

CLIMATOLOGICAL ASPECTS OF THE BRIER P-SCORE

HARRY R. GLAHN and DONALD L. JORGENSEN

Techniques Development Laboratory, Weather Bureau, ESSA, Silver Spring, Md.

ABSTRACT

The P -score has come into widespread usage for the evaluation of probability forecasts of weather events. Verification data for operational Weather Bureau forecasts of probability of precipitation are used to study the tendency for the P -score to be a function of the precipitation climatology of the place for which the forecasts are made. It is found that simple models which give an expected P -score as a function of climatology can be useful in comparing different sets of forecasts and in assessing the general level of skill of the operational probability forecasts made by the Weather Bureau.

1. INTRODUCTION

The P -score introduced by Brier (1950) has come into widespread usage for the evaluation of probability forecasts of weather events. It is generally agreed that Brier's (1950) statement, ". . . (the use of the P -score) cannot influence the forecaster in any undesirable way . . .," is true. In other words, if a forecaster feels, for a given set of conditions, that the probability of precipitation is 40 percent, then 40 percent should be his forecast. In this context, Murphy and Epstein (1967) prove that the P -score is a "proper" scoring system.

Notwithstanding this desirable characteristic, the P -score, in common with other verification indices, does not completely fulfill all verification requirements of probability forecasts. If one is interested in comparing the goodness of forecasts made at different places or different seasons, he should be aware that the P -score does not take into account the difficulty of the forecasting problem as reflected in the climatology of the points for which the forecasts are made. Thus, one must be very cautious in comparing P -scores for forecasts made under one set of conditions with P -scores for forecasts made under other conditions.

During the past several years, P -scores for precipitation forecasts have been computed routinely for a network of 100 stations scattered throughout the conterminous 48 states (Roberts et al. 1967, 1969). The widespread geographical distribution of the stations together with a grouping of the P -scores by season provides an opportunity to study these scores for a broad range of climatic regimes. In this paper, some of the observed relationships between the P -score and climatology are discussed.

2. THE P -SCORE

The P -score as defined by Brier (1950) is

$$P\text{-score} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^n (f_{i,j} - E_{i,j})^2$$

where $f_{i,j}$ is the probability of the j th of r exhaustive and mutually exclusive events for the i th of n cases in the

sample, and correspondingly $E_{i,j}$ takes the value of 1 if the event occurred and 0 if it did not. Of course

$$\sum_{j=1}^r f_{i,j} = \sum_{j=1}^r E_{i,j} = 1$$

for each i .

When there are only two possible events ($r=2$), the score becomes

$$P\text{-score} = \frac{2}{n} \sum_{i=1}^n (f_i - E_i)^2$$

where f_i and E_i now refer to either of the two events. For example, in verifying forecasts of the probability of occurrence of precipitation, f_i is that probability for a particular case i , and E_i is 1 if precipitation occurred and 0 otherwise.

It is becoming common practice in the precipitation verification programs of the Weather Bureau to compute a number which is one-half of the P -score:

$$P = \frac{1}{n} \sum_{i=1}^n (f_i - E_i)^2.$$

The score P will be used in this paper; it has a possible range of 0 for a sample of perfect forecasts ($f_i=1$ when $E_i=1$, and $f_i=0$ when $E_i=0$) to 1 for the worst possible forecasts ($f_i=1$ when $E_i=0$, and $f_i=0$ when $E_i=1$).

3. DEFINITION OF SKILL

Much has been written about "skill" in regard to the verification of weather forecasts. We do not wish to dwell here on what is the best measure of skill. It does seem that in order to measure skill, the forecasts must be compared with some standard (Sanders 1963). This standard or "no skill forecast" can logically, in many circumstances, be climatology.

It can be seen that P is actually the mean square error of the set of forecasts. If the climatological probability

$$C = \frac{1}{n} \sum_{i=1}^n E_i$$

computed on the sample is used as a forecast for each of the n cases, the resulting score

$$P_c = \frac{1}{n} \sum_{i=1}^n (C - E_i)^2 = C(1 - C)$$

is the sample variance of the binary variable E , or the mean square error of climatological forecasts.

