding effectiveness. One model assumes the most damaging particles from a solar flare are protons having an average energy near 50 MeV. Assuming an unshielded 68.1 kilogram space worker is exposed to these particles one can estimate a dose rate by using the greater than 10 MeV integrated flux. The expression is (4.4 X 10$^{-4}$) (greater than 10 MeV flux) = dose rate (REM/ hour). This method makes several restrictive assumptions to yield a very rough estimate. The limited accuracy of these models makes them useful for mission planning but not practical for monitoring daily space operations. For safety and greater accuracy, daily operations rely on in-situ measurements of accumulated dose and dose rates.

15.3.4.2 Biological Effects

Ionizing radiation produced by energetic particle bombardment of the human body causes changes at an atomic level. The extent and location of change will determine the degree of damage. The radiant energy released can alter the electrochemical make-up of a cell by a restricting movement of protein in and out of the cell or by producing toxic substances. These alterations can be lethal to the cell, particularly if enzymes (which are catalysts for all biological reactions) are destroyed. Organs and tissue are made up of like cells, so extensive cellular damage or destruction can lead to tissue or organ failure. The result is radiation sickness or death.

The three most sensitive organ systems are the blood, digestive, and central nervous systems. The most sensitive tissues include the gonads, the skin, and the lens of the eye. Damage to cells by alteration of the chemical processes is considered indirect, as the resulting physical change usually occurs slowly.

A more direct physical change in the cell results when high doses of radiation are received over relatively short periods. Such acute exposure results in more immediate effects. These effects may be brought on by destruction of the cell nucleus which contains genetic material and mediates the cellular activity. Such effects are unlikely in space workers except during nuclear detonations in space or large solar particle events. Acutely exposed workers would need immediate medical attention, since major organ systems would be affected.

Radiation damage may be reversible or irreversible depending on dose, dose rate, and tissue or organs affected. Small doses which kill a few red blood cells may be reversible, since red blood cells are replaced about every 120 days. Larger doses may affect the bone marrow which produces these cells, and may not be reversible. The end result would be anemia or even death.

Irreversible damage brought on by prolonged, low dosages usually results in delayed effects. This makes minimum dose thresholds difficult to establish. Based on conventional types of ionizing radiation in the space environment, delayed effects include cancer, developmental abnormalities in newborn, genetically related ill heath, lens cataracts, shortened life span, and impairment of fertility. The average yearly dose of 170 millirem to earth bound individuals leads to about 6000 cancer deaths, or roughty 0.2% of the total yearly cancer deaths in the U.S. Insufficient data makes a similar dose comparison impossible for space travelers.

To further complicate the difficulty in assessing biological damage for a given dose, studies show that individuals have varying degrees of tolerance based on sex and stamina. To circumvent this problem, most estimates of radiation effects are based on a population sample with the number of affected individuals expressed as a percentage. Table 15.4 relates accumulated dose to probable effect on a sample population.

| DOSE (REM) | PROBABLE EFFECT |
|---|---|
| 0-50 | No obvious effect except, possibly, minor blood changes. |
| 50-100 | Radiation sickness in 5-10% of exposed personnel. No serious disability. |
| 100-150 | Radiation sickness in about 25% of exposed personnel. |
| 150-200 | Radiation sickness in about 50% of exposed personnel. No deaths anticipated |
| 200-350 | Radiation sickness in nearly all personnel. About 20% deaths. |
| 350-550 | Radiation sickness. About 50% deaths. |
| 1000 | Probably no survivors. |

Table 15.4. Probable radiation dose effects for a sample population (Cladis et. al., 1977).

Using a similar technique, models have been developed which express astronaut sickness per unit of radiation. One model estimates that for each 10,000 space travelers exposed to 10 missions of 40 REM each, between 320 and 2000 additional cancer deaths in excess of normal expectations might occur in later life. While these estimates are subject to a great deal of uncertainty, they highlight some of the severe consequences which can befall the poorly protected space traveler.

### 15.3.5 Space Medicine

The complexities of space operations require highly skilled astronauts working at near peak efficiency. Since workers experiencing even slight radiation sickness could make small errors resulting in disasterous consequences, protection of the space traveler is necessary for any mission's success. The goal of space medicine has been to meet this need by monitoring radiation dosage, assessing biological effects, and establishing maximum acceptable dose accumulation based on mission requirements and worker safety.

Combining in-situ measurements and analytical modeling, NASA officials have formulated estimates of the radiation dose for various near-earth regions. Unfortunately, state-of-the-art models are not sophisticated enough to account for dynamic magnetospheric processes or individual solar particle events. Both of these variables are capable of considerable modification of the dose estimates. Model estimates are, nevertheless, useful in establishing an average or baseline dose.

Recognizing that no region of space is free from radiation, mission planners are forced to accept some risk for each space journey. These risks change for each mission, depending on time, location, and the individual's health. Consideration of these influences led the Space Science Board and NASA officials to establish radiation dose guidelines (Table 15.5) for manned spaceflight exposure for a variety of mission durations. By combining selective mission planning with a proper degree of shielding, it was determined the man can perform effectively in space with minimal radiation sickness risk.

| Mission Operational Dose | Skin | Eye | Bonemarrow |
|---|---|---|---|
| 30 Day Max | 75 | 37 | 25 |
| Quarterly Max | 105 | 52 | 35 |
| Yearly Max | 225 | 112 | 75 |
| Career Limit | 1200 | 600 | 400 |

Table 15.5. Established dose guideline in REM (Space Science Board, 1970).

#### 15.3.5.1 Radiation Dose Monitoring

In-situ radiation measurements have been commonly used to monitor accumulated dose for space workers. On board spectrometers are used to measure high energy particle flux and spectral data. Dosimeters are used to

measure skin and depth dose rates. Hand held radiation survey meters are also used to measure REM dose rates. Once collected, the data can be transmitted to earth for analysis. Dose levels in excess of prescribed limits would probably cause mission curtailment or replanning. This method has the advantage of near real-time monitoring; however, it is an after-the-fact method having little predictive value.

These types of on-board systems may not provide adequate safety margins. During EVA, for example, several additional minutes may be required, after reaching exposure limits, to move to a more suitable shelter. Such necessary delays could result in over exposure.

15.3.5.2 Methods of Protection

Several methods are available to mission planners to minimize radiation dose exposure. The "best" method is usually a compromise between mission profile, weight limitations, and crew health considerations.

Mission timing, when possible, can be useful. Mission timing may involve scheduling the construction of a space station for the least hazardous period of the solar cycle or minor changes to an existing mission such as termination of an EVA after very energetic solar flare activity. Unfortunately; this method can be very impractical based on need and committment of funds and manpower.

Orbital choice can be a viable means of reducing radiation exposure. In general, low equatorial orbits are the least hazardous, while geosynchronous orbits (GEO), especially during solar particle events, expose the crewmember to the greatest levels of ionizing radiation. Low earth orbits (LEO) inclined to the equator expose crew members to low-energy electrons and high energy protons primarily in the South Atlantic anomaly (SAA). The electrons can be stopped by minimal shielding, but the protons will produce significant levels of radiation. Flight paths in LEO are such that some orbits will pass through the SAA while others will not. Flight trajectories through the SAA will require careful mission planning such as EVA curtailment to prevent over exposure. Table 15.6 highlights expected radiation doses for several orbits. Transfer ellipse orbits (TEO) are trajectories which take crew members from LEO out to GEO and lack. During large solar events, it may be beneficial to move from GEO to LEO to take advantage of the shielding afforded by the upper atmosphere and magnetic field.

The most common method of protection is through the use of shielding. Shielding may be magnetic, electric (used primarily for equipment), or a substance such as aluminum (Al). Aluminum seems to be best suited for shielding in space applications, because it is dense, lightweight, easily fabricated, flexible, and relatively inexpensive. Materials such as lead offer better shielding at the expense of more weight. A spacesuit provides adequate shielding for LEO. For GEO, the trapped electron population has sufficient energy to penetrate a spacesuit. Extra vehiclar activity at GEO will require additional shielding and careful mission planning.

| Mission Phase | Dose Equivalent (REM) |
|---|---|
| LEO: | |
| average daily dose at solar minimum | 0.15 - 0.30 |
| at solar maximun | 0.08 - 0.15 |
| TEO: | |
| average one-way trip from LEO TO GEO | 0.02 - 1.0 |
| GEO: | |
| average daily dose at solar minimum (worst case) | 0.43 |

Table 15.6 Expected Radiation Doses for Specific Orbits (from M. White, 1980).

Vehicle shielding is another means of limiting the radiation hazard. External skin thicknesses range from 0.8 mm Al for satellites to as high as several millimeters of Al. A large manned space structure would typically have about 5-8 grams/cm$^2$ Al of shielding surrounding the habitat or work area. For most orbits, 5 gm/cm$^2$ Al represents a safe compromise between worker health and weight limitations. GEO receives the greatest shielding for various mission profiles. In GEO, for shielding greater than 1-2 gm/cm$^2$ Al, the bremsstrahlung dose becomes the primary radiation source. Figure 15.10 shows the estimated daily dose at GEO for various thicknesses of aluminum shielding.

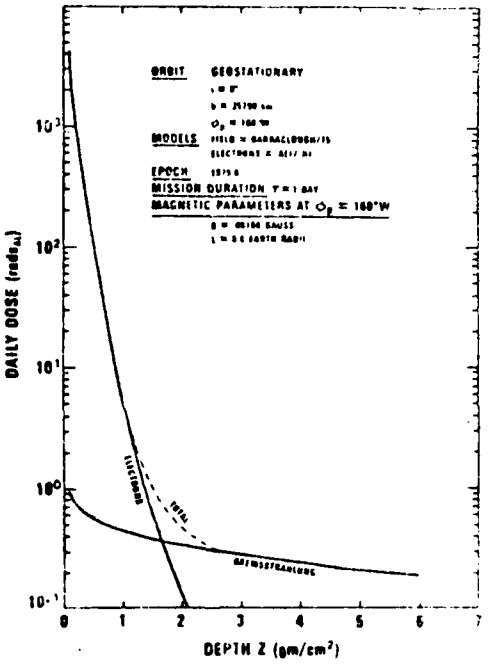



Figure 15.10 Electron and Bremsstralung Dose Estimates (Stassinopoulos, 1980).

The need for greater protection during large solar particle events has resulted in several additional short-term protection schemes. One method involves orienting the surface of greatest shielding on the space vehicle towards the highest particle flux. Another method, the storm cellar approach, would allow heavy shelding in one region of the space vehicle. Personnel would then move into this "storm cellar" during hazardous solar events. Other less costly methods involve rearrangement of existing equipment in an attempt to protect the crew.

## 15.4 Summary

Spacecraft operations are ushering in a host of new problems. Electrostatic charging, and subsequent discharges, can befuddle operators of high altitude satellites by generating false commands or damaging critical systems. Plasma injection and photoelectron emission are the primary sources of this problem. Plasma injection during a geomagnetic disturbance is also major source of upper atmosphere density variations. Changes in solar emission, meteor showers, and stratospheric warming can also alter the densities encountered by low orbit spacecraft. Resulting orbit changes can cause large tracking and positioning errors. While near-earth orbit problems can confuse or damage spacecraft and alter their orbits, they can be lethal to astronauts. If the orbit is selected to avoid the trapped radiation belts, a low orbit provides maximum protection via atmospheric shielding. The multitude of radiation sources make orbit planning particularly important. This need is reemphasized by the limitation in on-board shielding due to launch weight constraints. In all these areas, the observations are limited and their exact meanings are not yet clear. An expanded space program should help solve these and numerous other SESS-related problems.

