# The Verification of Probability Forecasts

FREDERICK SANDERS[1]

*Massachusetts Institute of Technology, Cambridge*

(Revised manuscript received 23 December 1966)

ABSTRACT

Brier's scoring procedure for the evaluation of probability statements is analyzed to show two important aspects of the forecasting process. The first is a *sorting* process in which the forecaster assigns each prediction to one of a set of ordered classes of likelihood of occurrence of the meteorological event. The second is a *labeling* process in which he assigns a numerical value to each class. This value is intended to be the relative frequency (or probability) of occurrence of the event for the predictions in that class. When forecasts are evaluated relative to the use of simple statements such as climatological probability, the Brier score is shown to consists of a *sorting gain* and a *bias* (or mislabeling) *penalty*. Evidence is presented to show that meteorological forecasts made by humans have appreciable sorting skill and suffer little bias penalty. The relevance of the bias penalty is attacked and defended.

## 1. Introduction

One of the factors which spurred this author's interest in probability forecasting was the assertion by Brier and Allen (1951) that a certain proposed verification procedure (Brier, 1950) could not be "played." What do we mean when we say that a verification procedure cannot be played? Perhaps that no strategy can be employed which will optimize the verification score at the expense of the utility of the forecasts. Though the utility of forecasts is not the main subject of inquiry in this discussion, it is pertinent to note that Hunt (1963) recently demonstrated a direct correspondence between the Brier score and overall operational value provided that the probability threshold for operational decision were randomly distributed. Aside from the question of utility, the concept of unplayability must refer to a circumstance in which the strategy for optimizing the score does not obscure or suppress the full amount of scientific skill which is contained in the forecaster, be he objective system or human individual. This assertion, however, leaves undefined what we mean by "scientific skill."

## 2. Analysis of the Brier score

Let us proceed with this question in a roundabout manner by analyzing the Brier score. The analysis, similar to that presented by Sanders (1963), discloses two rather specific aspects of forecast performance which seem to comprise a reasonably adequate definition of skill. As originally proposed, the score, expressed as an average over $N$ forecasts, is

$$F=\frac{1}{N}\sum_{i=1}^{N}(f_i-O_i)^2. \tag{1}$$

Here $f_i$ is the forecast probability of occurrence in the $i$th instance and $O_i$ takes on the value one or zero, depending on whether the event in fact occurs or does not occur. The object of the game is obviously to keep the score as low as possible, by being as certain as possible and, not the least, by being right as often as possible. Now suppose that we require that $f$ can take only certain discrete values, say, integral numbers of chances in 10. Then we may divide the $N$ forecasts into eleven categories, representing forecast probabilities of 0, 1, 2,$\cdots$10, chances in 10. The average score for the $M_k$ forecasts in the $k$th category is

$$F_k=\frac{1}{M_k}\sum_{i=1}^{M_k}(f_k-O_{ki})^2=(f_k-\bar{O}_k)^2+\bar{O}_k(1-\bar{O}_k), \tag{2}$$

where the overbar refers to an average over the cases in this category.

The two quantities on the right side of (2) each have a distinct meaning. The first, $(f_k-\bar{O}_k)^2$, measures the amount of bias in the forecasts in this category (provided $M_k$ is large enough). That is, this contribution to the score is smaller the less difference there is between the stated probability $f_k$ and the relative frequency of occurrence $\bar{O}_k$. The second, $\bar{O}_k(1-\bar{O}_k)$, measures the degree of certainty in the forecasts. It is largest when $\bar{O}_k$ is 0.5 and vanishes when $\bar{O}_k$ is either zero or one. Now the average score for the whole sample is obtained by averaging the mean scores for each category, weighted by the number of forecasts in that cate-

[1] Present affiliation: National Hurricane Research Laboratory, Coral Gables, Fla.

gory. Thus,

$$F=\overline{\overline{(f_k-\bar{O}_k)^2}}+\overline{\overline{\bar{O}_k\,(1-\bar{O}_k)}}, \tag{3}$$

where

$$\overline{\overline{(\ )}}\equiv\frac{1}{N}\sum_{k=1}^{11} M_k[\ \ ].$$

In the process of making these forecasts, the score suggests to us that two things have been done. First, all instances have been *sorted* into eleven categories of *qualitative* likelihood of occurrence. Second, each of these categories has been *labeled* with a certain *quantitative* probability. The Brier score will be lower 1) the more cases have been put in categories in which the *observed* relative frequency of occurence $\bar{O}_k$ is close to zero or one, and 2) the closer the correspondence between the *forecast* relative frequency of occurrence $f_k$ and the *observed* relative frequency $\bar{O}_k$.

