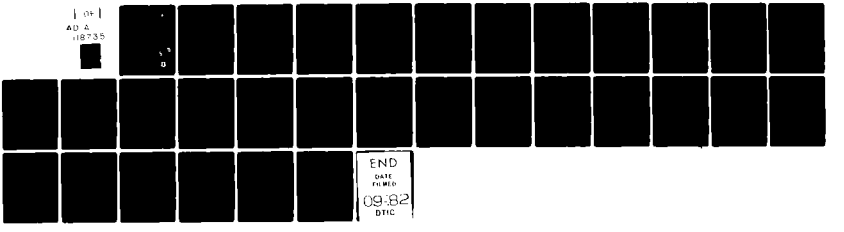
AD-A118 735 AIR FORCE GEOPHYSICS LAB HANSCOM AFB MA F/G 4/2
THE B-G SYSTEM OF EVALUATING FORECASTS.(U)
JAN 82 I I GRINGORTEN, A R BOEHM
UNCLASSIFIED AFGL-TR-82-0006 NL



(12)

AFGL-TR-82-0006
ENVIRONMENTAL RESEARCH PAPERS, NO. 761

The B-G System of Evaluating Forecasts

IRVING I. GRINGORTEN
ALBERT R. BOEHM, Maj, USAF

4 January 1982

METEOROLOGY DIVISION PROJECT 6670
AIR FORCE GEOPHYSICS LABORATORY
HANSCOM AFB, MASSACHUSETTS 01731

AIR FORCE SYSTEMS COMMAND, USAF

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER AFGL-TR-82-0006	2. GOVT ACCESSION NO. AD-A118735	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle) THE B-G SYSTEM OF EVALUATING FORECASTS		5 TYPE OF REPORT & PERIOD COVERED Scientific. Interim
		6. PERFORMING ORG. REPORT NUMBER ERP No. 761
7 AUTHOR(s) Irving I. Gringorten Albert R. Boehm, Major, USAF*		8 CONTRACT OR GRANT NUMBER(s)
9 PERFORMING ORGANIZATION NAME AND ADDRESS Air Force Geophysics Laboratory (LYT) Hanscom AFB Massachusetts 01731		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62101F 66700908
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Geophysics Laboratory (LYT) Hanscom AFB Massachusetts 01731		12. REPORT DATE 4 January 1982
		13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

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)

18. SUPPLEMENTARY NOTES

*Affiliation - USAF-ETAC
Scott AFB, IL 62225

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

Forecast verification
Forecast scoring
Forecast evaluation

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

To score a forecast of an event, like the noontime temperature at a city airport, a system has been proposed that would measure the improvement of the forecast over simple climatic information. If the climatic probability of the forecast (F) is P_F, and that of the observed event (V) is P_V, then the assigned score (s_{FV}) becomes $s_{FV} = -\ell n(1 - P_1)P_2 - 1$, where $P_1 = \min(P_F, P_V)$ and $P_2 = \max(P_F, P_V)$. A test of the forecasters' skill, however, requires a large set of forecasts and corresponding verifications.

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (Continued)

It has been found desirable to evaluate additionally each score by the likelihood that it could have been achieved by chance. The likelihood of a chance score (LCS) has been determined, based on the climatic probabilities (P_F, P_V). It reduces to zero for a perfect forecast; it would be expectedly equal to 1/2 in pure chance or guesswork, and would be 1.0 for a completely erroneous forecast. The excess number (n) of forecasts, whose LCS's are less than expected by chance, becomes the indicator of forecast skill. Chi-square becomes a readily available criterion of the significance of the number (n) as opposed to the expected number by chance.

Preface

Forecast scoring has had a history of frustration, as well as a measure of success, with little universal agreement on how it should be achieved. Consequently, interest in the subject has increased and decreased intermittently over the years. In April 1979, a special panel of the Committee on Atmospheric Sciences of the National Research Council (NRC) wrote a report on "Long-Range Weather Forecast Evaluation," which contained a recommended scoring system for the forecasts of weekly average temperature at several locations in the United States, forecast one month in advance. The NRC report became a starting point for a renewed effort on scoring methods, both at United States Air Force Environmental Technical Applications Center (USAFETAC) and Air Force Geophysics Laboratory (AFGL). The authors of this report, representing the last two organizations, offer a procedure that shows promise on extended forecasting. The prior interest in the subject by Col Gary Atkinson, AWS, and the several earlier investigations by Lt Col Gerald J. Dittberner and others at ETAC are acknowledged both for their inspiration and their guidance.

Contents

Illustrations

Tables

Tables

The B-G System of Evaluating Forecasts

1. INTRODUCTION

The effectiveness of weather forecasting, 12 hours to several days in advance, is well-established. But how much longer in advance are forecasts worthwhile? How shall we decide which of several competing predictions deserves the highest rating? Indeed, how should forecasts be judged? By accuracy, by usefulness or by something else that we might call the skill or expertise of the forecaster? Should forecasters be allowed to hedge, by stating probabilities or otherwise? What constitutes a verification? How are errors in the forecast measured?

The whole subject of scoring and evaluating forecasts is riddled with problems, complications, doubts, and skepticism. Usually efforts to devise a scheme of scoring, objective or otherwise, have suffered rejection, sometimes by the most professional of meteorologists. At the same time the need for such evaluation has been inescapable and will not fade away.[1] Moreover, there has been no dearth of schemes, new[2] or old.[3] In particular, an acceptable objective system

(Received for publication 30 December 1981)

1. Nap, J. L., Van den Dool, H. M., Oerlemens, J. (1981) A verification of monthly weather forecasts in the seventies, Monthly Weather Review 109: 306-312.
2. Gulezian, Dean P. (1981) A new verification score for public forecasts, Monthly Weather Review 109:313-323.
3. Gringorten, Irving I. (1965) A measure of skill in forecasting a continuous variable, J. Appl. Meteorol. 4:47-53.

has been sought for long-range forecasting in order to answer the question about the very existence of effective expertise in the forecasts (see Appendix A).

Forecasters might be rated for their accuracy, usefulness or professional skill. The degree of accuracy, by and large, is secondary to the utility of the forecasts. But measure of utility, in turn, is elusive. It would require a knowledge of customers costs and or profits, which are much too changeable, to say nothing about differing and conflicting interests. We are left with the goal to determine, or uncover, the professional skill in a set of forecasts. The B-G system presented here is limited to this specific objective.

Basically, climatological information such as the climatic frequency of cloud cover from clear to overcast should be available to a forecaster. In the case of temperature, say at Minneapolis, noontime in January, the climatic information will include the frequency distribution of the temperatures, ranging in this case from -36°C with 1/100 of 1 percent probability through the median of -11°C up to 10°C with 1/100 of 1 percent probability of exceedance.