# CHAPTER 16

## SUN AND CLIMATE

Energy from the sun drives the atmospheric motions which constitute weather. In the simplest model, uneven solar heating causes the hot tropical air to rise and travel towards the high latitudes where it cools and descends. More extensive theories usually employ a three cell model with distinct areas of rising and falling air causing long term climatic conditions. Examples are the deserts that prevail in latitudes characterized by subsiding air. The air travelling in large scale convective cells is acted on by the Coriolis force due to the rotation of the earth. This leads to the familiar flow patterns about centers of low and high pressure.

Climatic change occurs, with varying regularity and intensity. Perhaps the most striking example of climatic variation is an ice age. Several times in the past million years, great ice sheets have advanced into what is now the northern United States. The most recent ice sheet receded only twenty thousand years ago. It is not necessary to have an ice age in order to cause catastrophic upheaval in the everyday lives of people. In the last millennium, many cases of seemingly minor climatic variation have radically affected the lives of many people. Greenland (perhaps intentionally misnamed by Erik the Red in order to obtain more volunteers for his new colony) certainly cooled considerably during the early Middle Ages, becoming uninhabitable for all but the most hardy adventurers. Europe suffered a similar cooling trend in the same period.

A more immediate example, which has done much to promote interest in the direct relationship between solar and climate cycles, is the drought cycle on the high plains that occurs approximately every twenty-two years. Twenty-two years is about double the well known, but not universally accepted, eleven year solar cycle. Furthermore, the droughts tend to occur at solar minimum (Lansford, 1979). This possible connection is an enticing reason to relate solar and climatic variation.

### 16.1 The Early Search For a Solar - Terrestrial Connection

The notion of a connection between a solar occurrence and terrestrial phenomena probably began in 1859 (Wilcox, 1976). R.C. Carrington, while engaged in his daily task of mapping sunspots, saw a very brilliant solar flare. The flare lasted only minutes, and all evidence was gone from the solar disk soon thereafter. Carrington reported to the Royal Astronomical Society that a moderate, but very marked, disturbance of the geomagnetic field was observed within two minutes of the time of the white light flare. In the early hours of the following morning, a great geomagnetic storm began. Carrington did not claim that the flare and the magnetic storm were connected, although he no doubt suspected it.

In the following years, considerable work was devoted to comparing the variation of geomagnetic activity with the varying number of sunspots throughout the eleven year cycle. By 1885, all leading astronomers had accepted the relationship between the solar cycle and geomagnetic disturbances.

Yet, opposition to this relationship was still in evidence in other branches of science.

In 1892, Lord Kelvin, in an address to the Royal Society, pointed out that to produce eight hours of a not very severe geomagnetic storm, the sun must do as much work in sending out magnetic waves in all directions in space as it actually does in four months of regular heat and light production. Kelvin stated that this result demonstrated that any connection between geomagnetic storms and solar magnetic activity is purely coincidental. Kelvin went on to say that the connection between sunspots and magnetic activity, and their cycles, is similarly coincidental. The acknowledgement of solar-terrestrial relations remains a subject of contention even today.

Direct observation of the sun was not the only method employed during the last century to establish a solar-terrestrial climatic link.

As mentioned earlier, the high plains of the United States seem to be afflicted with drought that occurs at regular intervals. The region's history over last century and a half was interestingly reported by Lansford in 1979. In 1820, an expedition led by Major Stephen H. Long reported the plains east of the Rockies to be "The Great American Desert." Later explorers, such as Major John Wesley Powell, expressed similar views. It was obvious to these observers that the area bounded by Abilene, Texas; Dodge City, Kansas; South Platte, Nebraska; and the farmland of California was incapable of supporting agriculture without massive irrigation efforts.

In the early 1880's, however, such pessimistic views were less prevalent. The railroads were bringing settlers encouraged by occasional heavy rains, and the claims of railroad promoters that the 'rain belt' was moving westward. Many homesteaders made decent livings on the plains during the 1880s. Late in the decade, however, rainfall grew scarce again. By 1890, the plains were in the grip of another severe drought. Many farmers, accustomed only to eastern farming methods, were wiped out. Once-thriving communities along railroad lines were reduced to ghost towns. Some hardy farmers were able to adapt their techniques to the variable climate. Drought-resistant crop varieties were developed, and farmers learned to cultivate semi-arid areas of the West successfully in the relatively higher rainfall years. The pattern of droughts every 20-22 years became well established to plains residents, who became proficient at living with it.

Other areas of the Southwest are so dry that even dryland farming is hardly ever possible. Fortunately, these regions have rivers which can be tapped for irrigation water. The rivers are fed by melting snow from the mountain ranges, but snow fall is far from constant on a year to year basis. The Reclamation Act of 1902 authorized the Bureau of Reclamation to build dams and develop irrigation systems in the West. Competition between states for river water was spirited. In 1922, seven western states signed the Colorado River Compact, which apportioned shares of the river to each state. Unfortunately, the compact was based upon highly optimistic assumptions of the amount of water available in the Colorado River. Flow records went back only a few years. A more detailed study of the river history indicates that the first twenty years of this century were the wettest in the Colorado River

Basin in the last two centuries (Lansford, 1979). The problem of adequately apportioning river water still exists. Western states and Mexico have claims on the Colorado which, when totaled, far exceed the amount of water in the river. Clearly, a detailed history of the climate is needed in order to properly and effectively allocate such resources in the future.

Andrew E. Douglas, an astronomer at the University of Arizona, began a study early in this century to map the climatic history of the Southwest. Douglas chose dendroclimatology, the study of climate through tree rings, as his method. Tree rings are formed by a layer of plant tissue just under the bark which produces large, thin-walled cells at the start of each growing season. This causes a distinct boundary between the first wood formed in the new season, and the last wood formed in the prior season. Douglas proposed that the relative thicknesses of tree rings in trees growing in harsh climates would mirror changes in climate. That is, during a drought year the ring would be very thin, but during a moist year the ring would be much thicker. Douglas was able to extend his tree ring research back several hundred years. His pioneering work continues at the Tree-Ring Research Laboratory at the University of Arizona. Modern techniques have correlated tree rings with temperature and moisture, and have reconstructed climatic conditions for 10 year periods over much of the western United States. Scientists at the Tree-Ring Lab have used data from 40 sites across the West to reconstruct the occurrence of drought back to 1700AD. The data shows a repeated pattern of widespread drought approximately every 22 years. The most interesting aspect of this finding, from the point of view of this chapter, is that the droughts coincide with the so-called double sunspot cycle. This cycle, also known as the Hale Cycle, is a 22-year pattern of rising and falling numbers of sunspots present on the solar disk. The cycle is known as double, because the interval between the maximum number of sunspots through minimum number and back to maximum is about 11 years. This 11 year cycle has apparently persisted since the end of the Maunder Minimum. A mechanism to explain the connection between sunspot and drought cycles has not been developed. Statistically, the evidence is strong, and few people dispute the likelihood of another drought occurring in the same 22 year cycle. However, until a workable mechanism explaining the connection is advanced, we cannot predict the droughts on a scientific basis.

## 16.2 The Sun During Recorded History

One of the problems of both solar and meteorological research is differentiating between normal and unusual. For example, we commonly employ rainfall measurements of the last thirty or forty years to infer a climatological average rainfall rate. After a very dry year, we say the rainfall was well below normal. Are we justified in claiming our thirty year data base sufficient to determine what is, and is not, normal? If we could expand our annual rainfall data to include several centuries we would probably come up with a completely different definition of normal. The same is true for solar activity. Can we realistically base our ideas of solar cycles on data going back little more than one hundred years? Eddy (1976) sought to address this problem by compiling solar data, both direct and inferred, as far back in time as possible.

It was assumed until recently that the eleven year sunspot cycle was regular and repeatable. The cycle may be only a temporary feature of recent history. According to Eddy, there is little evidence to suggest the eleven year cycle existed before about 1700AD. It seems much more likely that the sun exhibits changes of behavior over periods of hundreds to thousands of years. Furthermore, these periods of solar variation are probably not periodic but stochastic. Much of the work done to relate the eleven year cycle to climatic variation may have been in vain. Eddy's primary data sources are the carbon 14 ($C^{14}$) records of the past 5000 years and a detailed study of the Maunder Minimum.

### 16.2.1 Radiocarbon Dating

The $C^{14}$ isotope is produced in the upper atmosphere by galactic cosmic ray (GCR) bombardment. The GCR flux is not constant, so the production rate of the $C^{14}$ isotope varies in time. An important modulator of terrestrial $C^{14}$ production is solar activity. During periods of relatively high solar activity, the earth is shielded from GCR and vice versa. The shielding efficiency is a function of IMF and geomagnetic field strength. The strength of the earth's magnetic field varies by about a factor of two over a period of approximately 10,000 years, and does not strongly affect $C^{14}$ numbers.

Given a record of past levels of atmospheric $C^{14}$, it should be possible to deduce the history of solar activity. Such a record exists in carbonaceous fossils and in trees, where $C^{14}$ is assimilated as $CO_2$ during photosynthesis. Individual tree rings preserve a record of the prevailing $C^{14}$ to $C^{12}$ abundance ratio in the lower atmosphere at the time they were formed. This record can be read in living trees, such as bristlecone pines of the Sierra Nevadas, to around 3000BC. Well-preserved dead wood can extend this record back to before 5000BC.

A complication arises in the use of $C^{14}$ dating. There is a delay of 10-50 years between the instantaneous changes in upper atmospheric $C^{14}$ levels, and the resultant $C^{14}$ changes in the biosphere. This delay tends to wash out short-term changes, such as the eleven year cycle. The delay also displaces all effects in time. If we plot the $C^{14}$ content against time, and remove trends due to the varying geomagnetic intensity previously mentioned, we see two distinct increases in the level of $C^{14}$ production occurring over the last few centuries. Increased $C^{14}$ production implies lower solar activity. These two periods are known as the Maunder and Sporer Minima. Around 1200AD, the level of $C^{14}$ production shows a marked decrease. Solar activity presumably was quite high. This evidence corroborates historical data. Greenland was being colonized around this time, and wine producing grapes were grown in England. During the thirteenth century, the average temperatures fell considerably. Writers of the time tell of miserably long winters and short cool summers. Greenland iced over, wine production in England ceased, and the people of northern Europe were forced to adapt themselves to a harsher climate (which persisted for well over a century).

