DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

WEATHER BUREAU
F. W. REICHELDERFER, Chief

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 1950
Issued April 15, 1950

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

INTRODUCTION

Verification of weather forecasts has been a controversial subject for more than a half century. There are a number of reasons why this problem has been so perplexing to meteorologists and others but one of the most important difficulties seems to be in reaching an agreement on the specification of a scale of goodness for weather forecasts. Numerous systems have been proposed but one of the greatest arguments raised against forecast verification is that forecasts which may be the "best" according to the accepted system of arbitrary scores may not be the most useful forecasts. In attempting to resolve this difficulty the forecaster may often find himself in the position of choosing to ignore the verification system or to let it do the forecasting for him by "hedging" or "playing the system." This may lead the forecaster to forecast something other than what he thinks will occur, for it is often easier to analyze the effect of different possible forecasts on the verification score than it is to analyze the weather situation. It is generally agreed that this state of affairs is unsatisfactory, as one essential criterion for satisfactory verification is that the verification scheme should influence the forecaster in no undesirable way. Unfortunately, the criterion is difficult, if not impossible to satisfy, although some schemes will be much worse than others in this respect.

It is the purpose of this paper to discuss one situation where it appears to be possible to devise a verification scheme that cannot influence the forecaster in any undesirable way. This is the case when forecasts are expressed in terms of probability statements. The advantages of expressing the degree of assumed reliability of a forecast numerically have been discussed previously [1, 2, 3, 4] so that the purpose here will not be to emphasize the enhanced usefulness of such forecasts but rather to point out how some aspects of the verification problem are simplified or solved.

VERIFICATION FORMULA

Suppose that on each of n occasions an event can occur in only one of r possible classes or categories and on one such occasion, i, the forecast probabilities are f_{i1}, f_{i2}, . . . f_{ir}, that the event will occur in classes 1, 2, . . . r, respectively. The r classes are chosen to be mutually exclusive and exhaustive so that

$$\sum_{j=1}^{r} f_{ij}=1,\ i=1,2,3,\ldots n \tag{1}$$

A number of interesting observations can be made about a vertification score P defined by

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

where E_{ij} takes the value 1 or 0 according to whether the event occurred in class j or not. Before discussing this score in detail it will be instuctive to consider an illustrative example. Table 1 shows 10 actual forecasts ($n=10$) of rain or no-rain ($r=2$) in which a probability or confidence statement (f_{ij}) was made for each forecast. In the table, in accordance with the definition of E_{ij}, unity is placed in the rain column and zero in the no-rain column if rain occurs; and if the event is no-rain, unity is placed in the no-rain column and zero in the rain

877086—50——1

column. According to formula (2) the score P for these forecasts is

$$P=\frac{1}{10}(0.7^2+0.1^2+0.2^2+\ldots+0.1^2)=0.19$$

From consideration of this example or of formula (2), it is obvious that the score P has a minimum value of zero for perfect forecasting and a maximum value of 2 for the worst possible forecasting. Perfect forecasting is defined as correctly forecasting the event to occur with a probability of unity or 100 percent confidence. The worst possible forecast is defined as stating a probability of unity or certainty for an event that did not materialize (and also, of course, stating a probability of zero for the event that did materialize).

It is also easy to show that if $p_1, p_2, \ldots p_r$, are the relative frequencies that the event occurred in classes $1, 2, \ldots r$, then the minimum score that can be obtained by forecasting the same thing on every occasion is when

$$f_{ij}=p_j, \qquad i=1, 2, \ldots n \tag{3}$$

This will minimize the score P for constant values of $f_{1j}=f_{2j}=\ldots f_{nj}$ and the mean value of the score will be

$$P'=1-\sum_{j=1}^{r} p_j^2 \tag{4}$$

In the example given here (table 1) there are two classes, rain or no-rain, so $r=2$. It rained 3 out of 10 times so if the forecaster had no skill in differentiating one occasion from another, thus making the same forecast each time, he should put down 0.3 for the probability of rain and 0.7 for the probability of no-rain in order to get the best score. Of course in actual practice he doesn't know in advance the relative frequencies $p_1, p_2, \ldots p_r$ so he would ordinarily use the best estimates of the p_j based on climatological studies. The important point is, however, that if he has some skill in forecasting an average departure from the climatological probabilities he should make use of it. Thus, in the example given, a forecast of 0.3 probability for rain on every occasion would give a score

$$P'=1-(0.3^2+0.7^2)=0.42$$

If for these same 10 forecasts a climatological probability of say 0.2 for rain had been used on every occasion the corresponding score is 0.44. Thus the forecaster receives credit for recognizing or forecasting a departure from the normal conditions through the period even though he may not be able to distinguish one occasion from another within the period.