Since a perfect score is zero, the fractional amount by which the forecasts f_i improve on the climatological forecasts C is

$$\frac{P_c - P}{P_c} = K.$$

That is, of the possible improvement P_c , the actual fractional improvement is K . Viewed in this light, K is an adaptation of the Skill Score introduced by Heidke (1926) and has been discussed by Sanders (1963), Hughes (1965, 1967, 1968), and Roberts et al. (1967, 1969). It can also be interpreted as the reduction of variance of E due to the forecasts f .

4. A MODEL FOR COMPARING P-SCORES (MODEL 1)

Although P_c is a unique function of the climatological mean C , it is reasonable to postulate that the skill of experienced forecasters as defined by K does not vary with that aspect of climatology embodied in C . This requires that for several sets of forecasts made under conditions of differing C , $K = \text{constant}$, and

$$P = P_c(1 - K) = aP_c.$$

With sufficient samples of data, the constant a can be estimated by least squares, and the resulting equation can be used to estimate P for equally skillful forecasts for any C :

$$\hat{P} = (1 - K)P_c = aC(1 - C),$$

the variance P_c of E being reduced by the fraction K leaving the remainder \hat{P} .

The least-squares estimate of a is

$$a = \frac{\overline{P_c P}}{\overline{P_c^2}}$$

where a bar denotes an average over the sample.

5. A LOOK AT THE DATA

The data used in this study were compiled as a part of a continuing program to evaluate the performance of forecasts at selected Weather Bureau Offices (Roberts et al. 1967, 1969). P and C for local forecasts made at 100 stations within the conterminous States expressing the probability of precipitation for the three periods, today (1200-0000 GMT), tonight (0000-1200 GMT),

TABLE 1.—The K -score and the reductions of variance RV_Q and RV_L achieved by quadratic estimation $\hat{P} = (1 - K)P_c$ and linear estimation $\hat{P} = a_0 + a_1C$, respectively, for each of several subsamples.

Sample	No. of cases	K	RV_Q	RV_L
2 yr, both seasons, 1st period	398	0.363	0.686	0.587
" " " " 2d "	399	.191	.790	.708
" " " " 3d "	399	.110	.888	.792
" " summer 1st "	198	.299	.786	.762
" " " " 2d "	199	.136	.878	.863
" " " " 3d "	199	.078	.921	.888
" " winter 1st "	200	.429	.687	.554
" " " " 2d "	200	.241	.779	.716
" " " " 3d "	200	.144	.868	.757
1st yr, summer 1st "	100	.293	.820	.782
" " " " 2d "	100	.122	.912	.891
" " " " 3d "	100	.059	.926	.881
2d " " 1st "	98	.305	.723	.723
" " " " 2d "	99	.149	.833	.824
" " " " 3d "	99	.095	.920	.901
1st yr, winter 1st "	100	.412	.756	.630
" " " " 2d "	100	.230	.846	.743
" " " " 3d "	100	.133	.917	.791
2d " " 1st "	100	.443	.540	.449
" " " " 2d "	100	.251	.679	.673
" " " " 3d "	100	.153	.782	.705

and tomorrow (1200-0000 GMT) for 2 yr, April 1966 to March 1968, were available. The forecasts were issued in the early morning, about 0500 LST, and the statistics are stratified by season (April-September and October-March).

For each of several subsamples, the K -score was calculated. These scores are shown in table 1 along with RV_Q the reduction of variance of the P scores achieved by the quadratic curve $\hat{P} = aC(1 - C)$ and RV_L the fraction of the variance of the P scores explained by the linear relationship $\hat{P} = a_0 + a_1C$.

A portion of the data is plotted in figures 1 and 2. Figure 1 contains the P versus C values for the first period for the second-year winter season. In figure 2 are the values for the third period for the first-year summer season. These figures also include the quadratic and linear estimates \hat{P} and \hat{P} and the climatological score P_c .

