NWS-CR-TA 88-26

CRH SSD
JULY 1988

CENTRAL REGION TECHNICAL ATTACHMENT 88-26

AWS TECHNICAL LIBRARY
FL 4414
SCOTT AFB IL 62225
2 0 JUL 1988

A NEW MEASURE OF FORECAST SKILL USING THE BRIER SCORE

Richard P. McNulty
National Weather Service Forecast Office
Topeka, Kansas

1. Introduction

Verification of probability forecasts has been a topic of discussion for many years. One of the most widely used verification scores for probability of precipitation (POP) is the Brier score (Brier, 1950). A slightly modified version of this score (Sanders, 1963) is used by the National Weather Service (NWS). The National Verification Plan (NVP) (NWS, 1982) defined the Brier score in terms of this modification:

$$B = \frac{1}{N} \sum_{i=1}^{N} (P_i - \delta_i)^2 \qquad (1)$$

where N total number of forecasts;
P_i POP associated with the ith forecast; and
δ_i equals 1 if rain occurs, or equals 0 if no rain occurs (for the ith forecast).

The quantity $(P_i - \delta_i)$ is the difference between the forecast probability and the perfect forecast probability for the ith forecast. A perfect forecast probability is 1 (or a 100 percent chance) when it rains and 0 when no rain occurs. Thus equation (1) is an error measure, specifically the mean squared error for a forecast probability relative to a perfect forecast probability.

The Brier score, B, ranges from 0 for a series of perfect forecasts (i.e., 100 percent forecast POP for every rain event, 0 percent forecast POP for every no rain event) to 1 for a series of totally bad forecasts. Thus, the goal of any forecaster is to minimize his/her Brier score.

2. Measuring Skill

The Brier score by itself is an error measure, but says nothing about forecast skill. The quantity traditionally used to measure skill is the skill score. The Glossary of Meteorology (Huschke, 1959) defines skill score as:

> "...an index of the degree of skill of a set of forecasts, expressed with reference to some standard such as forecasts based upon chance, persistence or climatology."

One form of skill score measures the improvement of the forecast or forecast score over a reference value or a score based on the reference level. This takes the general form:

$$S = \frac{R - F}{R} \qquad (2)$$

where S is the skill score, R is the reference and F is the forecast. This form is used when the forecaster wants to minimize F (as with the Brier score). Positive values of S represent improvement over the reference level; negative values of S represent less skill than the reference level.

Choosing the reference level is somewhat arbitrary. The Glossary of Meteorology suggests chance, persistence or climatology. Two of these will be discussed below. The purpose of this paper is to suggest another reference level. All these reference levels can be considered "levels of no forecast skill", i.e., a forecaster selecting a POP based on the reference level needs to know little or nothing about the current meteorological situation.

3. Reference Levels

A. Chance

If the forecast probability is chosen at random, and if a sufficiently large number of forecasts are made, the number of forecasts associated with each probability value should be uniformly distributed. Similarly with a sufficiently large number of forecasts, it can be assumed that the rain events are also uniformly distributed.

If r discrete probability values are forecast, then N/r forecasts and n/r rain events (where n is the number of rain events in the sample and $0 <= n <= N$) are associated with each discrete POP value. POP values are expressed as:

$$P_i = \frac{i}{r-1} \quad i = 0, 1, 2 \ldots r-1. \qquad (3)$$

For the uniformly distributed case, equation (1) can be written:

$$B = \frac{1}{N} \left[\frac{n}{r} \sum_{i=0}^{r-1} (P_i - 1)^2 + \frac{N-n}{r} \sum_{i=0}^{r-1} P_i^2 \right] \qquad (4)$$

Due to the definition of P_i in equation (3), it can be shown that:

$$\sum_{i=0}^{r-1} (P_i - 1)^2 = \sum_{i=0}^{r-1} P_i^2 \qquad (5)$$

Equation (5) reduces equation (4) to:

$$B = \frac{1}{r} \sum_{i=0}^{r-1} P_i^2 \qquad (6)$$

For r = 11, P_i equals 0, 0.1, 0.2, ..., 0.9 and 1.0, and B = 0.35. Thus, for a random choice of P_i and a sufficiently large sample, the Brier score would equal 0.35. If chance were used as a reference standard, Brier scores less than 0.35 would indicate skill.

B. Climatology

Brier and Allen (1951) were one of the early references to suggest that the climatological frequency of precipitation be used as a reference standard. The NVP lists the "Brier score improvement over climatology" as one of its measures. This choice is based on a property of the Brier score described below.