TABLE 1.—*Example of forecasts stated in terms of probability*

Occasion i	Rain		No rain	
	Forecast probability f_{i1}	Observed E_{i1}	Forecast probability f_{i2}	Observed E_{i2}
1	0.7	0	0.3	1
2	.9	1	.1	0
3	.8	1	.2	0
4	.4	1	.6	0
5	.2	0	.8	1
6	0	0	1.0	1
7	0	0	1.0	1
8	0	0	1.0	1
9	0	0	1.0	1
10	.1	0	.9	1

In addition to encouraging the forecaster to minimize his score P by getting the forecasts exactly right and stating a probability of unity, he is encouraged to state unbiased estimates of the probability of each event when he cannot forecast perfectly. A little experience with the use of score P will soon convince him that he is fooling nobody but himself if he thinks he can beat the verification system by putting down only zeros and unities when his forecasting skill does not justify such statements of extreme confidence. And in the complete absence of any forecasting skill he is encouraged to predict the climatological probabilities instead of categorically forecasting the most frequent class on every occasion.

COMPARISON OF FORECAST AND OBSERVED PROBABILITIES

When a series of forecasts has been made using probability statements a study can also be made to determine whether the forecast probabilities are related to the relative frequency of the events' occurrence. An example of this type of comparison is shown in table 2 (based on a more extended series of such forecasts), which suggests a relationship between the forecast and observed probabilities but indicates that the forecaster should modify or adjust his scale to improve the forecasts. However, knowledge of a good relationship between forecast and observed probabilities is not sufficient to indicate how useful the forecasts are, for it is also necessary to know the frequency with which forecasts are made in the various categories. In general, the most useful forecasts are those which fall into the extreme classes shown in table 2. The score P depends on both the frequency distribution of the forecast probability statements and the correlation between the forecast and observed probabilities. In other words, in order for P to become smaller the correlation must increase and the proportion of forecasts in the extreme classes must increase.

TABLE 2.—*Verification of a series of 85 forecasts expressed in terms of the probability of rain*

Forecast probability of rain	Observed proportion of rain cases	Forecast probability of rain	Observed proportion of rain cases
0.00–0.19	0.07	0.60–0.79	0.40
0.20–0.39	.10	0.80–1.00	.50
0.40–0.59	.29		

CONCLUSIONS

During the past several months the score P has been used in the verification of some experimental precipitation forecasts for 6-hour periods made by members of the Short Range Forecast Development Section of the Weather Bureau. Although initially there was some criticism of this scoring method, most of it seemed to be a result of lack of understanding of the method or due to the disillusionment of finding so little ability to call for precipitation with much certainty or confidence. However, a few questions have been raised that indicate further study is needed regarding this particular scoring method and it is hoped that this paper will stimulate others to investigate this phase of the verification problem.

REFERENCES

1. G. W. Brier, "Verification of a Forecaster's Confidence and the Use of Probability Statements in Weather Forecasting," *Research Paper No. 16*, U. S. Weather Bureau, Washington, February 1944.
2. W. E. Cooke, "Weighting Forecasts," *Monthly Weather Review*, vol. 34, No. 6, June 1906, pp. 274–275.
3. Cleve Hallenbeck, "Forecasting Precipitation in Percentages of Probability," *Monthly Weather Review*, vol. 48, No. 11, November 1920, pp. 645–647.
4. Saul Price, "Thunderstorm Today?—Try a Probability Forecast," *Weatherwise*, vol. 2, No. 3, June 1949, pp. 61–63, 67.