From table 1 and figures 1 and 2, it is apparent that:

- 1) The quadratic curve \hat{P} fits the data well over the complete range of C . That is, for each interval of C , \hat{P} is near the mean of P .
- 2) The linear line \hat{P} does not fit the data as well; it generally overestimates P for both low and high C and underestimates it in midrange. Also, a very wet climate ($C > 0.5$) cannot be accommodated by the linear estimate. It should be about as easy to forecast for a station with $C = 0.7$ as for a station with $C = 0.3$. The quadratic estimate recognizes this, while the linear does not. In

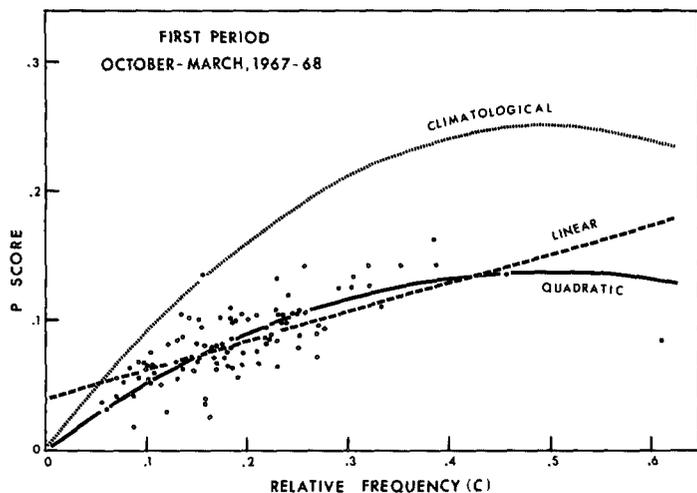


FIGURE 1.—Chart giving the relationships between the score P and the sample relative frequency. The dashed line and the solid line give the least-squares fit to the observed data in terms of a linear and a quadratic relationship, respectively. The dotted line indicates the score for climatological forecasts. Data are for the first-period forecasts for the winter season, 1967-68.

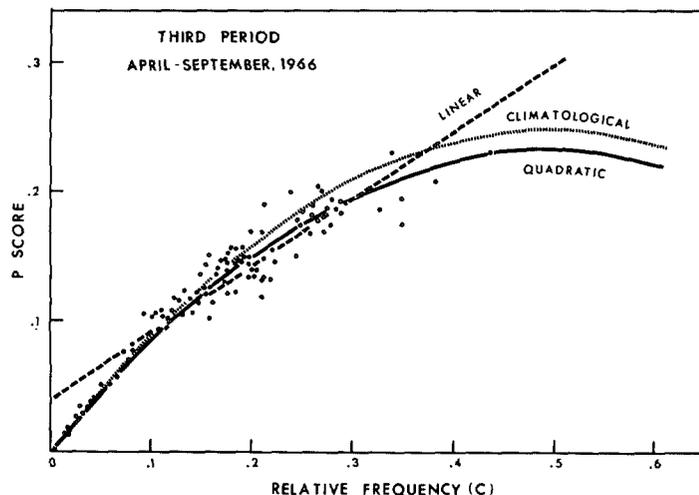


FIGURE 2.—Same as figure 1 except for the third-period forecasts for the summer season, 1966.

every subsample $RV_Q > RV_L$, even though the quadratic fit uses only 1 degree of freedom while the linear uses 2 degrees of freedom.

3) The skill as described by K is higher in the winter than in the summer, probably because the precipitation occurs in more organized patterns in winter than in summer.

4) The skill is very low for the third period in the summer. The improvement in P_c is only about 8 percent.

5) The improvement in P_c for the first period in winter is as high as 43 percent.

6) The skill was not much different for the 2 yr for which data were available. The greatest difference was for the third period in summer.

7) The skill for the third period in winter was slightly greater than the skill for the second period in summer.