For the purpose of this paper, the skill of the forecaster is defined as his ability to recognize and to quantify the probability of departure of a future event from the normal climatic frequency of the event. If, in the case of Minneapolis temperature, he sees a strong probability of the later temperature -15°C, and so forecasts, then a subsequent verification of -15°C should earn him a maximum positive score, and somewhat less if subsequently the temperature is -16°C or -14°C.

In recent years, much has been said, as well as done, about probability forecasting. This has necessitated another type of skill—the ability to state valid probabilities. A line of demarcation should be drawn between this skill and the ability of the forecaster to sharpen the prediction toward a higher degree of certainty of one outcome. If, say, there is a 30 percent climatic probability of rain, then the quotation of 50 percent probability of rain is to be considered as a sharpening of the odds on rain. There could be some kind of verification of the 50 percent quotation, but it is not the intended goal of this paper.

This paper seeks to evaluate the total of forecast statements. Such evaluation is especially desirable for long-range forecasts, seasonal or annual, since the verification of probability statements would be meaningless. If the weather element to be predicted is the seasonal winter average temperature, predicted in the previous fall season, the forecaster is called upon to give his best single estimate of the subsequent winter seasonal average temperature. If he chooses -8°C as his forecast at Minneapolis, he will imply a warm winter, since the median winter temperature is -11°C.

For this scoring system, there is a mandatory premise that distinguishes it from most, if not all other, scoring systems: The unskilled forecaster should

not benefit from any unskilled strategy when making his forecast. No quantity should be forecast with greater expectation of a score than any other quantity unless it be done for meteorological reasons, extended beyond the simple knowledge of climatic frequencies. There must be no long-term advantage in an unskilled choice. For example, if the unskilled forecaster predicts no-rain (NR), simply because it is more frequent, he should earn the same average score, in 100 such forecasts, as he would if he predicted rain (R). Symbolically, if the climatic frequencies are P(NR) and P(R) for no-rain and rain, respectively, then the scores, S(NR) and S(R) should satisfy

$$P(NR)\,S(NR) = P(R)\,S(R) = 1 \qquad (1)$$

A skilled forecaster, on the other hand, must pursue the analysis of the synoptic weather situation until he detects a trend or a better-than-usual probability of one future event. Persistence, as a strategy, is examined in Section 5.

The B-G system, presented herein, provides a score for each individual forecast and an evaluation of a sufficiently large collection of forecasts.

2. SCORE

If the weather element to be forecast consists of two alternatives, A and not-A, with climatic frequencies p and (1-p), respectively, then in accordance with the above premise, a correct forecast should earn the score s_A or $s_{not\text{-}A}$ given by

$$s_A = 1/p$$
$$s_{not\text{-}A} = 1/(1-p) \qquad (2)$$

Suppose, on the other hand, that the weather element is a variable (X) whose values range continuously from low to high values with cumulative probability $P(X \leq x)$ symbolized as p. Let the predicted event (F) have climatic cumulative probability P_F, and the verifiably observed event (V) have climatic cumulative probability P_V. If the variable (X) is divided dichotomously at x where the climatic probability is p, then: for p less than both P_F and P_V, F is a correct forecast of not-X, and the score is

$$s_p = 1/(1-p)$$

For p greater than one of (P_F or P_V), but less than the other, F is an incorrect forecast and the score is

$$s_p = 0$$

For p greater than both P_F and P_V, F is a correct forecast of X, and the score is

$$s_p = 1/p$$

For all dichotomous divisions the score average (s'), when $P_V < P_F$, is given by

$$s' = \int_0^{P_V} dp/(1-p) + \int_{P_F}^{1} dp/p$$

That is,

$$s' = -\ln\{(1 - P_V)P_F\} \quad \text{for} \quad P_V < P_F$$

Similarly,

$$s' = -\ln\{(1 - P_F)P_V\} \quad \text{for} \quad P_V \geq P_F$$

For an unskilled forecast the expected value of s' is 1.0. A score of zero, however, is preferred for no skill. Therefore, the score (s_{FV}) for the forecast (F) and verification (V) is chosen to be:

$$\begin{aligned} s_{FV} &= -\ln\{(1 - P_V)P_F\} - 1 \quad \text{for} \quad P_V < P_F \\ &= -\ln\{(1 - P_F)P_V\} - 1 \quad \text{for} \quad P_V \geq P_F \end{aligned} \tag{3}$$

To illustrate the scores (s_{FV}), consider the following: In Figure 1 it is assumed that the median has been forecast, (that is, $P_F = 0.5$). The score is maximum when the forecast is exactly correct; it is still positive when, for verification, $0.27 \leq P_V \leq 0.73$; it is negative otherwise and lowest when either the lowest or highest extreme verifies. In Figure 2, it is assumed that the lower quartile has been forecast ($P_F = 0.25$). The score again is maximum for an exactly correct verification. In Figure 3, it is assumed that the forecast calls

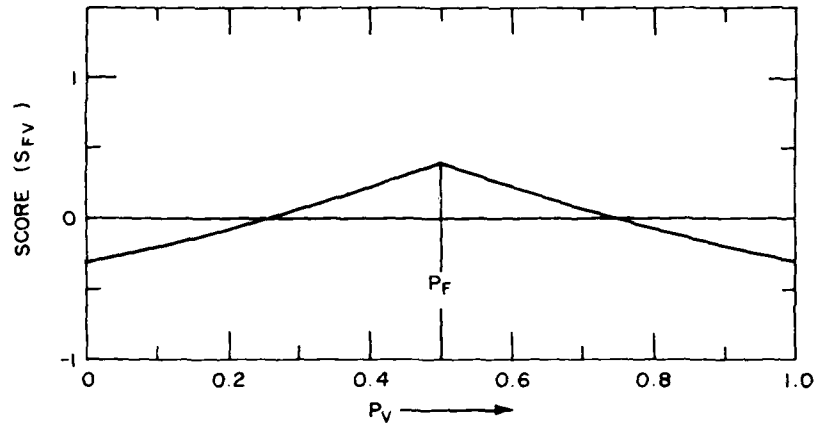


Figure 1. The Score (s_{FV}) Plotted Against the Probability (P_V) of the Observed Event (V) When the Forecast (F) is the Median ($P_F = 0.5$)

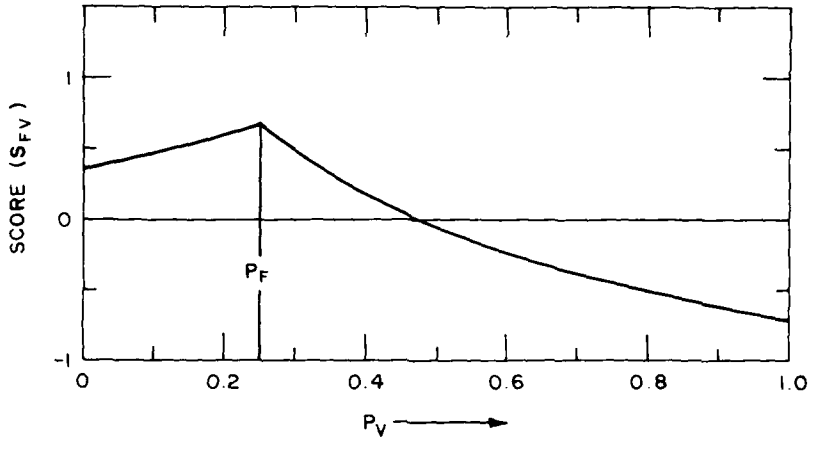


Figure 2. The Score (s_{FV}) Plotted Against the Probability (P_V) of the Observed Event (V) When the Forecast (F) is the Lower Quartile ($P_F = 0.25$)