Other distinct changes in the $C^{14}$ production rate are evident in the past few thousand years. Unfortunately, modern technology has significantly reduced the value of $C^{14}$ study. The burning of fossil fuels has introduced significant levels of $CO_2$ into the atmosphere. This causes an abrupt drop in $C^{14}$ concentration due to the dramatic increase in $C^{12}$ production from fossil fuel burning. This result overwhelms the solar information in the modern radiocarbon record. Radiocarbon data since the middle nineteenth century cannot be employed to determine levels of solar activity. Nor can it be used to give a present standard of solar behavior to compare with the past. This is one reason why we cannot say with certainty whether the modern era represents normal or abnormal solar behavior. The trend in the earlier $C^{14}$ data curve suggests we are now in, or are approaching, a maximum level of solar activity. This trend is mirrored in the overall envelope of the sunspot cycle curve. It seems that we are now in a period of abnormally high solar activity, a level similar to perhaps only 10% of the past five millennia.

### 16.2.2 The Maunder Minimum

The years 1645-1715 are known in solar history as the Maunder Minimum. During this period, solar activity was apparently very low in comparison to the years before and since. The Maunder Minimum appears in the $C^{14}$ trend previously discussed. Fortunately, it is recent enough that other comparison observations are available for the time, and provide a check for $C^{14}$ data.

Eddy employs the Maunder Minimum as a cornerstone in his development of solar history. The period 1645-1715 was a time of unique solar behavior. Sunspots were very rare, and solar activity hovered at a level lower than that characteristic of the minimum in present 11-year cycles. For periods of up to ten years, no sunspots were seen, and none were observed in the solar northern hemisphere for 32 years. Reports of visible aurorae throughout Europe fell sharply during the Maunder Minimum, and rose abruptly after it. The solar electron corona was severely weakened or absent altogether. Observers of the sun at total eclipses during the Maunder Minimum reported very little light around the moon, and what light there was appeared of uniform breadth. The few sunspots that were reported during this period were isolated features, always at low latitudes. This whole period seemed, in Maunder's words, "a prolonged sunspot minimum." It seems impossible to determine whether an 11-year cycle was in effect during or in the thirty-five years prior to the Maunder Minimum. It is very possible that the early or mid 18th century saw the beginning of the present 11 year cycle series (Eddy 1976). Eddy also notes that solar rotation, as determined from sunspot transit times, was apparently faster just before the onset of the Maunder Minimum than is the present solar rotation rate. Equatorial regions seemed to rotate approximately 3% faster then than now.

### 16.2.3 Long Period Cycles

Eddy (1976) correlated data from the past 5000 years and presented it in tabular form (Table 16.1).

Table 16.1 Major Solar Excursions Over the Past 5000 Years.

| Feature | Beginning & End in Radiocarbon Record | | Probable Extent in Real Time | | Amplitude: 14C Correction | |
|---|---|---|---|---|---|---|
| 1. Modern Maximum | AD 1800? | _ _ _ | AD 1780? | — — — | ? | ? |
| 2. Maunder Minimum | AD 1660 | AD 1770 | AD 1640 | AD 1710 | -1.0 | -1.0 |
| 3. Sporer Minimum | AD 1420 | AD 1570 | AD 1400 | AD 1510 | -1.0 | -1.1 |
| 4. Medieval Maximum | AD 1140 | AD 1340 | AD 1120 | AD 1280 | 0.7 | 0.8 |
| 5. Medieval Minimum | AD 660 | AD 770 | AD 640 | AD 710 | -0.6 | -0.7 |
| 6. Roman Maximum | AD 1 | AD 140 | 20 BC | AD 80 | 0.6 | 0.7 |
| 7. Grecian Minimum | 420 BC | 300 BC | 440 BC | 360 BC | -2.0 | -2.1 |
| 8. Homeric Minimum | 800 BC | 580 BC | 820 BC | 640 BC | -2.1 | -2.0 |
| 9. Egyptian Minimum | 1400 BC | 1200 BC | 1420 BC | 1260 BC | -1.5 | -1.4 |
| 10. Stonehenge Maximum | 1850 BC | 1700 BC | 1879 BC | 1760 BC | 1.6 | 1.3 |
| 11. Summerian Maximum | 2700 BC | 2550 BC | 2720 BC | 2610 BC | 1.7 | 1.3 |

From Eddy (1976)

Eddy quotes an analysis by Damon (1976) suggesting that the Sporer and Maunder Minima may actually be parts of a single minimum in a long cycle of roughly 2500 years. Damon applied statistical analysis to the $C^{14}$ data to search for obvious cyclic effects. In the past 2000 years, he notes significant power at 56, 69, 182, and 400 years. In the two millenia before that, the significant periods were 286 and 500 years. Finally, in the period of 4000 to 6000 years before the present, the periods were 100, 286, and 1000 years. These findings are preliminary and, thus, not too much credence should yet be given them. However, one conclusion does seem reasonable. Long-term solar behaviour is variable; nice, clear-cut periodic cycles that have recently been ascribed to solar behavior are almost centainly valid only for short periods, if at all.

## 16.3 Global Climatological Trends

We have discussed the apparent connection between the prolonged period of relatively low solar activity and a cooler climate during the Maunder Minimum, and the evidence for the high plains drought cycle in concert with the double sunspot cycle. Attempts have also been made to relate solar trends with climate trends in many more locations and instances than those already considered. If the drought cycle is a direct result of solar variability, it seems reasonable to assume that other regions of the world should follow similar climate trends.

Many attempts have been made over the years to correlate sunspot number trends with various earthly phenomena. As long ago as 1801, Herschel suggested that sunspots indirectly controlled the London wheat prices, because of his observations that rainfall was directly proportional to the sunspot number. Comparisons have been made between the Dow Jones Index and sunspot number. One study even went as far as to suggest a connection between sunspot trends and the height of hemlines in women's skirts. Naturally, no scientific respect is accorded to such claims.

A study of the connection between sunspot cycles and trends in such parameters as temperature and rainfall yields mixed results. The correlation between sunspot number and annual rainfall amount may be positive, negative, or nonexistent (Herman and Goldberg, 1978). A positive correlation would imply increasing rainfall amount as sunspot number increases, a negative correlation implies the opposite. Correlations are expressed as decimal portions of plus or minus unity. For example, the strongest possible positive correlation is + 1, the strongest negative is - 1. Values near zero indicate no correlation at all.

At equatorial latitudes, there is a positive correlation between rainfall and sunspot number. On the average, more rain falls in solar maximum years than in solar minimum years. In a study of some middle latitude stations, the reverse seems to be true (Herman and Goldberg, 1978). It must be stressed that these studies used only a few stations. It is possible that quite different results would be obtained by using a larger data base. Another complication arises in some cases where the same stations are used, but at different times. For example, a study by Clayton in 1923 showed a positive correlation for annual rainfall at Hobart, Tasmania and Cairns, Queensland. A study of the same parameter at the same stations, but employing a much larger time period, 1880-1960, supported Clayton's findings at Hobart, but contradicted them at Cairns (Herman & Goldberg, 1978). Is it possible that the rainfall amount at Cairns changed because of some solar effect? It is hardly likely. What is much more plausible is that the weather at Cairns is not dominated by solar activity at all. Many other studies of rainfall rate as a function of sunspot number have been made. It is sufficient to say that correlations seem genuine in some regions, but rarely for extended periods of time. The correlation can appear to be very strong for a period of time and then disappear. For example, a study early in this century showed strong correlation, +.88, between the mean annual water level of Lake Victoria and the sunspot number. The correlation deteriorated considerably after 1930. Some areas show correlations at certain times of the year, and not at others.

The Ohio River near Cincinatti is higher in spring and early summer during solar minimum years, but no correlation is evident for the remaining seasons. Herman and Goldberg (1978) do suggest that the annual rainfall may be dependent on the 11-yr cycle, at least in equatorial and upper (i.e., greater than 40°) latitude.

Attempts to correlate temperature with sunspot cycles have also produced mixed results. One study for the years 1804-1910, showed that the global mean annual temperature was lower at sunspot maximum than at solar minimum.

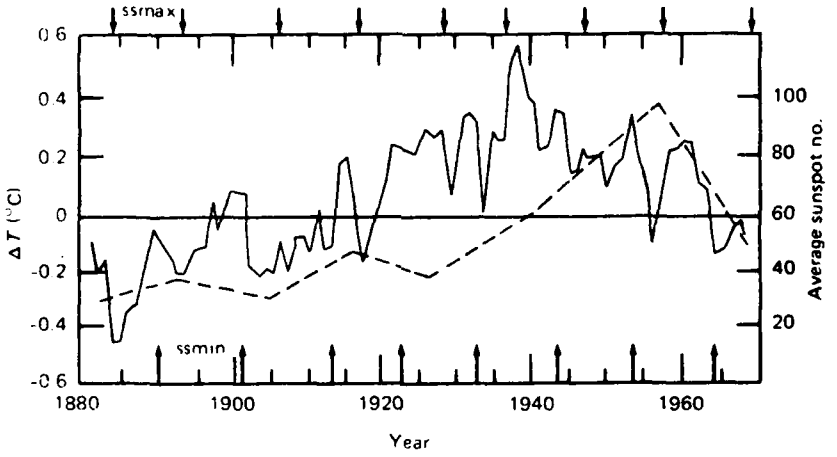



Figure 16.1 Northern Hemisphere Annual Mean Temperatures for 1880-1968 (solid line), compared to 11-year mean of annual sunspot numbers plotted on center (maximum) year (broken line). Years of ssmax are indicated by arrows at top, ssmin years by arrows at bottom (from Herman and Goldberg, 1978).

Unfortunately this finding tends to be contradicted in more recent years. In fact, the solar minimum of 1964 coincides with a temperature minimum rather than the maximum predicted. Perhaps this trend reversal is related to a long-term trend in solar activity. Other studies have attempted to correlate atmospheric pressure trends with solar cycles. Some results are encouraging, others are not.

Because of the conflicting results obtained when comparing solar activity with parameters measured at stations, different approaches have been tried. One study attempts to correlate movement of pressure centers with solar activity. Since major pressure centers, such as the Icelandic Low and the Azores High, exert great control over the general atmospheric circulation, it is perhaps more worthwhile to compare their positions at different times of the solar cycle than it is to record results only from fixed observation points. Results do show that the Azores High and Irelandic Low gradually migrated northward between 1889 and 1940 and then began moving to the southeast. The annual mean temperatures in the northern hemisphere displayed a similar pattern. Thus, there may be a connection between the pressure centers, temperatures, and an 80 or 100 year trend in solar activity.

We may summarize this section by saying that certain locations have, at certain times, displayed correlations between one or more meteorological parameters and sunspot cycles of varying length. It seems that for short solar trends, such as the 11-year cycle, the results for a given region vary considerably. For longer term cycles, we run into the problem of insufficient data. To accurately measure parameters against long term solar trends we need data going back many years. This data, in the detail required, is rarely available. The only solution seems to be to gather more data and continue the search for related patterns of terrestrial and solar trends. Indeed, there are those who attach very little credence to existing studies that purport to relate climate trends with solar cycles. Gerety et. al. (1977) compared temperature and precipitation records from 92 stations of wide geographical distribution against a solar analysis based on the 11-year cycle. The results varied considerably, and did not seem to fit into any geographically coherent pattern. The authors went so far as to say that the possibility of solar-terrestrial relationships in individual records resulting from chance could not be ruled out. Gerety and his co-authors present strong arguments to show that evidence of a short-term solar-terrestrial climatotological relationship is probably illusory. They could not rule out the possibility of a connection if longer solar cycles are employed. As stated previously, we need more data to effectively evaluate longer term solar cycles.