8) All stations but one improved on climatology for the first period for the second-year winter (fig. 1). However, many stations failed to do this for the third period for the first-year summer season (fig. 2).

6. A MODEL INCORPORATING PERSISTENCE (MODEL 2)

Although $\hat{P} = aP_c$ gives a value with which P for a given station can be compared, there are other factors affecting the difficulty of forecasting at that particular station. Some of these factors may be purely local and can never be accounted for by a simple model. However, it is possible that an element of climatology, persistence, can be used to further "standardize" the P -scores. Let us define a persistence measure

$$Z = \frac{P[AB]}{P[AUB]}$$

where A and B are precipitation events in two consecu-

tive 12-hr periods, respectively, $P[AB]$ is the probability or relative frequency of both events, and $P[AUB]$ is the probability of either of the two events. We now define a relationship

$$P' = P_c(b_0 + b_1Z)$$

and fit the coefficients b_0 and b_1 by least squares.

The necessary equations are

$$b_0 = \frac{\overline{P_c P} \overline{P_c^2 Z^2} - \overline{P_c P Z} \overline{P_c^2 Z}}{\overline{P_c^2} \overline{P_c^2 Z^2} - \overline{P_c Z} \overline{P_c^2 Z}}$$

and

$$b_1 = \frac{\overline{P_c^2} \overline{P_c P Z} - \overline{P_c P} \overline{P_c^2 Z}}{\overline{P_c^2} \overline{P_c^2 Z^2} - \overline{P_c Z} \overline{P_c^2 Z}}$$

The variable Z was evaluated from data made available by Jorgensen (1967) and represents a climatological value of persistence.

Table 2 gives the constants and corresponding reductions of variance for the persistence model for several subsamples. By comparing values of RV_Q in table 1 with corresponding values of RV in table 2, one sees that the persistence factor adds over 5 percent for the subsample containing the first period for both seasons and 3 percent for the subsample containing the second period for both seasons. However, for the other subsamples the increase was less than 2 percent. The persistence factor, as defined, would be expected to be of more help in the early part of the forecast interval, namely, the first period.

Even though the persistence factor is of most help when the data are not stratified by season, for the first and second periods the overall reduction of variance,

$$\frac{\sum_{i=1}^2 n_i S_i^2 RV_{Q_i} + \sum_{i=1}^2 n_i (\bar{P}_i - \bar{P})^2}{NS^2}$$

TABLE 2.—The constants and reductions of variance for the persistence model $P' = P_c(b_0 + b_1Z)$

Sample	b_0	b_1	RV
2 yr, both seasons, 1st period	0.896	-0.810	0.740
" " " " 2d "	1.029	-.679	.821
" " " " 3d "	1.043	-.481	.901
" " summer 1st "	0.885	-.654	.800
" " " 2d "	1.027	-.571	.898
" " " 3d "	1.005	-.294	.923
" " winter 1st "	0.684	-.314	.677
" " " 2d "	.867	-.300	.785
" " " 3d "	.973	-.326	.875

where

n_i = season sample size,

S_i^2 = variance of season sample,

RV_{Q_i} = reduction of variance of season sample,

N = combined sample size,

S^2 = variance of combined sample,

\bar{P}_i = mean P for the sample, and

\bar{P} = mean P of combined sample,

is greater for model 1 with stratification, being 0.814 and 0.835 for the first and second periods, respectively, than for the persistence model with no stratification. For the third period, the overall reduction for model 1 with stratification by season is 0.900, about the same as the persistence model with no stratification.

7. A HIGHER ORDER MODEL (MODEL 3)

Even though the quadratic model appears to fit the data very well (as indicated in figs. 1 and 2), if the K -score is plotted versus C , a definite relationship is noted. The same two samples shown in figures 1 and 2 are plotted in figures 3 and 4. Although nothing more complicated than a linear trend is apparent, the relationship could just as well be quadratic and symmetric about $C=0.5$. After the relationship is seen in figures 3 and 4, it can also be noticed in figures 1 and 2 as a slight tendency for the points to lie above the solid line at low C .