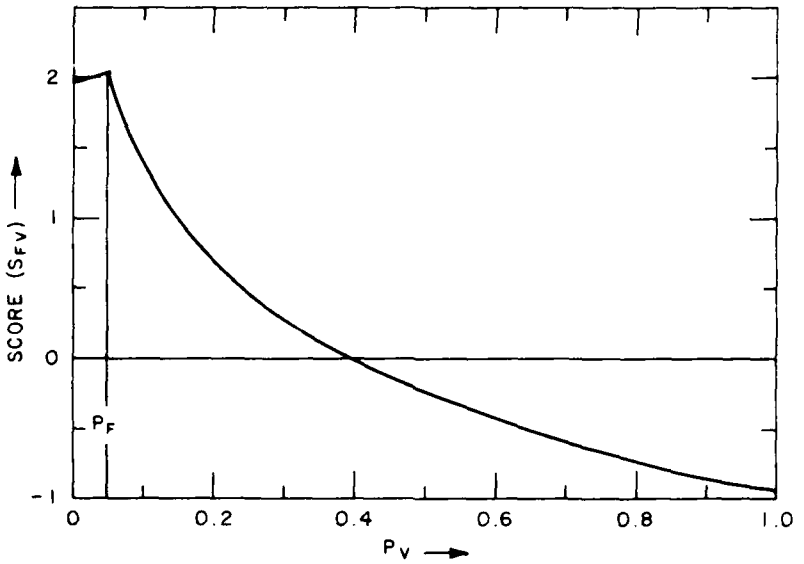


Figure 3. The Score (s_{FV}) Plotted Against the Probability (P_V) of the Observed Event (V) When the Forecast (F) is the Five Percentile (P_F = 0.05)

for an extreme low condition to develop. If correct, or nearly correct, then the score is large; otherwise the score drops rapidly to negative values with increasing error. In all cases, however, the average, or expected, no-skill score is zero.

Figure 4 shows the scores (s), from the lowest possible to the perfect versus the forecast (represented by P_F). When the forecaster predicts the median (P_F = 0.5), he will earn a positive score when the median event verifies, and lose if some unusual or extreme event verifies, although his gain or loss will be modest. On the other hand, if the forecaster predicts an unusual event, low or high, he can earn an unusually high score (s), while risking a greater negative score. Whether he forecasts the median or an extreme event, however, if he is only guessing, or uses an unskilled system of forecasting, the expected score is 0.0. At all times the perfect score is achieved when $P_V = P_F$. The average of the perfect scores is 1.0.

The foregoing scores [Eq. (3)] can be used when the predictand weather element varies continuously, theoretically from minus infinity to infinity, with cumulative probabilities (P_F, P_V) well defined. However, when we must begin with a

category, such as a substantial probability of no rain, or clear sky (0/10), or end with a substantial probability of overcast (10/10), we must modify the scoring system, while adhering to the premise that unskilled forecasting must not benefit from a no-skill strategy (Appendix B).

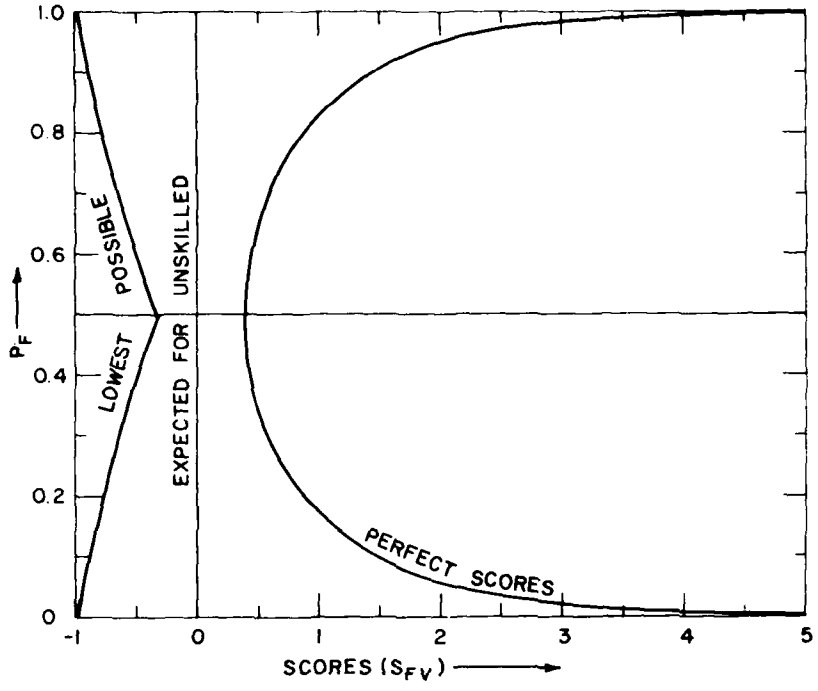


Figure 4. The Range of Scores (s) From the Lowest Possible to the Perfect Score, as a Function of the Forecast (P_F)

3. EVALUATION

In the preceding section, each pair of forecast and verification (P_F, P_V) results in a score (S_{FV}). Evaluation of a whole forecast system, however, must depend on a large set (N) of forecasts and verifications. Among N forecasts, there will be high and low scores. Since the actual score depends on the forecast itself, as well as on its excellence, it is desirable to examine the score of the forecast by its place in the set of possible scores (Figures 1, 2, or 3). The forecaster whose scores are consistently among the upper 10 percent is clearly better than the forecaster who can claim only that his scores are among the upper 25 percent.

Figure 5 shows the scores for the whole range of verifications ($0 \leq P_V \leq 1$) for the forecast ($P_F = 0.35$). For the verification (P_{V2}), a score (s_o) by pure chance can exceed the earned score (S_{FV2}) with frequency P_{V2}. If the verification is closer to the forecast (P_F), then the score (s_{FV}) is not exceeded quite as often by chance. For the same score (s_{FV}) there can be two verifications:

If $P_V > P_F$ then, for the same score, $P_{V\ell} = 1 - (1 - P_F) P_V/P_F$

If $P_V \leq P_F$ then, for the same score, $P_{Vu} = (1 - P_V) P_F/(1 - P_F)$

As seen from the diagram, the score s_{FV} can be exceeded by chance with probability $(P_V - P_{V\ell})$ or $(P_{Vu} - P_V)$.