## 16.4 Solar-Terrestrial Weather Connections

The last section examined possible connections between solar activity and long-term climate variations. Long-term was generally considered to be eleven years and longer. Recent studies have attempted to link solar activity with short-term weather, on an almost daily basis.

Wilcox et. al. (1973) related the IMF structure to the total average area of high positive vorticity centers, observed during winter at the 300mb level. The IMF structure is conveyed radially outward from the sun by the solar wind and results in an interplanetary sector structure with the field polarity either towards or away from the sun in each sector. Adjacent sectors having opposite field polarities are separated by a very narrow boundary (Wilcox et. al., 1973). The vorticity employed in this study is an integrated vorticity area taken from absolute vorticity maps north of $20^{o}N$. Wilcox used the vorticity area index defined as the sum of the area, in square kilometers, over which the absolute vorticity (i.e., relative vorticity plus the Coriolis parameter) exceeded $20 \times 10^{-5}$ sec $^{-1}$ plus the area over which the vorticity exceeded $24 \times 10^{-5}$ sec $^{-1}$. A prominent feature of the vorticity maps was the association of regions of high positive vorticity with low pressure troughs, and, hence, significant weather. Wilcox examined the response of this hemispheric vorticity index to the passage of sector boundaries. The study showed that, on the average, the vorticity index starts to decrease about 1 1/2 days before the sector boundary passes the earth, reaching a minimum value about 1 day after boundary passage. The index then increases for the next 2 1/2 days. This effect was observed only in the winter months.

Later efforts related solar flares to a vorticity index at 500mb. The vorticity at the 500mb level is a rough measure of the strength of cyclonic activity over the Northern Hemisphere. An increase in vorticity area index (VAI) is noticeable in the first two days after a large flare. Then, after the geomagnetic storm which generally follows a major flare, there is a sharp decline in VAI, (Olson et al, 1975). Olson suggested an average sequence of events following a major solar flare. These results are primarily for the upper troposphere in the winter season. Olson defined the day of the large solar flare as day 0. Then, by day 1 or 2, the Northern Hemisphere VAI increases sharply, by 5-10% above its background level. By day 2 or 3 the geomagnetic storm has started, and may persist for several days. By day 3 or 4, the VAI has decreased to 5-10% below initial background level. By day 5 or 6, the VAI has recovered its original value. A factor not considered in this study was the disk position of the flare. In 1976, Wilcox and others refined the data to filter out the possibility of meteorological processes inducing geomagnetic activity. This was done by measuring sector boundary passage from spacecraft; beyond any tropospheric meteorological input. Again the results showed a vorticity area index response to geomagnetic activity following a major solar flare. It must be stressed that these results are statistical, rather like the high plains drought cycle and Hale cycle concurrence. A mechanism to explain the connection is needed.

As in the search for solar-terrestrial climate correlation, the scientific community is not in accord regarding the solar-terrestrial weather link either. Williams and Gerety, (1978), extended Wilcox's study period of 1963-73 an additional 4 years. Williams and Gerety claim the VAI response to sector boundary passage was not evident in the years 1974-77. The period, 1974-77, is analogous to 1963-66 in terms of the 11 year solar cycle. Both are around solar minimum. Williams and Gerety suggest that the years 1963-73 yielded an anomalous result of unusually high statistical significance. In other words, evidence for the VAI-geomagnetic link as an indicator of a sun-weather correlation is specious at best.

Efforts to link solar effects with short-term terrestrial responses continue, particularly behind the Iron Curtain. Bucha (1980) reports a strong positive correlation, 0.78, between geomagnetic activity and agricultural production in Central Europe and Canada. Bucha suggests that the energy of an intense geomagnetic storm may be sufficient to act as a trigger, at least in winter, to affect vorticity patterns in the troposphere. This suggestion is in line with the earlier work of Wilcox and others. Bucha goes beyond statistical analysis however, and proposes a mechanism for this effect. Bucha quotes earlier studies claiming electric currents of $10^4$ - $10^6$ Amps generated in the auroral oval during magnetospheric sub-storms as evidence that electrical conductivity of the auroral oval is considerably increased. As the energy dissipates, the temperature in the auroral oval will increase. Assuming maximum currents of $10^6$ Amps, a temperature increase in the center of the auroral oval, i.e., over the geomagnetic pole, of $13^oC$ is possible (Bucha, 1977, 1980). This corresponds to actual measured surface temperature increases during geomagnetic storms. Since the temperature increases, the

atmospheric pressure decreases, thus generating a low pressure area over the geomagnetic pole. This low pressure area, with a probable associated increase in positive vorticity, could significantly affect atmospheric circulation. According to Bucha (1980), an upper air cyclone above the geomagnetic pole may have its rotational energy increased by $10^{18}$ Joules over a 24 hour period. This should be enough coupling force to trigger the cyclonic process in the lower atmosphere.

Bucha's suggested mechanism is as follows. Given heating at altitudes of about 100km in the auroral oval, the velocities of the particles propagating towards the earth in the auroral oval increase. Then, adiabatic expansion of the medium leads to the generation of planetary pressure waves, and to their penetration through the auroral oval to the troposphere. This penetration is manifested as temperature changes observed at altitudes of 11-24km. At altitudes below 6km, a direct transformation of kinetic into thermal energy occurs due to oscillations and collisions of molecules in denser air layers carried along by the planetary wave propagating vertically downward. This leads to the marked warming of the air in the lowest layers of the atmosphere along the auroral oval, at first by 3-10°C. This warming reduces the atmospheric pressure and, thus, fuels cyclogenesis. This causes warm air masses to move northward leading to additional warming of up to 30°C. This marked temperature increase in especially prevalent in winter. This is because insolation is lower in winter, so the energy from the geomagnetic storm represents a proportionally much larger share of the total energy flux. Bucha claims high positive correlation between geomagnetic activity and auroral oval temperatures.

Bucha goes on to say that warming due to the geomagnetic storm takes place simultaneously along the whole auroral oval. However, the regions north of Siberia and in Canada, where the relatively low tropospheric temperatures exist, permit the planetary wave to propagate more easily. Within a few days following a geomagnetic storm, the auroral oval temperatures decrease gradually as areas of increased temperature move towards the geomagnetic pole. Temperature increases over the pole may be 20°C, with gradual penetration into the higher troposphere, and even into the stratosphere. The sudden increases in surface temperatures within the auroral oval result, after a lag time of 1-2 days, in considerable changes in the pressure patterns over the Northern Hemisphere. Bucha cites examples in which the major cyclones, such as the Icelandic and Aleutian lows, and the large anti-cyclone which dominates Eurasia during winter months, all showed considerable movement and strength changes following strong geomagnetic storms.

The evidence of Bucha (1980) is detailed and quite compelling. As mentioned in the previous section, it is difficult to relate solar activity with terrestrial climate responses if we restrict our measurements to fixed surface stations. If, however, we take into consideration the effects of solar-induced geomagnetic activity on major atmospheric pressure centers, more

consistent correlations may be forthcoming. It is these pressure centers, after all, that are responsible for the energy transfers that we observe as weather.

## 16.5 Other Possible Causes of Climate Variation

Until now, we have assumed the sun to be solely responsible for climate and weather changes. There are other suggested causes for climate change. A prominent theory relates climate to air pollution, both natural and man-made.

### 16.5.1 Volcanic Activity

The principle cause of natural atmospheric pollution is volcanic activity. According to Budyko (1969), a comparatively small variation in atmospheric transparency, due perhaps to an increase in volcanic dust, could be sufficient for the development of quaternary glaciation. The quaternary period is the geological age we are in now. Relatively small average temperature variations may be sufficient to trigger another advance of glacial ice. Such temperature variations occur quite rapidly. For example, it is well known that from the last years of the 19th century until about 1940 the average temperature increased. This trend has reversed since 1940. The temperature in the Northern Hemisphere increased by 0.6°C in the warming period, then decreased by the mid 1950's by 0.2°C (Budyko, 1969). These temperature changes may seem insignificant, but the growing season in England, for example, has been shortened by 2 weeks since the Second World War (Bryson and Murray, 1977).

Volcanic effects tend to be more dramatic. The 1815 eruption of Mt Tambora in Indonesia produced a stratospheric dust veil sufficient to reflect enough incoming solar radiation to lower average middle latitude temperatures in the Northern Hemisphere in 1816 1°C below normal. Conditions were so severe in some places that 1816 became famous as "the year without a summer." In some parts of England the summer was 3°C colder than normal. Certain regions had rain on all but 3 or 4 days from May to October. Canada and New England received widespread snow in June, and suffered frosts in every month of the year. Naturally, farms were especially hard hit. The food situation became so critical that riots broke out in Wales, (Bryan & Murray, 1977). The greatest volcanic eruption of modern times, that of Krakatoa in 1883, also affected the amount of solar energy reaching the earth. In fact, there have been periods of relatively high volcanic activity throughout history. We are in a period of quite high volcanic activity now. Yet this volcanic period did not begin with the general cooling trend that started around 1940. The cooling was already in progress. The issue is further complicated by the long stratospheric lifetime of volcanic ash. While volcanic ash is responsible for such phenomena as a "blue moon," it is probably not the sole cause of climatic change.

### 16.5.2 Continental Drift

Another possible cause of climate change, at least in local areas, is a shift in prevailing winds. Orogenesis due to tectonic activity can cause a deflection of prevailing winds, such that local climate changes considerably. Byran and Murray (1977) give several examples of this.

Related to mountain building is continental drift. As land masses migrate, atmospheric flow patterns are invariably affected. Obviously continental drift is not a cause of short-term climate change, but its importance in the modern perspective should not be totally discounted. For example, the Cinque Ports of southern England, which were famous in the Middle Ages, are now more than a mile from the coast. In certain key areas, it is not necessary to move land masses a mile in order to radically change weather systems.

### 16.5.3 Pollution

Opinion is still divided on Man's effect on climate. Our industrialized society puts millions of tons of material into the air each year. It has been estimated that human activity loads the atmosphere with as much dust as do the volcanoes. The dust reflects some sunlight, but it also absorbs some heat, which is in turn reradiated in all directions. Some people claim that this reradiation, known as the greenhouse effect, is sufficient to counter the loss of impinging solar energy due to reflection. Bryson and Murray (1977), have concluded, though, that the net effect of atmospheric dust is one of cooling. They go on to say that the dust is implicated in changing the patterns of prevailing westerly winds. These changes result in more varied weather around the globe, such as drought in monsoon dependent areas, and short growing seasons in the world's principle food producing regions.