A third model which accounts for this higher order relationship between P and C is

$$K = \frac{P_c - P}{P_c} = \gamma - \beta P_c$$

or

$$P = P_c(1 - \gamma + \beta P_c) = P_c(\alpha + \beta P_c)$$

where $\alpha = 1 - \gamma$.

This model is in the same form as the persistence model with Z replaced by P_c . The least-squares estimates of α and β are

$$\alpha = \frac{\bar{P}_c \bar{P} \bar{P}_c^2 - \bar{P}_c^2 \bar{P} \bar{P}_c}{\bar{P}_c^2 \bar{P}_c^2 - \bar{P}_c^3 \bar{P}_c^3}$$

and

$$\beta = \frac{\bar{P}_c^2 \bar{P}_c^2 \bar{P} - \bar{P}_c \bar{P} \bar{P}_c^3}{\bar{P}_c^2 \bar{P}_c^2 - \bar{P}_c^3 \bar{P}_c^3}$$

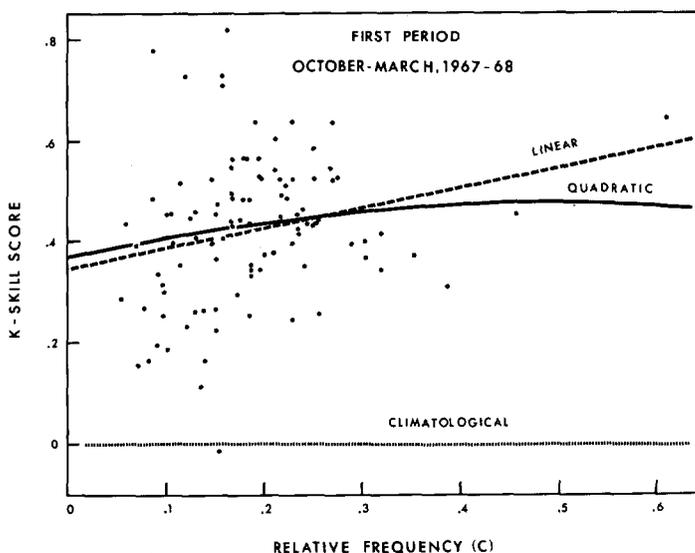


FIGURE 3.—Chart showing the values of K plotted against sample relative frequency for the first-period forecasts for the winter season, 1967-68. The dashed line gives the linear trend of K against frequency. The solid line gives a similar trend based on the higher order model (model 3). The dotted line gives the skill of climatological forecasts, that is, zero for all values of the frequency.

Table 3 gives estimates of α and β and the associated reductions of variance derived from the various sub-samples. The appropriate solutions are also plotted on figures 3 and 4 as solid lines.

A comparison of the data in tables 1, 2, and 3 reveals that the increased reduction of variance of P due to the higher order term is of the order of 1 percent or less. Also, there is little difference in reductions of variance for the persistence model and the higher order model. All of the regression curves defined in table 3 are concave downward as in figures 3 and 4 except for the first-year winter season, first period. In this case the trend is of little importance, and a linear fit explains only 2.7 percent of the variance of K .

The regression coefficients were estimated by minimizing the mean-square error of the estimate of P . Slightly different results obtain if the coefficients are estimated by minimizing the mean-square error of the estimate of K . A curve fitted by the latter method will, of course, fit the K -scores better in a least-squares sense. On the other hand, a curve fitted by the former method will fit the P scores better. Quite probably, a comparison of individual station values of P with the curve obtained from this model is about as much as can be legitimately done. Even then, local difficulties of forecasting may influence a particular station to lie consistently above (or below) the line.