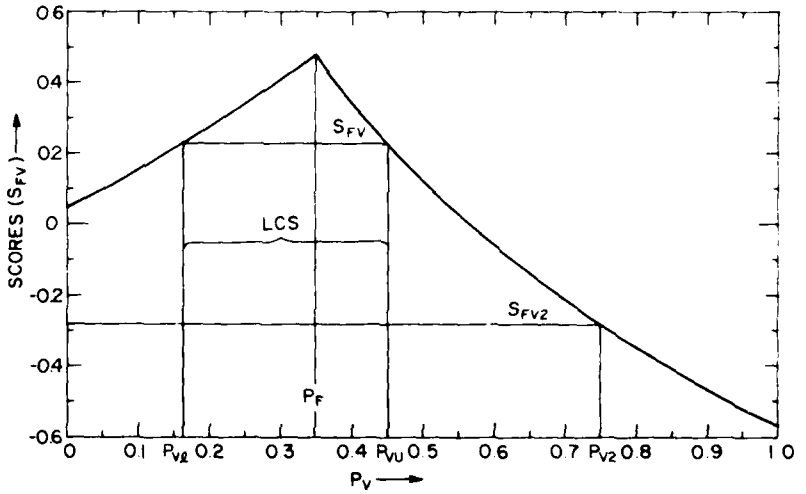


Figure 5. Illustrating the Distribution of Scores (s_{FV}) as a Function of Observed Event (P_V), and the Likelihood (LCS) of a Chance Score (s_o) to Exceed s_{FV} When the Forecast is P_F

In summary, we find the probability, $P(s_o \geq s_{FV})$, that a chance score (s_o) can equal or exceed the achieved score (s_{FV}), as follows (Figure 5), defining LCS as LCS = $P(S_o \geq S_{FV})$:

$$
\begin{aligned}
\text{LCS} &= P_V && \text{when } P_F < P_V \ , \ P_V \geq P_F/(1 - P_F) \\
&= (P_V - P_F)/P_F && \text{when } P_F < P_V \ , \ P_V < P_F/(1 - P_F) \\
&= (P_F - P_V)/(1 - P_F) && \text{when } P_F \geq P_V \ , \ P_V \geq 2 - 1/P_F \\
&= (1 - P_V) && \text{when } P_F \geq P_V \ , \ P_V < 2 - 1/P_F \ .
\end{aligned} \tag{6}
$$

The likelihood of a chance score (LCS), will be less for a more successful forecast. For an accurate forecast it would be identically zero; for a "complete bust" it would be 1.0. For N unskilled forecasts, we expect (N/10) forecasts to achieve scores for which LCS $\leq$ 0.1, (2N/10) forecasts with scores for which LCS $\leq$ 0.2, and so on. In general, we would expect (i N/10) unskilled forecasts, scores for which LCS $\leq$ i/10. In Figure 6 the number of forecasts (n) is plotted against LCS (P). The diagonal straight line gives the expected numbers by chance. For example, among 100 unskilled forecasts we expect 40 to have scores high enough to give LCS $\leq$ 0.4. In a test (Figure 6) in which 100 random numbers were used for the verification (P_V) against 100 random forecasts (P_F), the number of scores in the upper 10 percent were 9 (instead of 10); in the upper 20 percent there were 17 (instead of 20), and so on, as shown by the solid curve (Figure 6). The chi-square test (see Section 4) showed no significant differences between the solid curve and the diagonal straight line. In two other tests the forecasts were: (1) consistently the median; (2) consistently the lowest 1-percentile, with the results shown by the broken lines. Again the chi-square test did not reveal a significant difference.

In contrast with the tests illustrated in Figure 6, those in Figure 7 were performed with 100 pairs of forecasts and verifications when there were significant correlation coefficient (ρ) between them, simulating skillful forecasting. When ρ = 0.5, there were 18 forecasts, out of 100, whose scores were among the upper 10 percent; when ρ = 0.95, there were 49 such successful forecasts; when ρ = 0.99, there were as many as 71 forecasts that succeeded in achieving scores in the rare 10 percent bracket.

An evaluation of the set of N forecasts is accomplished by finding the 9 numbers (n_P) for P = 0.1(0.1)0.9, knowing that they must exceed the numbers (NP) expected by chance.

Alternatively, or as a supplementary evaluation, the average value of the probability, $\overline{\text{LCS}}$, will be useful especially in the comparison of two sets of forecasts. A proposed evaluation (E) of the set, ranging from -1.0 for worse-than-useless to 1.0 for perfect, is:

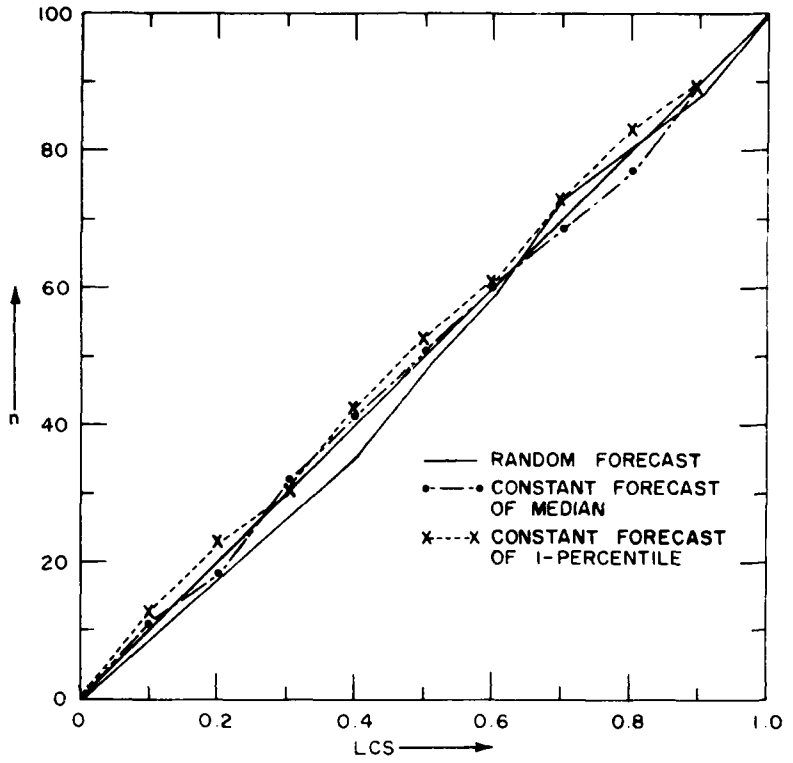


Figure 6. The Number (n) of "Forecasts" (out of 100) Whose Scores had Probability (LCS) of Being Equaled or Exceeded by Chance. All "forecasts" were made by unskilled procedures

$$E = 1 - 2(\overline{LCS}) \tag{7}$$

Zero for E would imply no skill. Consider the following example: For 75 years (1902 to 1976), at the University of Wisconsin, forecasts of the average temperature in the winter season (December, January, and February) were made for some 42 stations on or before 30 November of each year. The results for Winnipeg, Manitoba (Figure 8) show that the forecasts do demonstrate skill, or have informational value but qualitatively, since there was not a significant number of forecasts verifying at less than the 30-percentile. The results for Chicago, Illinois (Figure 9) are very similar, except that, without the supporting evidence of the Winnipeg performance, these results would have been viewed as indecisive, requiring further sampling to establish conclusively the forecasting skill (see Section 4).