### 16.5.4 Galactic Position

A more astronomical theory of climate change was described by Verschuur (1976). According to him, major climate changes, such as ice-age onsets, are caused by the solar system passing through gigantic dust clouds. Many galaxies, including ours, resemble a spiral in structure. Verschuur suggests that the arms of the spiral, which may be loosely or quite tightly wound, contain not only stars but also large accumulations of dust. Our solar system completes one revolution about the galactic nucleus every 250 million years. Each galactic orbit carries the solar system through the arms of our galaxy. This introduces varying concentrations of stellar dust into the solar system. Verschuur suggests that ice ages are a function of the solar system's position relative to a galactic spiral arm dust cloud. When in the cloud, solar brightness is reduced, thereby causing an ice age. Because the dust cloud is so thin it is not always possible to determine exact dust concentrations. Moreover, accretion of dust by the sun may significantly alter its brilliance or even produce the variability we know as the sunspot cycle. That other stars are known to have "starspots", flares, and undergo "activity cycles" provides additional credence to cosmic-scale variations in the sun. If these affect terrestrial climate, they should similarly affect the other planets.

For now, sufficient observations are lacking. As observational capabilities increase, we will, according to Verschuur, be able to plot and predict our passage relative to the dust clouds, and perhaps be able to predict future ice ages.

### 16.5.5 The Answer?

As you can see, there are many theories as to the cause of climate variation, and many conflicting claims with regard to a solar-terrestrial connection. This chapter certainly raises more questions than it provides answers. At present, our knowledge is not sufficient to identify, with certainty, all the links between the sun and the earth. What we can say is that knowledge of such links and their relative importance is vital to the survival of mankind.

The earth's climate is complex. It is probably not possible to apply generalizations to the entire surface of the globe. The number of parameters that constitute climate, and the number of permutations possible from those parameters, are large enough to thwart efforts to envision a single mechanism for a definite solar-terrestrial connection. Significant work on computer models of climate is continuing though. Since the end of World War II, the capabilities of computers have increased remarkably. Much of this increase can be traced directly to the needs of weather forecasters. The pioneering efforts of Jon van Neumann and others at Princeton and MIT in the late forties and early fifties provided the necessary computer capability to run the early barotropic and baroclinic models.

These prototype models were designed to forecast conditions from a few hours to, at most, a few days in the future. Nowadays the goal is to model the entire atmosphere and be able to run it both backwards and forwards in time for years, and even centuries. The main problem is still the compilation of necessary and sufficient data to give the model an accurate starting point and ensure worthwhile predictions.

Perhaps the most comprehensive model now in operation is at NASA's Goddard Institute for Space Studies. As many parameters as can be utilized are fed into the model, and the computer time-steps forward or backward similar to the way the commonly-used Limited Fine Mesh (LFM) model is generated. The Goddard model is so complex that a complete run can take several weeks. But even this model does not take into account all factors. The ocean's ability to transfer heat is not considered due to its complexity. This is an extremely important parameter to omit. Should the global temperature increase slightly, some oceanographers have speculated that the Gulf Stream would become sluggish due to a decrease in the temperature gradient between low and upper-middle latitudes. An enervated Gulf Stream would permit much more cold air to sweep over Northern Europe. Ironically, this would result in a much colder Britain and Scandanavia, even though the global average temperature had increased.

Current models, with their vast inputs, can actually suffer from too much initial knowledge. A single computer run, requiring weeks or possibly months, can be so complicated by detail that the result becomes meaningless.

Despite the difficulties inherent in such complicated models, many definite trends can be observed. For example, the cooling trend that began in the 1940's has apparently ended. The Goddard model predicts a warming trend throughout the 1980's, culminating in average temperatures above those experienced in 1930's. A warming trend due to the burning of fossil fuels, the greenhouse effect, is now given a higher probability of occurrence than it has had in recent years.

These results must still be viewed with caution. The complex interactions of the many climate parameters have yet to be satisfactorily modelled. Climate change is undoubtedly due to many factors; some, at least, solar related.

## 16.6 Summary

The search for solar-terrestrial connections has so far been a study of statistics. We can trace, with reasonable certainty, the climatic history of the past few millennia, but we encounter serious difficulties in correlating solar variation with climatic change. The Maunder Minimum provides the cornerstone in relating solar variation with climatic change, yet some observers even question the existence of the Maunder Minimum.

It is extremely difficult to correlate climatic parameters for a particular station with solar cycles for long time periods. Early claims of high correlation almost invariably fail to hold up as the observation time is extended. Hines (1976) warns against the pitfalls of stumbling onto an apparently significant statistical relationship, and then investing research time only to realize that the alleged correlation was no more than a statistical quirk. That does not mean that seemingly unrelated statistical matches should not be investigated. It does mean that a high degree of skepticism must be maintained. For example, Staupel (1980) argues that there is a higher mortality rate from cardiovascular failure in the morning hours of active solar days than on non-active days. This may be due to some physiological response to magnetic activity, or it may simply be a statistical coincidence. If research bears out Staupel, solar forecasters may, at some future date, be required to advise hospitals when geomagnetic storms are expected.

Perhaps the statistics we should be most concerned with are given by Brysan and Murray (1977). What they basically say is what we tend to consider as climatically normal is really not so. About 90% of the past million years have been colder than the present. In fact, since 1700AD, all 30 year periods have been colder than the 1931-60 period. We are in a cooling trend that began around 1940. Already, growing seasons have shortened appreciably. If the cooling trend continues, and if the world population increases, world food production, already straining to keep up with needs, cannot possibly be sufficient to meet future demands. If, as computer models suggest, we are entering a warming trend, then the chances of crop failure due to drought increase. In either case, any slight change in average temperatures will have a much greater effect in the future than it had in the past because of the much larger population. The need to understand and predict climatic variations, whether solar activated or not, has never been greater.

# GLOSSARY

A brief introduction to the jargon of space forecasting.

AFCC Job: Also AFCS or PROPO job. Air Force Communications Command request for radio propagation conditions forecast for a specific path or paths and date(s). Uses HFMUF3 software.

AGDB: Astrogeophysical Data Base: The primary SESS data storage file on AFGWC computers. (AFGWC * DATMANSESBAS)

Akasofu Parameter: Also, Epsilon. This number provides an estimate of the energy per unit time (ergs/sec) which the magnetosphere extracts from the solar wind. It is calculated using solar wind velocity and magnetic field data measured near the earth's orbit.

Altitude: Also, elevation. The angular distance of a celestial body above the horizon measured upwards towards the zenith. The zenith has an altitude of 90°.

Ap: A dimensionless number or index providing a linear measure of the level of disturbance of the planetary geomagnetic field. Ap is a 24-hour index, and ap is a 3-hour index. See also K.

Apogee: High point for an earth-orbit spacecraft (maximum distance from each in a closed orbit).

ATN: Astrogeophysical Teletype Network. The original teletype system designed to link solar observatories with the forecast center and primary customers. Now nearly obsolete.

Aurora: Visible aurora results primarily from emission lines of oxygen and molecular nitrogen excited by energetic particles which precipitate in the auroral oval. Radio aurora is often observed as well and results from field-aligned irregularities (due to particle precipitation) which produce scattering of radio waves.

AU: Astronomical Unit, the mean earth-sun separation. The currently accepted value is near 93 million miles.

Auroral Oval: The actual, real-time location of aurora. The oval location varies with time and the level of geomagnetic activity, and is actually more nearly a circle.

Auroral Zone: The locus of the average position of the midnight auroral oval. It defines the location of highest probability for the occurrence of aurora and varies with the level of geomagnetic activity.

AWN: Automated Weather Network, typically in reference to the high speed teletype system which links Air Weather Service sites worldwide via a computer switching system. At Global Weather Central, AWN generally refers to the high speed data link between GWC and the central AWN switch at Carswell AFB, Texas.

Azimuth: The angular displacement of a celestial body from true north measured around the horizon from north (0°) through east (90°). Its range is 0° - 359°.

Bistatic: Two station system. Typically with respect to a radiowave communications system employing a geographically separated transmitter and receiver.

Bottomside: That portion of the ionosphere below the level of maximum electron density (i.e. usually the portion of the ionosphere below about 400km).

Brilliance: An optical index of the intensity or energetic nature of a solar flare. The three categories are faint (F), normal(N), and brilliant(B), and are measured at or near 6562.8 Angstroms (hydrogen-alpha). Determined by light intensity when using a SOON videometer. Using a system without videometer, it is based on the Doppler shift observed: $\pm$0.4A(F), $\pm$0.6A(N), and $\pm$1.0A(B).

Burst: Usually at radio or x-ray wavelengths, it is a sudden, marked increase in emission with respect to background levels. Also, the name of an Air Weather Service code for reporting radio frequency bursts.

C-Event: The least energetic type of solar event (usually, a flare) routinely recorded. Normally defined by the peak x-ray emission observed in the 1-8 Angstrom band. For a C-event, peak emission is between $1.0\text{-}9.9\text{X}10^{-3}$ ergs/$cm^2$/sec. Often stated as C5.3, for example, if peak emission were $5.3\text{X}10^{-3}$ ergs/$cm^2$/sec. See also M and X events.

Celestial Equator: An extension of the earth's equator into space. It divides the sky into two hemispheres analogous to the north and south terrestrial hemispheres.

CM: Short for central meridian. In reference to the sun, it is the line connecting the north and south solar poles so as to divide the visible solar surface into halves.

Control Point: The point at which a radio wave is reflected/refracted from the ionosphere to permit long distance radio communication. Determined by the location of the transmitter and receiver and the electron density profiles of the ionosphere. There will be a control point for each "hop" of the path.

Coronal Hole: A region of low density and open magnetic fields in the solar corona. They are a primary source of the high speed solar wind streams observed near the earth. As such, they are closely related to the occurrence of geomagnetic disturbances.

CRT: Cathode Ray Tube. A common data entry device for computers; also, "scope".

Critical Frequency: The highest frequency reflected at normal incidence from a given ionospheric layer, it is approximately equal to the electron plasma frequency of the layer.

D-Region: The lowest, commonly referenced level of the ionosphere. It ranges from approximately 75 to 90km above the earth's surface and is primarily responsible for absorption of radio wave energy.

Declination: Angular distance of a celestial body north or south of the celestial equator-something like the "latitude" of a star, planet, etc. Essentially a constant for distant objects. Variations are significant only for nearby objects (e.g., sun, moon, planets).

Disk: The half of the sun visible from the earth at any given instant.

Dispersive: A medium in which effects vary with frequency (or wavelength) of an EM wave.

E-Region: The ionospheric layer ranging from about 90 to 150 km above the earth's surface and containing electrical currents known as electrojets. The level of sporadic E (ES) occurrence.

EDP: Electron Density Profile: Variation in the density of free electrons with altitude. It is a vertical sounding of the ionosphere and is usually constructed at least partially from theory.

EP: Energetic Particle data provided by a series of geostationary satellites. Electrons ranging from 30KeV to 2MeV and protons from 50KeV to 150MeV are measured. Also refers to the encrypted lines which bring the data to GWC directly from Buckley ANG Base, Colorado or by way of Patrick AFB, Florida.

eV: Electron volt - a measure of energy.

F-Region: The highest normally referenced level of the ionosphere ranging from 150km above the surface to the base of the plasmasphere (about 1000km). Separates into the F1 (lower) and F2 (higher) layers during daylight and possesses the highest electron density of the ionosphere.