Assume that P_i takes on a constant value, f. Equation (1) then becomes:

$$B = \frac{1}{N} \left[\sum^{n} (f - 1)^2 + \sum^{N-n} f^2 \right] \qquad (7)$$

where the summations are over the n "wet" forecasts and N-n "dry" forecasts. After some manipulations, equation (7) reduces to:

$$B = \frac{n}{N} (1 - 2f) + f^2 \qquad (8)$$

To find the value of f that minimizes B, the derivative of equation (8) is taken and set equal to zero:

$$\frac{dB}{df} = -\frac{2n}{N} + 2f = 0 \quad \text{or} \quad f = \frac{n}{N}$$

Thus, B is minimized when P_i is equal to the relative frequency of precipitation in the forecast sample. Figure 1a shows a plot of equation (8) for B versus f for various values of relative frequency (n/N). The minimum values of B are smaller than for the random case. The lowest value for any given n/N is:

$$B_o = \frac{n}{N} - \left(\frac{n}{N}\right)^2 \qquad (9)$$

The reason for suggesting climatology as a reference standard lies in the above analysis. Forecasting a constant rain probability equal to the climatological frequency of rain would minimize the Brier score for this "no skill" forecast.

4. A New Measure

If a Brier score was calculated for a very long series of forecasts, the frequency of rainfall during that series would approach the climatological frequency of rainfall. However, Brier scores calculated for operational forecasts are usually for periods of six months (warm or cold season) or one year. The relative frequency of rainfall during this shorter period can vary considerably from the climatological frequency. This raises a question: should the climatological frequency of precipitation be used as a reference standard for short period calculations of Brier score improvements over a reference level? This paper suggests the answer is no.

If a forecaster knows beforehand what the relative frequency of rainfall will be over a forecast period, he/she could forecast a constant POP equal to that relative frequency and minimize the Brier score (in a no skill sense). This argument suggests that equation (9) is a better reference standard for Brier score improvement calculations over short forecast periods. More specifically, the skill score derived from this standard would take the form:

$$S_o = \frac{B_o - B}{B_o} \qquad (10)$$

Figure 1b shows B_o versus relative frequency. It can be seen that B_o is maximum for a relative frequency of 50 percent and tapers to zero for all rain or no rain series. It is interesting to note that for all rain or no rain cases, a forecaster would have to forecast all rain or no rain, respectively, in order to not lose to the reference level. However, the forecaster cannot improve over the standard for these extreme cases.

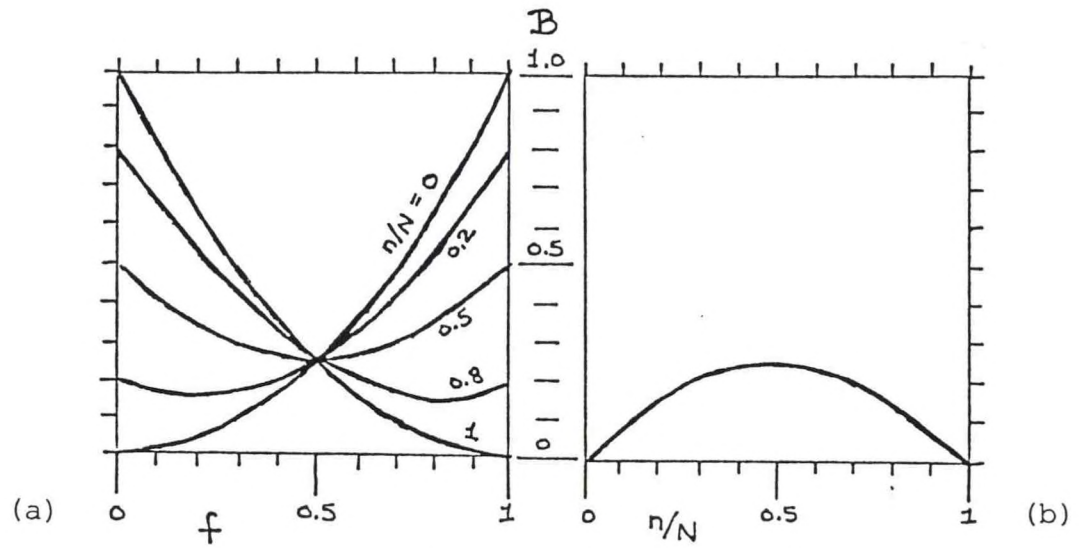

Fig. 1 (a) Plot of Brier score, B, versus the constant POP forecast, f; based on equation (8). Curves represent forecast samples with different relative frequencies of precipitation (n/N). (b) Plot of Brier score, B, versus the relative frequency of precipitation (n/N); based on equation (9). Curve shows the minimum Brier score for a constant POP forecast equal to n/N.