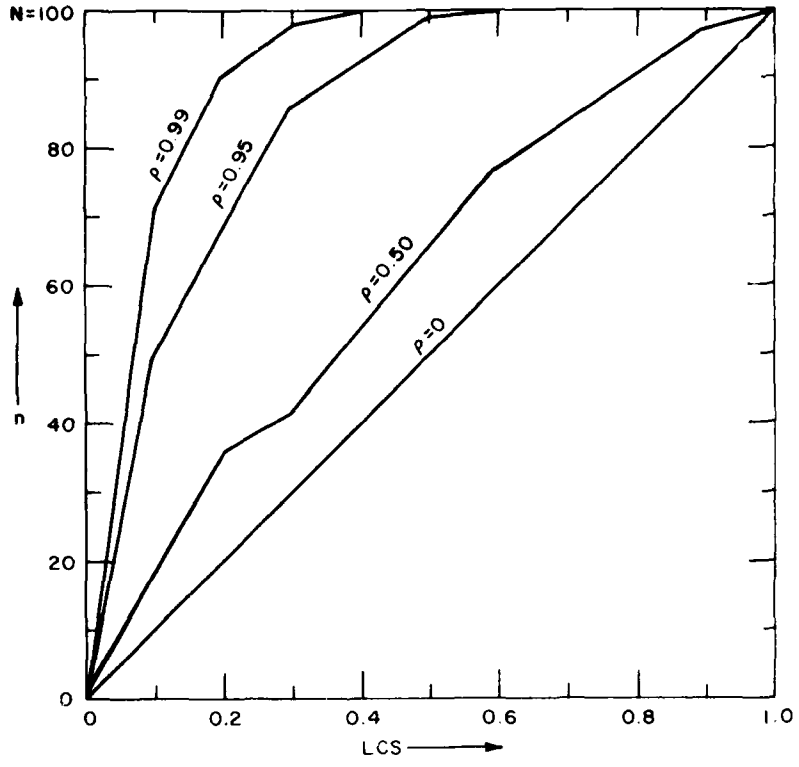


Figure 7. The Number (n) of "Forecasts" (out of 100) Whose Scores had Probability (LCS) of Being Equaled or Exceeded by Chance. The correlation coefficients (ρ) between "forecasts" and "observed" events were as shown

Again the problem of some categories of weather, such as no-rain, must be faced. Modifications of Eq. (6) were obtained on the assumption that there is a category (X) with climatic frequency ($P_{xu} - P_{xl}$, with much attention to details (Appendix B).

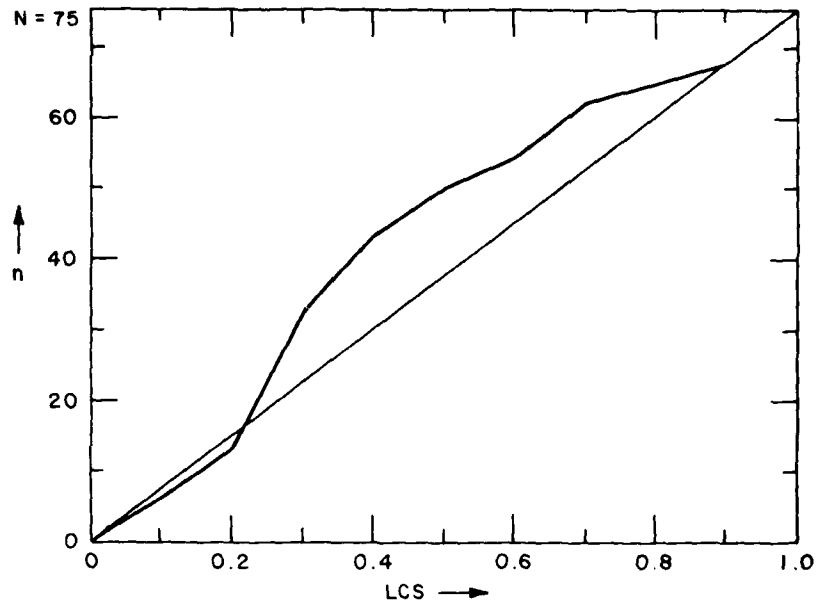


Figure 8. The Number (n) of Forecasts [out of 75 Forecasts (1903-1978)] Whose Scores had Probability (LCS) of Being Equaled or Exceeded by Chance. The example is for the average winter temperature at Winnipeg, Manitoba

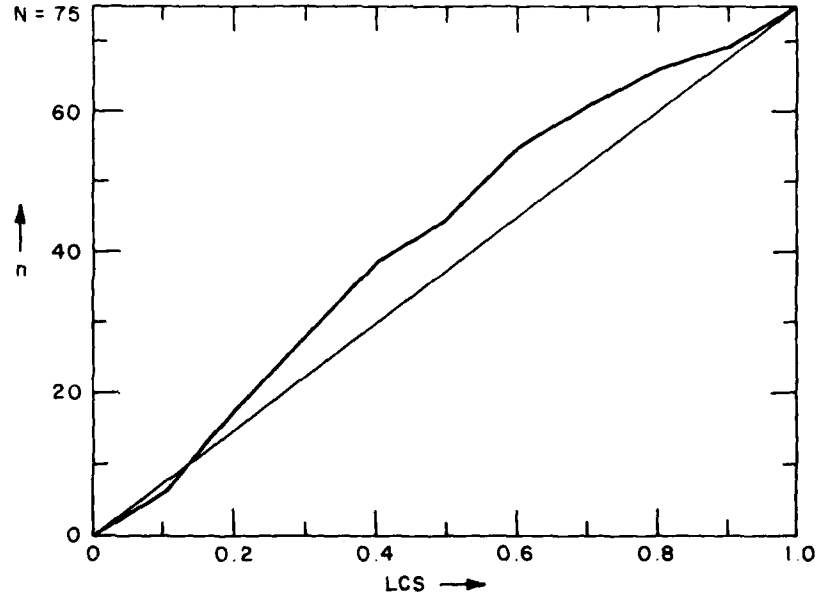


Figure 9. The Number (n) of Forecasts [out of 75 Forecasts (1903-1978)] Whose Scores had Probability (LCS) of Being Equaled or Exceeded by Chance. The example is for the average winter temperature at Chicago, Illinois