Flux: The amount of something (protons, x-rays, radio energy, popcorn, etc.) arriving at a detector of a certain size in a given time period, or passing through a specified area in a give time period.

foES: Critical frequency (of the ordinary wave) of the sporadic E layer. The highest frequency reflected by the layer at vertical incidence - depends on electron density of the layer.

foF2: Critical frequency (highest frequency reflected at vertical incidence of the F2 layer for the ordinary wave. Related to the maximum usable frequency (MUF) and determined by the electron density of the F2 region.

fmin: Minimum frequency observed by an ionosonde. Depends on the ambient ionosphere, ionosonde power and sensitivity, and the nearby radio environment.

4-D: A static four dimensional ionospheric model used by GWC to model the current state of the ionosphere using a variety of observations and climatology. "4D" refers to the dimensions of latitude, longitude, height, and time. (Accounts for diurual variation, but is not dynamic or predictive.)

F10: The radio flux observed at or near a wavelength of 10.7cm (2800MHz) by the Ottawa Radio Observatory at 1400Z, 1700Z, and 2000Z daily. The 1700Z value is recorded as the F10 for that day.

F10 Bar: Also F bar, this is the 90-day running mean value of the 1700Z F10.

Gamma: A Greek letter used in two ways. It is one classification used by Mt. Wilson and other solar observatories to identify magnetically complicated (polarities intermixed) sunspot groups. Magnetically, it is a unit of magnetic field strength, 1 gauss = $10^5$ gamma. The average surface strength of the geomagnetic field is about 1/2 gauss, while the average strength of the interplanetary magnetic field (IMF) is 5-25 gamma.

GDO: Global Weather Central Duty Officer. Usually a lieutenant colonel with acting, operational control of GWC.

GHz: Gigahertz: A unit of frequency equivalent to a thousand MHz, or a billion cycles per second. The corresponding wavelength (to 1GHz) about 30 cm (for electromagnetic radiation).

GOES: Geostationary Operational Environmental Satellite: A series of spacecraft designed to monitor weather and the near-earth space environment. They provide SESS observations of energetic particles (alternative to EP data) and x-rays. Operated by NOAA.

H-Data: Spacecraft observations of visual aurora provided to and manually analysed by SESS to detemine the equatorward boundary of the auroral oval. Also, a general reference to all solar data transmitted by teletype, where it carries a four letter heading (MANOP) beginning with "H".

H-Alpha: Hydrogen alpha refers to an absorption/emission line of the hydrogen atom at 6562.8 Angstroms-red light. It is the first Balmer series line. Most solar optical observations made by AWS are at this wavelength, which originates in the lower chromosphere.

HF: High Frequency, 3-30 MHz, radio wave band. Normally used for long distance communication by reflection/refraction in the ionosphere.

HLMS: The High Latitude Monitoring Station is jointly operated by the Air Force and NOAA near Anchorage, Alaska. Also refers to the data provided by this station which gathers and processes data from Thule, Greenland, and the Alaskan area. May also refer to the communications link which carries this data from Alaska to Carswell, where it is added to the AWN.

IMF: Interplanetary Magnetic Field.

Importance: An index of the size of an optical flare in millionths of the visible hemisphere of the sun. The earth would cover about 100 millionths of the visible hemisphere. Categories are 0 (subflare, less than 200), 1 (200-499), 2 (500-1199), 3 (1200-2400), and 4 (greater than 2400). It is measured at peak optical brilliance.

IP: Interplanetary, outside the magnetosphere.

IPP: Ionospheric Penetration Point, generally refers to the geographic point at which a radio wave passes through the altitude of about 400km in transit between a ground station and a satellite.

ISEE-3: Vehicle 3 of the International Sun Earth Explorer satellite system. It is located between the earth and sun, about a million miles from the earth. It is a source of real time observations of the solar wind and the interplanetary magnetic field (IMF).

ITS-78: Institute for Telecommunications Science ionospheric climatology intended for use in operations and planning of long-range HF communications. It uses data gathered in 1958 and 1964.

J-DATA: Spacecraft energetic particle data are provided under this name. Early spacecraft used the J-3 sensor which detected only electrons 20eV - 50KeV. The J-4 sensor will fly on new satellites, measuring both electrons and protons. The primary use of this data is determination of the auroral oval boundaries.

K: Also Kp. These are semi-logarithmic indices of geomagnetic activity for a 3-hour period. The "p" subscript is a planetary (opposed to a single station) index. K ranges from 0-9 and Kp in 27 steps from 0o to 9o with one-third unit steps (0o, 0+, 1-, 1o, 1+, etc).

K-Band: Users' designation for frequency band centered near 25 GHz.

KeV: Thousand electron volts; a measure of energy.

KHz: Thousand Hertz (cycles per second); a measure of frequency.

L-Band: Designation for a frequency band centered near 1415 MHz.

Limb: The visible edge of the sun. The east limb is that nearest the observer's eastern horizon at sunrise.

LUF: Lowest Usable Frequency, generally in reference to long range radio communication in the HF band. It is a function of D region absorption and numerous equipment parameters.

M-Event: Also, a minor event. See C-event, X-event. M-events are defined by peak X-ray emission of 1.0 - 9.9 X $10^{-2}$ ergs/$cm^2$/sec in the 1-8 Angstrom band. A fixed frequency (e.g. 245 MHz, 15,400 MHz, etc.) radio burst of 500-9999 SFU is considered an M-event if the associated x-ray burst is insufficiently strong.

Mag: Short for magnetic, geomagnetic, or magnetometer. Sometimes in reference to magnetic storm. See also Ap and K. A minor mag storm exists when ap is greater than or equal to 30. At and above an ap of 50, a major "mag" storm is said to exist.

MANOP: Manual of operations. A general term describing procedures for handing operational message traffic by teletype, phone, etc. Often, MANOP header. For SESS, this refers to a four letter code group at the beginning of

a message. The first letter describes the general data type (H for solar--heliospheric). The second letter specifies the class of data (I--ionospheric, E--event, etc.), and the last two letters identify the country of origin (US-United States; CN--Canada, etc.). HEUS would be a MANOP header for solar event data from a US station.

Max: Short for maximum. Often used in reference to the peak emission of a solar event or to the peak in the sunspot cycle (solar max). Sunspot cycle max is determined by the level of such indicies as the Solar Sunspot Number (SSN) which is roughly equivalent to ten times the number of spot groups plus the number of spots.

MeV: Million electron volts, a unit of energy. Electrons and protons with energies in the MeV range are considered "highly energetic" to "relativistic" (moving at speeds which are a significant fraction of the speed of light).

MFAC: Multiplication Factor. It is the ratio of the MUF/foF2 and is usually specified for a stated, single-hop path length (e.g. M3000 is the MFAC for 3000km). In actuality, it is an index of the height (hmax) of the maximum F2 layer electron density. The greater the height, the smaller the algebraic value of the MFAC. The typical range is 2.0 to 4.0 (400km to 200km, approximately).

MHz: Megahertz, millions of cycles for second, a measure of frequency. 3-30MHz is the HF frequency band.

Min: Minimum with reference to the background level of solar emission before or after a solar flare. Also used in reference to the solar cycle. Solar min occurs when indices such as the smoothed sunspot number or F10 reach a minimum.

Microwave: General reference to the frequency band extending from about 300 MHz to 500 GHz used by radars. Technically includes all frequencies above 3MHz.

Monostatic - Single station, usually with respect to a radio wave system. Transmitter and receiver are collocated, e.g. a radar.

MRF: Multipath Reliability Factor. A percentage of the MUF, intended to provide a frequency which guarantees a predetermined degree of freedom from multipath interference.

MUF: Maximum Usable Frequency, usually with respect to long range, HF propagation. It is related to the foF2 at the control point and is a theoretically determined number. By definition, it will likely be usable only 50% of the time.

Multipath: May include non-great circle propagation. Implies that the radio wave splits up and follows several different paths to the receiver. Since the paths may be of different lengths, the arrival time and phase via each path will vary. The result may be intermittent fading and/or reinforcement of the signal received.

Mystic Star: Similar to PROPO or AFCS jobs. High priority support/forecasts for usable HF frequencies on specified paths to Presidential Airways, Andrews AFB.

NCS: Network Control Station, usually with reference to the HF radio control station at Kennedy Space Center.

Neutral Line: Used in analysis of solar features, it is also (more accurately) called a magnetic inversion line. Neutral line is a misleading term, since a magnetic field does exist. This line separates solar magnetic fields of opposite polarity. It may be defined by a dark filament or a plage corridor and, by its complexity, (kinkiness) often indicates the flare probability in a given region. Often a preferred site of large flares.

NOAA: National Oceanic and Atmospheric Administration, a segment of the Department of Commerce which oversees the civilian components of SESS.

OTHB: Over-The-Horizon Backscatter, usually a radar system. Uses Doppler shift of moving targets to identify them against terrain. Operated in the HF band to permit ionospheric reflection of the radar beam.

P-Band: Designation for frequency band centered near 300 MHz.

Perigee: Point of closest approach to earth for earth-orbit spacecraft.

PCA: Polar Cap Absorption Event defined by level of absorption (0.5dB night, 2.0dB day) on a 30MHz riometer in the polar cap, usually Thule. Caused by high energy (greater than 5MeV) protons precipitated into the polar caps following a large solar flare. Also called a polar blackout because of its effects on HF radio propagation.

PCAF: Polar Cap Absorption Event Forecast. It is a color-coded condition or warning similar to weather warnings. Range is green (a PCA is unlikely); yellow (a region exists on the sun which, if it produces a large solar flare, will probably produce a PCA); red (a large, energetic flare has been observed, and a PCA is considered imminent); and in progress.

Polar: With respect to the AFGWC Polar Ionospheric Model, a modification of ITS-78 climatology to accurately reproduce auroral oval and trough electron density gradients. Input specifies the oval position. Output fields can be modified using IONMOD series software.

Q: Also Qe. An index used to specify auroral oval postion. It is a quarter hour index with a theoretical range of 0-10 in quasi-logarithmic steps. Qe data ranges from about -4 to +12 and is based on AFGWC analysis of DMSP observations. Theoretically, Qe = 2Kp -.35.

PROPO: See AFCC job.

RAD: A measure of absorbed radiation equivalent to 100 ergs per gram of body weight.

Radar Aurora: Radar returns reflected by Ionization resulting from particle precipitation near or in the auroral oval. It may not be identical with visual aurora in spatial or temporal extent and is aspect dependent.

Raday: Short for Radio Day. Just another term for a Zulu or GMT day (e.g., Raday 3 Nov goes from 03/0000Z Nov to 03/2359Z Nov).

Ramp: Also, pi-ramp, or ramp value. It is used with respect to a Faraday Rotation Polarimeter (used to measure total electron content-TEC). This instrument records the change in the signal polarization from the time it leaves the satellite until it reaches the antenna. The amount of polarization change (rotation of the polarization vector) is a function of magnetic field strength and ionospheric electron content between the satellite and the ground station. A full rotation of the vector is 360° (2 pi radians). Since the head can't be distinguished from the tail of the polarization vector, and since the exact number of rotations can't be known precisely (electron content is never zero) there is an ambiguity in the measurement by an unknown number (n) of half (1pi) rotations, called the n-pi ambiguity. As electron content changes, the polarimeter strip chart records rotations. Actually, the chart recorder only permits recording a pi rotation at a time, and these have the appearance of a ramp. The change in electron content required to alter the polarization by 180° is calculated from theory for each station (it is a function of position and time in addition to electron content) from so-called ROPLOK tables. These are ramp values.