5. An Example

The new skill score, S_o, was applied to the cold season (October 1987 to March 1988) precipitation statistics at WSFO Topeka. These results are shown in Table 1 as the percent improvement over BZERO. The skill score shows that forecasters improved over their "no skill" level by as much as 61.9 percent. All forecasters except one showed an improvement of at least 24 percent. Four forecasters had an improvement better than the station average. Although the forecaster who experienced the highest relative frequency of precipitation showed the greatest improvement, S_o shows little correlation with the relative frequency of precipitation in general.

Brier scores for the ten forecasters ranged from 0.038 to 0.117. These magnitudes of themselves are good and show an average probability error of less than 10 percent. It is interesting to note that the forecaster with the best Brier score also experienced the lowest relative frequency of precipitation. For reference, the forecasters are ranked by Brier score and skill score. The rankings for the skill score are considerably different than those for the Brier score.

For comparison, the percent improvement over the climatological Brier score (for October through March) and over the Brier score for random POP forecasts are included in Table 1. Using equation (7), the monthly climatological POP's for Kansas, and the actual frequency of wet forecasts, a climatological Brier score of 0.108 was calculated for the period October, 1987, through March, 1988. Nine of the ten forecasters improved over this climatological Brier score by as much as 64.8 percent. One forecaster lost to climatology by 8.3 percent. Note that the relative forecaster ranking for this climatology based skill score is the same as the Brier score ranking.

In Section 3A the Brier score for forecasts based on a random POP was shown to be 0.35. The percentage improvement (skill) over this value is also shown in Table 1. All forecasters improved over this random POP Brier score. Improvements ranged from 66.6 percent to 89.1 percent. Note that the relative forecaster ranking for this random POP based skills score is again the same as the Brier score ranking.

A total assessment of these statistics would look for the forecaster who showed both skill (high improvement or skill score) and low error (low Brier score). Since the skill scores based on climatology and chance show the same relative ranking as the Brier score, little additional information is provided by these statistics. The new skill score, on the other hand, provides a different basis for comparison.

Comparing the Brier score and S show that Forecaster C, D and E were all better than the station average for both scores. Forecasters D and E had a Brier score almost 0.02 better than C, whereas Forecasters C and E had skill scores about 10 percent better than D. Of the three forecasters, E appears to have the best combined results.

TABLE 1

Cold Season Precipitation Statistics for WSFO Topeka (10/87-3/88)

FCSTR	#FCST	#PCPN	FREQ	BRIER SCORE	BRIER RANK	BZERO	% IMPROVEMENT OVR CLIMT	CHNCE	BZERO	SKILL RANK
A	591	93	0.157	0.081	9	0.132	25.0	76.9	38.6	6
B	507	46	0.091	0.057	5	0.083	47.2	83.7	31.3	8
C	252	45	0.179	0.056	4	0.147	48.1	84.0	61.9	1
D	498	41	0.082	0.038	1	0.075	64.8	89.1	49.3	4
E	461	48	0.104	0.039	2	0.093	63.9	88.9	58.1	2
F	489	55	0.112	0.057	6	0.099	47.2	83.7	42.4	5
G	210	18	0.086	0.060	7	0.079	44.4	82.9	24.1	9
H	267	44	0.165	0.064	8	0.138	40.7	81.7	53.6	3
I	228	20	0.088	0.054	3	0.080	50.0	84.6	32.5	7
J	126	18	0.143	0.117	10	0.123	-8.3	66.6	4.9	10
all	3794	439	0.116	0.058		0.103	46.3	83.4	43.7	

Forecasters B, F, H and I were better than the station average for either Brier score or skill score, but not both. These forecasters would occupy the middle echelon in forecast ability. It is interesting to note that Forecast I had a better Brier score than Forecaster C. However, the considerably lower skill level dropped Forecaster I from the upper echelon into the middle level.

The remaining forecasters (A, G and J) showed worse than average scores in both categories. Both scores indicate that Forecaster J needs to work hardest on his/her POP forecasts.

6. Conclusions

This paper argues that a skill score based on Brier score improvement over a relative "no skill" reference standard is better than one based on improvement over climatology. The main argument in favor of this relative reference level is the difference between the climatological rainfall frequency and the actual rainfall frequency during an evaluation period. Use of the relative reference level attempts to minimize the effect of rainfall frequency on skill measurement. The relative reference level, B_o, gives the best Brier score possible for a constant POP forecast for the forecast sample experienced by a forecaster.

7. References:

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78, 1-3.

_____, and R.A. Allen, 1951: Verification of weather forecasts. Compendium of Meteorology. Amer. Meteor. Soc., 841-848.

Huschke, R.E., ed., 1959: Glossary of Meteorology. Amer. Meteor. Soc., 638 pp.

National Weather Service, 1982: National Verification Plan, 81 pp.

Sanders, F., 1963: On subjective probability forecasts. J. Appl. Meteor., 2, 191-201.