4. CHI-SQUARE TEST OF SIGNIFICANCE

Where N is the over-all number of forecasts, n_i is the number of forecasts (i = 0(1)9) whose values of LCS lie between i/10 and (i + 1)/10, then chi-square with 9 degrees of freedom is

$$X_9^2 = \sum_{i=o}^{9} (N/10 - n_i)^2/(N/10)$$

If the forecasts are divided dichotomously at $\sum_{j=o}^{i} n_j$, then

$${}_iX_1^2 = \left(NP_i - \sum_{j=o}^{i} n_j\right)^2 \Big/ \{P_i(1 - P_i)N\}$$

where

$$P_i = 0.1\,(1 + i) \quad , \quad i = 0(1)8$$

The 5 percent values of chi-square, as given in textbook tables, are

$$X_9^2(0.05) = 16.9$$

$$X_1^2(0.05) = 3.8$$

For the numbers (n_i) obtained in the no-skill tests (Figure 6) the chi-square values were obtained as shown (Table 1). For the numbers (n_i) obtained when there was a significant correlation (ρ) between forecast and observed events (Figure 7) chi-square (Table 2) increased, as expected, with correlation coefficient. The significance of chi-square also improved with sample size (Table 3). Clearly, for smaller ρ it will take larger samples to establish significant skill.

The foregoing test is one of several alternatives, and its choice here is an arbitrary one. Other tests apply to s_{FV} as opposed to LCS. Future work might suggest another choice for test of significance.

Table 1. Chi-Square Values With One ($_iX_1^2$) and Nine (X_9^2) Degrees of Freedom in Tests of Three Kinds of Unskilled Forecast Selections. Each sample size was 100. For comparison, under the no-skill null hypothesis, the 5 percent chi-square with one degree of freedom is 3.8; with 9 degrees of freedom it is 16.9

Forecast Selection	X_9^2	$_iX_1^2$								
		i = 0	1	2	3	4	5	6	7	8
Always random	5.0	0.1	0.5	0.8	1.0	0.2	0.0	0.4	0.0	0.4
Always median	3.4	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.6	0.0
Always 1 percentile	4.4	1.0	0.6	0.0	0.4	0.4	0.0	0.4	0.6	0.0

Table 2. Chi-Square Values With One and Nine Degrees of Freedom in Tests in "Forecasts" Whose Correlation Coefficient With the "Observed" Events Varied from 0.05 to 0.99. Each sample size was 100. The broken line divides significant chi-square from nonsignificant

Correlation between F and V	X_9^2	X_1^2								
		i = 0	1	2	3	4	5	6	7	8
ρ = 0.05	4.8	0.0	0.2	0.2	0.2	0.0	0.0	0.8	0.6	0.4
0.25	10.6	1.0	2.2	1.7	0.4	1.4	5.0	3.9	0.6	0.4
0.5	24.2	9.0	16.1	6.9	8.2	11.6	12.0	8.0	6.2	5.4
0.95	218.4	169.0	156.0	149.0	117.0	96.0	66.6	42.9	25.0	11.1
0.99	449.0	413.0	315.0	220.0	150.0	100.0	66.7	42.9	25.0	11.1

Table 3. Test of Chi-Square Values With One and Nine Degrees of Freedom in Tests of "Forecasts" Whose Correlation Coefficient With "Observed" Events is 0.25. The sample size was varied from 50 to 500. The broken line divides significant chi-square from nonsignificant at the 5 percent level

Sample Size	X_9^2	X_1^2								
		0	1	2	3	4	5	6	7	8
50	7.2	0.9	0.5	2.4	0.8	2.0	3.0	2.4	0.5	0.9
100	10.6	1.0	2.2	1.7	0.4	1.4	5.0	3.9	0.6	0.4
150	16.4	6.0	10.7	7.1	2.8	2.7	6.2	3.8	1.0	0.3
200	17.4	6.7	10.1	8.6	3.5	3.4	6.0	4.7	0.5	0.9
250	17.9	7.5	9.0	6.9	4.8	5.2	8.1	6.2	0.6	1.6
300	18.1	7.3	12.0	9.1	5.6	4.8	6.7	6.3	1.0	1.8
400	28.0	7.1	17.0	14.6	11.3	7.8	10.7	13.0	2.6	3.4
500	24.0	7.2	17.1	17.6	18.8	6.7	8.0	8.6	3.2	3.8

5. DISCUSSION

The scoring system of this paper is developed for the continuous variable and for a forecast stated specifically, not for a probability forecast. Errors in forecasting are measured by the likelihood, by nonskilled methods (LCS), of the achieved score. If the forecaster habitually earns scores in the upper 50 percent then we rate him as skillful (Appendix C). But we proceed further, and examine his consistency in earning scores exceeding the upper 40 percent, 30 percent, 20 percent or 10 percent. Scores for nearly perfect forecasting will exceed even the upper 1 percent. If the forecaster consistently earns scores in the upper 10 percent we must rate him exceptionally good.

Other scoring programs gather forecasts and verifications in categories such as the following: below normal, normal, and above normal, which have equal climatic frequencies. They score 1.0 for the correctly forecast category and zero for an incorrect category, and base evaluation on the number of correct forecasts. Such evaluation would be in accord with the premise as stated, but surely we must agonize over the tantalizing feature that a "below normal" prediction would be counted as an error when the observed verification is "low normal." If the weather should verify extremely low, a forecast of "above

normal" would be badly in error, but a score of 0.0 does not show the extent of the error. Conversely, there is never a high reward for a spectacularly good forecast, only a score of 1.0.

Significant persistence in the weather could make the simple unskilled strategy, that is, forecasting the present weather to prevail into the future, result in profitable scores (s_{pV}). To avoid rewarding this nonskilled strategy, the forecaster's performance can be judged against the performance of persistence treated as the standard. The advantage in using simple climatology or other unskilled strategy has already been eliminated. Chi-square can be found to test the significance of the difference between any curves (Figures 6, 7, 8, or 9). The quantity E could be modified:

$$E = 1 - \overline{P(s_o \geq s_{FV})}/\overline{P(s_o \geq s_{pV})} \quad .$$

Utility of the forecasts is not measured directly by this system. It is reasonable, however, to view all operations in a given climate as requiring adjustment to that climate. As a simple example, consider the selection of the clothes that one may wear. Since the skillful forecaster does provide the sign and extent of the departure of the weather from the climatic norm, our measure of that skill becomes, at least indirectly, a measure of the utility of the forecast.

Since the system of evaluation depends heavily upon climatic frequency distributions, prior knowledge of the probabilities (P_F, P_V) of predicted and observed events (F,V) should be accurate. If the climatological information is faulty, then clearly the errors in P_F and P_V would cause errors in the evaluation (Eq. 6). However, as long as such errors are less than 10 percent, the errors in LCS should also be less than 10 percent, which should make the errors in the counts (n_i, $i = 0, 9$) negligible for our purpose.