Razdow: A fully manual solar patrol telescope first used by SESS in the late 1960's. Replaced at most observatories by SOON equipment beginning in the 1970's.

REM: A unit measuring the biological impact of a given radiation dose.

Relativistic: Particles with sufficient energy to move at speeds which are a significant fraction of the speed of light (10% or more, usually).

Right Ascension: Angular displacement of a celestial body eastward from the vernal equinox. Measured along the celestial equator. Similar to the "east longitude" of an object. Varies significantly only for nearby objects (e.g., sun, moon, planets).

RSTN: Radio Solar Telescope Network, an acronym for the system of standardized, computer-controlled solar radio telescopes. Standard frequency compliment includes 15,400; 8,800; 4,995; 2,695; 1,415; 606; 410; and 245MHz. Additionally, a sweep frequency interferometric radiometer monitors the 25-75MHz bandwidth.

S/N: Signal to noise ratio. Each is usually specified in decibels above some reference level, so the ratio is usually dimensionless. The larger the better for a communicator.

S-Band: A radio frequency band centered near 2,695MHz.

Scintillation: A rapid variation in amplitude or phase of an EM signal (usually on satellite communications links). Frequencies are usually transionospheric (thosegreater than 30MHz). It is a consequence of variations in electron density along the line of sight.

SDO: Systems Duty Officer: Officer having operational responsibility for computer systems at Global Weather Central.

SEON: Solar Electro-optical Observing Network, acronym for combined SOON-RSTN observatory.

SESB: Space Environmental Support Branch, Air Force Global Weather Central. Also known as AFGWC/WSE.

SESC: Space Environment Services Center, NOAA, Boulder, Colorado. The civilian counterpart of SESB.

SESS: Space Environmental Support System, a blanket acronym referring to all components of the real-time solar-terrestrial patrol and forecasting network. Also used to identify solar-terrestrial components of larger organizations or data provided by this network.

SESSPROP: A military planning document to provide long-term computer software development guidance for the SESS component of AFGWC.

SFU: Solar (or Standard) Flux Unit: A measure of emitted radio energy equal to $10^{-22}$ watts/$m^2$/Hertz. Standard units for reporting solar radio flux and bursts.

SID: Sudden Ionospheric Disturbance, a real-time response to solar activity. In particular, a response to solar x-ray emission associated with a solar flare. Limited to the sunlit hemisphere. Effects result from varying the electron density of the D and/or E regions of the ionosphere.

SINPO: Name of a coded report of a 5-digit radio propagation analysis provided by NORAD HF backup paths. It represents observed conditions on a scale of 1 (awful) to 5 (normal) for Signal strength, Interference, Noise, Propagation conditions, and Overall circuit evaluation.

Skip Distance: Minimum distance at which sky wave propagation will provide reliable communications.

Skip Zone: Region where radiowave communication is impossible.

Slab: Slab nomogram or slab thickness calculation in reference to measuring total electron content. It is often convenient to assume that the net effect of all the electrons in a given column can be equated to the effect of a slab of electrons of constant density and a predetermined thickness.

Solar Wind: An extension of the sun's corona into interplanetary space. The solar wind is a low density ($5/cm^3$) plasma expanding at near sonic velocities (300-1000km/sec) outward from the sun. It carries wave and density structures, defines the interplanetary magnetic field, and shapes the magnetospheres of the planets.

SOLMF: Solar mean field polarity (magnetic) observations taken at the Stanford solar magnetograph. It provides a measure of the dominant magnetic polarity of the visible solar hemisphere. It is also the code type name.

SOON: Solar Observing Optical Network, a system of automated flare patrol telescopes capable of objective H-alpha and spectrographic observations of the sun. The follow-on system to the Razdow.

SPAN: Solar Proton Alert Network, a system of civilian solar observatories operated via Boulder from the late 60's thru the mid 70's. Intended primarily to support the manned space program, it included sites at Culgoora and Carnarvon (Australia), Boulder, and the Canary Islands. Razdows were the primary optical telescopes.

SSIE: Insitu spacecraft plasma sensor to measure electron temperature, scale height, and ambient electron density at 840km.

SSIMF: Solar sector inferred magnetic field polarity based on Vostok and/or Thule magnetometer observations. Also the name for the code in which this data is transmitted.

SSN: Sunspot Number or Solar Sunspot Number. Smoothed Zurich SSN refers to the number of sunspots and sunspot groups observed each day on the sun. The Effective SSN often used by AFGWC is an index used to measure the average state of the ionosphere with respect to the ITS (Institute for Telecommunications Services) 78 climatology, and is unrelated to the number of visible sunspots.

Substorm: A short-term (2-3 hours), highly-localized (in or near the auroral zone) geomagnetic disturbance. At latitudes equatorward of the auroral zone, such disturbances are termed geomagnetic bays. May also be called polar or auroral substorms when they occur near the auroral zone. When this occurs, increased D-region absorption and visual and radar aurora are often observed.

Sunrise: (and sunset) Defined by the sun's central point having a zenith distance of $90^{o}$ 50'. This is ground sunrise/sunset and differs from ionospheric sunrise/sunset. D region sunrise/sunset occurs for a solar zenith distance of about $102^{o}$.

Sweep: Short for Sweep Frequency Interferometeric Radiometer and the data provided by these instruments. They monitor the 25-75MHz radio spectrum and are capable of determining how a burst changes in time at each frequency in this range. Such bursts are generally classified by structure more than by peak intensity.

SWF: Shortwave Fadeout, the LUF enhancement resulting from a flare-associated x-ray burst and affecting sunlit HF circuits. This is a D region phenomenon, and the strength of the fadeout is inversely proportional to frequency. The larger the burst, the higher the frequency affected. An SWF is one type of SID.

TEC/TELCO: Total Electron Content as measured by a Faraday Rotation Polarimeter (commonly). This may not be equivalent to the actual column electron content over the station, because the polarimeter measures along a slant path to the satellite. It responds only to the electron density below 1000-2000 km.

Topside: In reference to the ionosphere, it is that portion above the height (400km) of maximum electron density. Sometimes taken to include the plasmasphere.

Twilight: For civil twilight, the center of the sun was a zenith distance of 96°. For a zenith distance of 102°, it is nautical twilight. Astronomical twilight exists when the center of the sun has a zenith distance of 108°.

URSI: Union of Radio Science, International. This is a blanket agency for the international exchange of solar-terrestrial data. Much of this data comes to GWC by way of World Data Center A at Boulder, Colorado. A number of codes, prefix U, have been devised for this data (e.g. UFOFH, UMAGE, etc.).

VI: Vertical Incidence, typically regards an HF ionosonde or the data provided by this type of system.

X-Band: Designation for frequency band centered near 8,800 MHz.

X-Event: See also C and M-event. A major solar event as determined by a discrete frequency radio burst with peak emission in excess of $10^4$ SFU or an x-ray burst with peak emission in excess of $1.0 \times 10^{-1}$ ergs/cm$^2$/sec.

Zenith: The point directly overhead or 90° above the horizon.

Zenith Distance: The angular separation between a celestial body and the zenith. The zenith distance of the horizon is, by definition 90°.

BIBLIOGRAPHY

Akasofu, S.-I., "Interplanetary Energy Flux Associated With Magnetospheric Substorms", Planetary and Space Science. Vol 27, 1978.

_____, Polar and Magnetosphere Substorms, Springer-Verlag New York Inc., New York, 1968.

Akasofu, S.-I. and Sydney Chapman, Solar Terrestrial Physics, University Press, Oxford, 1972.

Allen, C.W., Astrogeophysical Quantities, Third Edition, The Athlone Press, London, 1973.

Atwell, W., "Dosimetry in the Manned Space Program", Solar-Terrestrial Predictions Proceedings, Space Environment Laboratory, Boulder 1980.

Babcock, H.W., "Topology of Sun's Magnetic Field and the 22-year Cycle", Astrophysical Journal, Vol. 133, 1961.

Bray, R.J. and R.E. Loughhead, Sunspots, International Astrophysics Series, Vol. 11, Chapman and Hall Ltd., London, 1964.

Bruzek, A. and C.J. Durrant, editors, "Illustrated Glossary for Solar and Solar-Terrestrial Physics", Astrophysics and Space Science Library, Vol. 69, D. Reidel, Dordrecht, 1977.

Bryson, R. A. and T. J. Murray, Climates of Hunger, University of Wisconsion Press, Madison, 1977.

Bucha, V., "Weather and Climate Prediction in the Northern Hemisphere Based on Solar-Terrestrial Relations", Solar-Terrestrial Predictions Proceedings, Vol. 4, Space Environment Laboratory, Boulder, 1980.

Budyko, M. I., "The Effect of Solar Radiation Variations on the Climate of the Earth", Tellus, Vol. 5, 1969.

Bueche, F., Technical Physics, Harper and Row, New York, 1981.

Carrigan, Anne L. and Robert A. Skrivanek, "Aerospace Environment", Air Force Cambridge Research Laboratory, Bedford, 1974.

Castelli, J.P., G.A. Michael, and J. Aarons, "Flux Density Measurements of Radio Bursts of Proton Producing and Non-proton Flares", Journal of Geophysical Research, Vol. 72, 1967.

Chapman, S., "The Absorption and Dissociative or Ionizing Effect of Monochromatic Radiation in an Atmosphere on a Rotating Earth", Proceedings of the Physics Society, Vol. 43, 1931.

Chappell, C. R., K. K. Harris, and G. W. Sharp, "Ogo 5 Measurements of the Plasmasphere During Observations of Stable Auroral Red Arcs", Journal of Geophysical Research, Vol. 76, 1971.

_____, _____, _____, "The Dayside of the Plasmasphere", Journal of Geophysical Research, Vol. 76, 1971.

Cladis, John B., Gerald T. Davidson, and Lester L. Newkirk, editors, The Trapped Radiation Handbook, Defense Nuclear Agency, General Electric Company, Santa Barbara, 1977.

Coffey, Helen E., editor, Solar-Geophysical Data - Part I (Prompt Reports), No. 447, Environmental Data and Information Service, Boulder, 1981.

Davies, Kenneth, "Ionospheric Radio Propagation", National Bureau of Standards Monograph 80, U.S. Government Printing Office, Washington, 1965.

_____, "Recent Progress in Satellite Radio Beacon Studies with Particular Emphasis on the ATS-6 Radio Beacon Experiment", Space Science Reviews, Vol. 25, 1980.

Dodson, H. W. and E. R. Hedeman, "An Experimental Comprehensive Flare Index and Its Derivation for "Major" Flares, 1955-1969", Report UAG-14, World Data Center A for Solar-Terrestrial Physics, NOAA/EDIS, Boulder 1971.

Eddy, J. A., "The Sun Since the Bronze Age", Proceedings of the International Symposium on Solar Terrestrial Physics, Boulder, 1976.

Eddy, J. A., P. A. Gilman, D. E. Trotter, "Anomalous Solar Rotation in the Early 17th Century", Science, Vol. 198, 1977.