8. LONG-TERM CLIMATOLOGY AS A STANDARD

Some may consider it unfair to the forecaster to compare his forecasts to the sample climatology, since the sample climatology gives a better P -score than the long-term

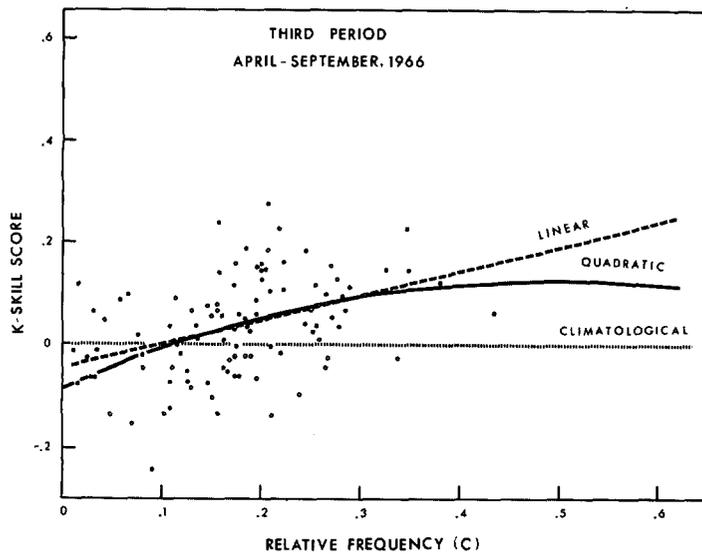


FIGURE 4.—Same as figure 3 except for the third-period forecasts for the summer season, 1966.

TABLE 3.—The constants and reductions of variance for the higher order model for each of several subsamples. Sample climatology (C) was used as a basis for comparison.

Sample	α	β	RV
2 yr, both seasons, 1st period	0.667	-0.173	0.687
" " " " 2d "	.932	-.742	.801
" " " " 3d "	.976	-.507	.892
" " summer 1st "	.778	-.454	.790
" " " 2d "	1.011	-.929	.890
" " " 3d "	1.020	-.571	.925
" " winter 1st "	0.578	-.039	.667
" " " 2d "	.817	-.335	.782
" " " 3d "	.946	-.532	.875
1st yr, summer 1st "	.788	-.472	.825
" " " 2d "	1.089	-1.353	.932
" " " 3d "	1.088	-0.858	.936
" " " 1st "	0.768	-.425	.728
" " " 2d "	.928	-.474	.837
" " " 3d "	.935	-.177	.921
1st " winter 1st "	.517	.423	.761
" " " 2d "	.806	-.210	.847
" " " 3d "	.868	-.007	.917
2d " " 1st "	.632	-.438	.549
" " " 2d "	.827	-.447	.686
" " " 3d "	1.021	-1.016	.814

climatology (unless the two are equal) and is not available to the forecaster when he makes his forecasts. If desired, the three models given above can be used for comparing a set of forecasts with the long-term climatology. In each case, the equations for the coefficients would be the same except that P_c would be replaced by P_L where

$$P_L = C(1 - C) + (L - C)^2 = P_c + (L - C)^2;$$

L and C are the long-term climatology and the sample climatology, respectively.

The coefficients for model 3 were computed with P_L substituted for P_c ; they are given in table 4. For all subsamples and for nearly the whole range of C , model 3 shows the skill K computed with P_L as a base to be greater than that computed with P_c as a base. The difference is not great, being generally less than 0.02; however, for the third period this may make the difference between positive and negative skill for certain values of climatology (fig. 2).

9. CONCLUSIONS

Three models have been presented with which individual station values of P can be compared. The persistence model is relatively difficult to apply and does little or no better than model 3 with seasonal stratification. Model 3 does fit the data slightly better than model 1 and can be used to advantage. Each of the three models is definitely better than the linear model and furnishes a rough means of standardizing the P -scores. However, any conclusions resulting from comparisons of this kind must be made with caution.

Perhaps the best use that can be made of such fitted curves is to help provide an answer to the general question, "How skillful are the operational Weather Bureau fore-

TABLE 4.—The constants and reductions of variance for the higher order model for each of the several subsamples. Long-term climatology (L) was used as a basis for comparison.