6. OTHER KEY POINTS

While there is a formula for scoring of forecasts (Eq. 3), the scores need not be calculated for the primary goal of evaluating a set of forecasts. The primary statistic has become LCS = $P(s_o \geq s_{FV})$, given by Eq. (6). The evaluation (E) depends upon the average value of LCS. The test of significance is done with our old friend, in fact everybody's old friend, Chi-square, on the number of forecasts that succeed in pinpointing the future events.

Appendix A

Comments on "Long-Range Weather Forecast Evaluation"
By a Special Panel of The National Research Council,
4 April 1979

A National Research Council (NRC) special panel of the Committee on Atmospheric Sciences has written a report, dated 4 April 1979, on "Long Range Weather Forecast Evaluation," which we (AFGL/LYT) have examined, at the request of the Air Weather Service. The NRC report contains a suggested scoring system for the forecasts of weekly average temperature at several locations in the United States, forecast one month in advance.

The method of scoring is directed at uncovering significant skill in 182 forecasts (F) of weekly average temperature (Y) of 26 alternate weeks of one year at seven widely scattered stations in the United States. To be considered skillful, the forecasts must improve on persistence (G), the weekly average temperature obtained at the time the forecast (F) is due.

In the procedure the three variables (Y, F, G) are standardized, or normalized, into (y, f, g), which should make them all have zero mean (0.0) and unit standard deviation (1.0). However, the forecaster's values (F) may be deliberately biased, and all values (Y, G) are to be found by sampling. The paper, therefore, allows for nonzero means $(\bar{y}, \bar{f}, \bar{g})$ and nonunity variances symbolized as [yy], [ff], [gg]. The covariances are [gy], [gf], and [yf].

The primary statistics for evaluation are the squares of the correlation coefficients (cc). If ρ_{yg} is the cc between y and g, and $\rho_{y;fg}$ is the multiple cc of y on f and g, then the skill score (S) is given by the difference:

$$S = \rho^2_{y;fg} - \rho^2_{yg} \tag{A1}$$

Surprisingly, the NRC paper avoids the term correlation and calls S the "incremental fractional reduction in forecast error variance," given by

$$S = \frac{\{[yg]\,[gf] - [yf]\,[gg]\}^2}{[gg]\,[yy]\,\{[gg]\,[ff] - [gf]^2\}} \tag{A2}$$

This number is always positive, skillful forecasts or otherwise. In its purest form, if $[yy] = [ff] = [gg] = 1$ and if persistence is useless, so that $[yg] = [fg] = 0$, then

$$S = [yf]^2 \tag{A3}$$

which reveals that the score, or statistic for evaluation, is basically the sum total of the products of each pair of forecast (f) and verified event (y). Each term (yf) of 182 such terms contributes to the measure of the skill (S) and therefore is effectively the score for that single forecast. It becomes useful, therefore, to examine a table of such "scores" for the individual forecasts (Table A1).

Table A1. The NRC Scores (f is the standardized deviate of the forecast; y is the standardized deviate of the verified event)

f \ y	-3.0	-2.0	-1.0	0	1.0	2.0	3.0
-3.0	9.00	6.00	3.00	0	-3.00	-6.00	-9.00
-2.0	6.00	4.00	2.00	0	-2.00	-4.00	-6.00
-1.0	3.00	2.00	1.00	0	-1.00	-2.00	-3.00
0	0	0	0	0	0	0	0
1.0	-3.00	-2.00	-1.00	0	1.00	2.00	3.00
2.0	-6.00	-4.00	-2.00	0	2.00	4.00	6.00
3.0	-9.00	-6.00	-3.00	0	3.00	6.00	9.00

After the collection of one year's data and calculation of the number S, two questions are raised: Is S significantly large? Can we conclude that the extended forecasts are truly skillful? The NRC panel proposes to find the probability that S will be equaled or exceeded in a population of values (S_o) for nonskilled forecasts. To find one value for S_o, the 182 values of y and the 182 values of g are entered together with 182 randomly selected values for f in Eq. (A2). Repeated 10,000 times this exercise produces a probability distribution of values for S_o. If S is large enough to lie within the upper 5 percent of the S_o-values, then the forecasts might be considered significantly skillful.

A criticism of the NRC approach is that it can be played. There has been no intentional device incorporated into the system to eliminate the advantages of an unskilled strategy by hedging or otherwise. Let us examine the "scores" in Table A1.

Faced with the need to make a forecast for the following month, the forecaster can examine the potential rewards and penalties in Table A1. If he is completely uncertain, he might be tempted to forecast the median temperature (f = 0), since he will not risk punishment, whatever extreme develops. If he leans to colder weather, he ought not to lean too far, because by choosing "moderate cold" (f = -1) he will gain points substantially if it becomes very cold without his having to forecast "very cold." His reward is not greatest for an accurate forecast.

In contrast, the scores, by the method of this paper, are presented (Table A2) for the same forecasts and verifications as in Table A1.

Table A2. The Gringorten Scores (f is the standardized deviate of the forecast; y is the standardized deviate of the verified event)

f \ y	-3.0	-2.0	-1.0	0	1.0	2.0	3.0
-3.0	5.61	2.78	0.84	-0.31	-0.83	-0.98	-0.9973
-2.0	2.78	2.81	0.86	-0.28	-0.80	-0.95	-0.98
-1.0	0.84	0.86	1.01	-0.13	-0.65	-0.80	-0.83
0	-0.31	-0.28	-0.13	0.39	-0.13	-0.28	0.31
1.0	-0.83	-0.80	-0.65	-0.13	1.01	0.86	0.84
2.0	-0.98	-0.95	-0.80	-0.28	0.86	2.81	2.78
3.0	-0.9973	-0.98	-0.83	-0.31	0.84	2.78	5.61

Appendix B

Scoring and Evaluation With a Category (For Example, No Measurable Rain)

Figure B1 is the same as Figure 2 of the text except that an interval in the distribution between P_{xl} and P_{xu} corresponds to a class of weather (X) whose climatic probability is $(P_{xu} - P_{xl})$. There should be one score for this category $({}_xs_{FV})$, although not necessarily the arithmetic average. It is such that, overall, $E({}_xs_{FV}) = 0$.