Eis, K. E., J. A. Klobuchar, and C. Malik, "On the Installation, Operation, Data Reduction, and Maintenance of VHF Electronic Polarimeters for Total Electron Content Measurements", Air Force Geophysics Laboratory Instrumentation Papers, No. 256, Hanscom AFB, 1977.

Eis, K. E. and R. C. Richard, "An Observer's Manual for the Air Force Swept Frequency Interferometric Radiometer", Air Force Geophysics Laboratory Instrumentation Papers, No. 263, Hanscom AFB, 1978.

Forbush, S. E. and D. Venkatesan, "Diurnal Variation in Cosmic Ray Intensity, 1937-1959, at Cheltenham (Fredricksburg), Huancayo, and Christchurch", Journal of Geophysical Research, Vol. 7, 1960.

Gibson, Edward G., The Quiet Sun, National Aeronautics and Space Administration SP-303, Washington, 1973.

Glasstone, S., editor, The Effects of Nuclear Weapons, U.S Government Printing Office, Washington, 1962.

Glasstone, S. and P. Dolan, The Effects of Nuclear Weapons, Third Edition, U.S. Government Printing Office, Washington, 1977.

Hardy, D. A., W. J. Burke, M. S. Gussenhoven, N. Heinemann; and E. Holeman, "DMSP/F2 Electron Observations of Equatorward Auroral Boundaries and Their Relationship to the Solar Wind Velocity and the North-South Component of the Interplanetary Magnetic Field", Journal of Geophysical Research, Vol. 86, 1981.

Hawkins, Gerald S., "Ionospheric Electron Content and Radio Scintillation During Magnetospherically Quiet Periods in 1970-71", Air Force Cambridge Research Laboratory Technical Report 74-0160, L.G. Hanscom Field, 1974.

Herman, John R. and Richard A. Goldberg, Sun, Weather, and Climate, National Aeronautics and Space Administration SP-426, Washington, 1978.

Hess, W. N., Radiation Belt and Magnetosphere, Ginn and Co., 1968.

Hess, W. N. and G. D. Mead, editors, Introduction to Space Science, Gordon and Breach Science Publishers, Inc., New York, 1968.

Hines, C. O., "Cause-Effect Inferences in Geophysical Statistical Studies", Proceedings of the International Symposium for Solar-Terrestrial Physics, Boulder, 1976.

Howard, R. and B. J. LaBonte, The Astrogeophysical Journal Letters, 1 July 1980, and report thereof in Mercury, Vol. 10, 1981.

Hundhausen, A. J., "Interplanetary Shock Waves and the Structure of Solar Wind Disturbances", Solar Wind, NASA SP-308, Washington, 1972.

Jacchia, Luigi G., "The Earth's Upper Atmosphere-I", Sky and Telescope, Vol. 49 1975.

_____, "Atmospheric Structure and its Variations at Heights above 200km", COSPAR International Reference Atmosphere 1965, North-Holland Publishing Company, Amsterdam, 1965.

_____, "The Earth's Upper Atmosphere-II", Sky and Telescope, Vol. 49, 1975.

Jordan, Stuart, editor, The Sun as a Star, National Aeronautics and Space Administration SP-450, National Technical Information Service, Springfield, 1981.

Joselyn, J. A. and P. S. McIntosh, "Disappearing Solar Filaments: A Useful Predictor of Gemagnetic Activity", Journal of Geophysical Research, Vol. 86, 1981.

Kildahl, Karl J.N., "Frequency of Class M and X Flares by Sunspot Class (1969-1976)", Solar-Terrestrial Predictions Proceedings, Vol. 3, Space Environment Laboratory, Boulder, 1980.

Kokubun, Susumu and Robert L. McPherron, "Substorm Signatures at Synchronous Altitude", Journal of Geophysical Research, Vol. 86, 1981.

Kundu, Mukul R., Solar Radio Astronomy, Interscience Publishers, New York, 1965.

Langsford, H., "Tree-Rings: Predictions of Drought?", Weatherwise, Vol. 32, 1979.

Leighton, R. B., "Transport of Magnetic Fields on the Sun", Astrogeophysical Journal, Vol. 140, 1964.

_____, "A Magneto-Kinematic Model of the Solar Cycle," Astrogeophysical Journal, Vol. 156, 1969.

Lundstedt, H., J. M. Wilcox, P. H. Scherrer, "Solar Acceleration of Solar Wind: Influence of Active Region Magnetic Field", Institute for Plasma Research Preprint, Stanford, 1981.

Manley, James A., "Short Term HF Forecasting and Analysis", Air Force Global, Weather Central Technical Note 81/001, Offutt AFB, 1981.

Matsushita, S., "A Study of the Morphology of Ionospheric Storms", Journal of Geophysical Research, Vol. 64, 1959.

MITRE Corporation, Space Environment Monitoring Final Report, ESD-TR-72-20, Bedford, 1972.

McIntosh, Patrick S., "The Birth and Evolution of Sunspots: Observations", The Physics of Sunspots, Sacramento Peak Observatory, Sunspot, 1981.

Mendillo, Michael, Michael J. Buonsanto, and John A. Klobuchar, "The Construction and Use of Storm Time Corrections for Ionospheric F-Region Parameters", Effect of the Ionosphere on Space Systems and Communications, Naval Research Laboratory, Washington, 1975.

Muldrew, D. B., "F Layer Inization Troughs Deduced from Alouette Data ", Journal of Geophysical Research, Vol. 70, 1965.

National Research Council, Solar-Terrestrial Research for the 1980's, National Academy Press, 1981.

Nuclear Regulatory Commission," Radiation Dose Levels for Apollo Crew Members, File Memo FA 2-10-67", Publication 1487, 1967.

Olson, R. H., W. O. Roberts, C. S. Zerefors, "Short Term Relationships Between Solar Flares, Geomagnetic Storms and Tropospheric Vorticity Patterns", Nature, Vol. 257, 1975.

Opp, Albert G., "Penetration of the Magnetopause beyond 15.6 Re during the Magnetic Storm of January 13-14, 1967: Introduction", Journal of Geophysical Research, Vol. 73, 1968.

Pasachoff, Jay M., Contemporary Astronomy, W. B. Saunders Co., Philadelphia, 1977.

Prochaska, Robert D., "Geomagnetic Index Calculation and Use at AFGWC", Air Force Global Weather Center Technical Note 80/002, Offutt AFB, 1980.

Prolss, G. W., "Magnetic Storm Associated Perturbations of the Upper Atmosphere: Recent Results Obtained by Satellite-Borne Gas Analyzers", Reviews of Geophysics and Space Physics, Vol. 18, 1980.

_____, "Latitudinal Structure and Extension of the Polar Atmospheric Disturbance", Journal of Geophysical Research, Vol. 86, 1981.

Raloff, Janet, "EMP: A sleeping Electronic Dragon", Science News, vol. 19, 1981.

_____, "EMP: Defensive Strategies", Science News, Vol. 19, 1981.

Rastogi, R. G., J. P. Mullen, and E. Mackenzie, "Effect of Geomagnetic Activity on Equatorial Radio VHF Scintillation and Spread F", Journal of Geophysical Research, Vol. 86, 1981.

Ratcliffe, J. A., An Introduction to the Ionosphere and Magnetosphere, Cambridge University Press, Cambridge, 1972.

Rishbeth, Henry and Owen K. Garriott, Introduction to Ionospheric Physics, International Geophysics Series, Vol. 14, Academic Press, New York, 1969.

Roberts, William M. and Rayner K. Rosich, editors, "Ionospheric Predictions", Telecommunications Research and Engineering Report 13, U.S. Government Printing Office, Washington, 1971.

Starkey, Roland J., Jr., "The Renaissance in Submarine Communications Part III: The ELF Odyssey 1958-1979", Military Electronics/Countermeasures, January 1981.

Stassinopoulos, E. G., "The Geostationary Environment", Journal of Spacecraft and Rockets, Vol. 17, 1980.

Stoupel, E., "Solar-Terrestrial Prediction: Aspects for Preventive Medicine", Solar-Terrestrial Predictions Proceedings, Vol. 4, Space Environment Laboratory, Boulder, 1980.

Svalgaard, L., "Evidence For Sun-Weather Relations", Proceedings of the International Symposium on Solar-Terrestrial Physics, Boulder, 1976.

Thomas, Barry T. and Edward J. Smith, "The Structure and Dynamics of the Heliospheric Current Sheet", Journal of Geophysical Research, Vol. 86, 1981.

U.S. Naval Observatory, The Astronomical Almanac 1982, U.S. Government Printing Office, Washington, 1981.

Valley, Shea L., editor, Handbook of Geophysics and Space Environments, McGraw-Hill Book Company, Inc., New York, 1965.

Vastenhoud, J., "Bandsaving Techniques in Broadcasting", World Radio TV Handbook, Vol. 34, Billboard Limited, Hvidovre, 1980.

Verschuur, G. L., "Dust Clouds and Ice Ages", Astronomy, 1976.

Wand, R.H. and J. V. Evans, "Morphology of Ionospheric Scintillation in the Auroral Zone", Effect of the Ionosphere on Space Systems and Communications, Naval Research Laboratory, Washington, 1975.

Watts, J. W., Jr. and J. J. Wright, "Charged Particle Radiation Environment for Spacelab and Other Missions in Low Earth Orbit, Revision A.", NASA Technical Memo TMX-73358, National Aeronautics and Space Administration, Washington, 1976.

White, M. R., "Environmental Assessment for the Satellite Power System Concept Development and Evaluation Program", Department of Energy Evaluation Report 0089, Department of Energy, Washington, 1980.

Wilcox, J. M., "History of Solar-Terrestrial Relations as Deduced from Spacecraft and Geomagnetic Data", Proceedings of the International Symposium on Solar-Terrestrial Physics, Boulder, 1976.

Wilcox, J. M., P. H. Scherrer, L. Svalgaard, W. O. Roberts, and R. H. Olson, "Solar Magnetic Sector Structure: Relation to Circulation of the Earth's Atmosphere", Science, Vol. 180, 1973.

Wilcox, J. M., L. Svalgaard, and P. H. Scherrer, "On the Reality of a Sun-Weather Effect", Journal of Atmospheric Science, Vol. 33, 1976.

Wilcox, J. M. and N. F. Ness, "Quasi-Stationary Corotating Structure in the Interplanetary Medium", Journal of Geophysical Research, Vol. 70, 1965.

Williams, R. G. and E. J. Geretz, "Does the Troposphere Respond to Day-to-Day Changes in the Solar Magnetic Field?", Nature, Vol. 275, 1978.

Wolfe, John H., "The Large-Scale Structure of the Solar Wind", Solar Wind, NASA SP-308, Washington, 1972.

Wright, J. W., "Dependence of the Ionospheric F-Region on the Solar Cycle", Nature, Vol. 194, 1962.

## AWS DISTRIBUTION

DISTRIBUTION:

1WW (3)
2WW (10)
3WW (1)
5WW (7)
7WW (3)
AWS/DN (1)
2WS (1)
AFGWC (4)
USAFETAC/DN (1)
USAFETAC/OL-A (1)
USAFETAC/TSK (5)
AUL (1)
3350 TCHTG/TTMV (5)

★ U.S. GOVERNMENT PRINTING OFFICE: 1983-664-000/6006