Sample	α	β	RV
2 yr, summer, 1st period	0.763	-0.406	0.796
" " " 2d "	.992	-.853	.892
" " " 3d "	1.006	-.546	.930
" " winter 1st "	0.584	-.123	.657
" " " 2d "	.803	-.323	.790
" " " 3d "	.927	-.489	.867
1st yr, summer, 1st "	.772	-.428	.827
" " " 2d "	1.060	-1.211	.934
" " " 3d "	1.073	-0.831	.937
2d " " 1st "	0.751	-.373	.738
" " " 2d "	.917	-.456	.839
" " " 3d "	.918	-.135	.931
1st " winter, 1st "	.533	.258	.748
" " " 2d "	.795	-.218	.853
" " " 3d "	.854	-.017	.908
2d " " 1st "	.639	-.511	.539
" " " 2d "	.812	-.409	.696
" " " 3d "	1.010	-.996	.809

casts of probability of precipitation?" For instance, the constants can be evaluated for model 3 and the resulting curves of K plotted as a function of C for successive yearly samples of third-period summer forecasts. The general level of the curves for a particular value of C shows the general level of skill for that value of C . If the curves tend over the years to become higher (or lower) for a particular range of values of C , then this would represent an increase (or decrease) in skill.

ACKNOWLEDGMENTS

We are grateful to Charles F. Roberts, John M. Porter, and Geraldine F. Cobb of the Technical Procedures Branch of the Weather Bureau for making available the verification data on

which this investigation is based. Without these data, this study could not have been carried out.

REFERENCES

- Brier, Glenn W., "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, Vol. 78, No. 1, Jan. 1950, pp. 1-3.
- Heidke, Paul, "Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst," (Successful and Unsuccessful Methods of Calculations of Wind Force at Storm-Warning Centers), *Geografiska Annaler*, Vol. 8, Srenska sällskap et för antropologi och geografi, Stockholm, 1926, pp. 301-349.
- Hughes, Lawrence A., "On the Probability Forecasting of the Occurrence of Precipitation," *Weather Bureau Technical Note* 20-CR-3, ESSA, Weather Bureau Central Region, Kansas City, Mo., Nov. 1965, 36 pp.
- Hughes, Lawrence A., "Probability Verification Results (6-Month and 18-Month)," *ESSA Technical Memorandum* WBTM CR-17, Weather Bureau Central Region, Kansas City, Mo., June 1967, 22 pp.
- Hughes, Lawrence A., "Probability Verification Results (24 Months)," *ESSA Technical Memorandum* WBTM CR-19, Weather Bureau Central Region, Kansas City, Mo., Feb. 1968, 29 pp.
- Jorgensen, Donald L., "Climatological Probabilities of Precipitation for the Conterminous United States," *ESSA Technical Report* WB-5, U.S. Department of Commerce, Washington, D.C., Dec. 1967, 60 pp.
- Murphy, Allan H., and Epstein, Edward S., "A Note on Probability Forecasts and 'Hedging,'" *Journal of Applied Meteorology*, Vol. 6, No. 6, Dec. 1967, pp. 1002-1004.
- Roberts, Charles F., Porter, John M., and Cobb, Geraldine F., "Report on the Forecast Performance of Selected Weather Bureau Offices for 1966-1967," *ESSA Technical Memorandum* WBTM FCST-9, U.S. Department of Commerce, Washington, D.C., Dec. 1967, 52 pp.
- Roberts, Charles F., Porter, John M., and Cobb, Geraldine F., "Report on Weather Bureau Forecast Performance 1967-68 and Comparison With Previous Years," *ESSA Technical Memorandum* WBTM FCST-11, U.S. Department of Commerce, Washington, D.C., Mar. 1969, 44 pp.
- Sanders, Frederick, "On Subjective Probability Forecasting," *Journal of Applied Meteorology*, Vol. 2, No. 2, Apr. 1963, pp. 191-201.

[Received July 8, 1969]