For $P_{xl} < P_V \leq P_{xu}$ and for $P_F \geq P_V$,

$$ {}_xs_{FV} = [-\ln\{(1 - P_{xu})P_F\} - 1] + \left[1 + \frac{1 - P_{xl}}{P_{xu} - P_{xl}} \cdot \ln \frac{1 - P_{xu}}{1 - P_{xl}}\right] \qquad \text{(B1)} $$

which corresponds to the probability (P'_V) given by

$$ P'_V = 1 - \left\{e^{-(1+{}_xs_{FV})}\right\}/P_F \qquad \text{(B2)} $$

For $P_F < P_V$,

$$ {}_xs_{FV} = [-\ln\{(1 - P_F)P_{xu}\} - 1] + \left[1 - \frac{P_{xl}}{P_{xu} - P_{xl}} \cdot \ln \frac{P_{xu}}{P_{xl}}\right] \qquad \text{(B3)} $$

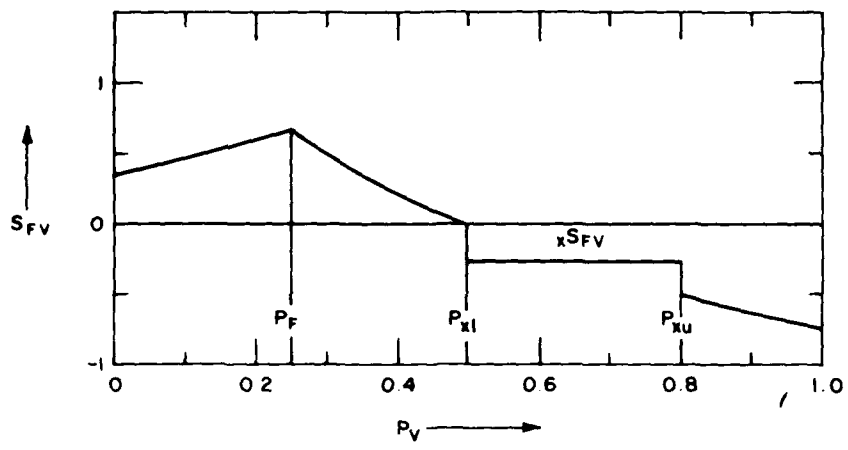


Figure B1. Showing the Distribution of Scores When There is a Category of Events from P_{xl} to P_{xu}

which corresponds to the probability given by

$$P'_V = e^{-(1+{}_xs_{FV})}/(1 - P_F) \tag{B4}$$

In the case of no-rain, with frequency $P_{NR} = P_{xu}$, $P_{xl} = 0$, the score for a forecast of a certain amount of rain, when in fact it does not rain, should be

$${}_{NR}s_{FV} = [-ln\{(1 - P_{NR})P_F\} - 1] + \left[1 + \frac{1}{P_{NR}} \cdot ln\,(1 - P_{NR})\right] \tag{B5}$$

Up to, and including, the categorized event, the cumulative probability should be P_{xu}. With this in mind, the likelihood, $LCS = P(s_o \geq s_{FV})$, that a chance score (s_o) can equal or exceed the achieved score (s_{FV}), follows. First, a few more terms need to be defined, in addition to those terms already given:

$$P'_{vu} = (1 - Pv')P_F/(1 - P_F) \tag{B6}$$

$$P_{v'l} = 1 - (1 - P_F)P_{v'}/P_F \tag{B7}$$

If $P_F < P_V$

If $P_V \geq P_F/(1 - P_F)$, then $LCS = P_V$

If $P_V < P_F/(1 - P_F)$

If $P_V \leq P_{xl}$, then LCS = $(P_V - P_F)/P_F$

If $P_V > P_{xl}$

If $P_V \leq P_{xu}$ (actually $P_V = P_{xu}$) then LCS = $P_{xu} - P_{vl}$

If $P_V > P_{xu}$

If $P_{xu} < P_{Vl}$, then LCS = $P_V - P_{vl}$

If $P_{xu} \geq P_{vl}$

If $P_{vl} > P_{xl}$,

If $s_{FV} \geq {}_x s_{FV}$, then LCS = $P_V - P_{xu}$

If $s_{FV} < {}_x s_{FV}$, then LCS = $P_V - P_{xl}$

If $P_{vl} \leq P_{xl}$, then LCS = $P_V - P_{Vl}$

If $P_F \geq P_V$

If $P_V \geq 2 - 1/P_F$

If $P_V \leq P_{xl}$

If $P_{xl} \geq P_{vu}$, then LCS = $P_{vu} - P_V$

If $P_{xl} < P_{vu}$,

If $P_{vu} \leq P_{xu}$,

If $s_{FV} > {}_x s_{FV}$, then LCS = $P_{xl} - P_V$

If $s_{FV} \leq {}_x s_{FV}$, then LCS = $P_{xu} - P_V$

If $P_{vu} > P_{xu}$, then LCS = $P_{vu} - P_V$

If $P_v > P_{xl}$

If $P_v \leq P_{xu}$ (actually $P_v = P_{xu}$) then LCS = $P_{v'u} - P_{x_1}$

(Note: If $P_{v'u}$ computes greater than 1.0, set it to 1.0)

If $P_v > P_{xu}$, then LCS = $(P_F - P_V)/(1 - P_F)$

If $P_V < 2 - 1/P_F$

If $P_V > P_{xu}$, then LCS = $1 - P_V$

If $P_V \leq P_{xu}$

If $P_V > P_{x\ell}$ (actually $P_V = P_{xu}$) then LCS = $1 - P_{x\ell}$

If $P_V \leq P_{x\ell}$, then LCS = $1 - P_V$

Appendix C

Program Steps in The B-G System of Evaluating Forecasts of Continuous Weather Elements and Chi-Square Tests of Significance

Step 1. Initialize: $N = 0$

$n_i = 0 \qquad i = 0(1)9$

$\sum_j (LCS)_j = \sum P(S_o \geq S_{FV}) = 0$

Set N_s: sample size

Step 2. Enter the (next) forecast and verification, respectively, as P_F, P_V.

Step 3. Find $LCS = P(S_o \geq S_{FV}) = P_V$, if $P_F < P_V$, $P_V \geq P_F/(1 - P_F)$

$= (P_V - P_F)/P_F$, if $P_F < P_V$, $P_V < P_F/(1 - P_F)$

$= (P_F - P_V)/(1 - P_F)$, if $P_F \geq P_V$, $P_V \geq 2 - 1/P_F$

$= 1 - P_V$, if $P_F \geq P_V$, $P_V < 2 - 1/P_V$

Step 4. Add $P(S_o \geq S_{FV})$ to $\sum_{j=1}^{N} \{P(S_o \geq S_{FV})\}_j$

Add 1 to N

Find i = int $\{10. P(S_o \geq S_{FV})\}$

Add 1 to n_i

Step 5. If $N \geq N_s$, go to Step 6.

If $N < N_s$, go to Step 2.

Step 6. Find $\overline{P(S_o \geq S_{FV})} = \sum_{j=1}^{N} \{P(S_o \geq S_{FV})\}_j / N$

Find $E = 1 - 2 \cdot \overline{P(S_o \geq S_{FV})}$

Step 7. Find $X_9^2 = \sum_{i=0}^{9} (N/10 - n_i)^2 / (N/10)$

Find $p_i = (i + 1)N/10 \qquad i = 0(1)8$

Find $_iX_1^2 = \left(Np_i - \sum_{j=0}^{i} n_i \right)^2 / \{p_i(1 - p_i)N\}$


ATE
LME