At this point an illustration is in order. In Fig. 1 we find an analysis of some 11,000 probability forecasts made mainly by the author in the synoptic laboratory program at MIT in 1955–56. Examination of the sorting process shows that the forecaster was frequently able to identify a large category of instances in which the event rarely happened but only a small category in which the event nearly always occurred. A study of the success in labeling shows that the forecaster underestimated the relative frequency of occurrence when the stated probability was less than 2 chances in 10 and overestimated the relative frequency when the forecast probability was greater than this amount. We are left in somewhat of a vacuum. Is this result good, bad or indifferent? These forecasts referred to a large variety of surface weather occurrences, most of which had a small climatological likelihood of occurrence. The few cases in the categories of high relative frequency of occurrence of the event, in Fig. 1, referred almost exclusively to those few events for which the climatological likelihood of occurrence was high.

### 3. Comparison with a control forecast

A more meaningful way of evaluating forecasts is to compare them with some simple control forecast, such as the climatological probability of occurrence or persistence expectancy (a conditional probability based on knowledge of occurrence or non-occurrence of the event in the most recent past interval). Let us then introduce the control probability $r$ and define the predicted departure of the forecast probability from the control value by $d\equiv f-r$, and the departure of the observed relative frequency of occurrence from the control probability by $E\equiv O-r$. Then our sample of $N$ forecasts can be split up into possibly as many as 21 categories, representing forecast departures ranging from $-1.0,\ -0.9,\ \cdots,\ 0,\ \cdots,\ +1.0$. The average score for

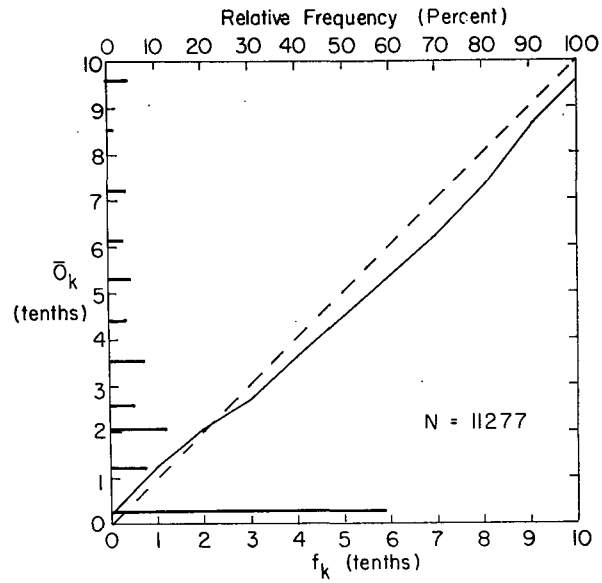


FIG. 1. Analysis of instructor's forecasts 1955–56.

the forecasts in the $k$th of these categories is

$$F_k=\frac{1}{M_k}\sum_{i=1}^{M_k}(d_k-E_{ki})^2. \tag{4}$$

The average control score for these same forecasts is simply

$$C_k=\frac{1}{M_k}\sum_{i=1}^{M_k}E_{ki}^2,$$

since here $d_k\equiv 0$. The amount of (hopeful) improvement over the control shown by these forecasts is

$$C_k-F_k=\bar{E}_k^2-(d_k-\bar{E}_k)^2. \tag{5}$$

Again, the two terms on the right side of (5) can be readily interpreted. The first, $\bar{E}_k^2$, is larger the greater the difference between the control probability and the relative frequency of occurrence of the event among the cases in this category. It vanishes when the difference vanishes. The second, $(d_k-\bar{E}_k)^2$, represents a penalty which is larger the greater the difference between the forecast probability departure and the departure of the relative frequency of occurrence. It vanishes if the difference vanishes, i.e., if the forecast probabilities are perfectly unbiased. The amount of improvement over the control for the entire sample of $N$ forecasts is obtained by averaging the mean gains (or losses) for each departure category, weighted by the number of cases in that category. Thus,

$$C-F=\overline{\overline{\bar{E}_k^2}}-\overline{\overline{(d_k-\bar{E}_k)^2}}. \tag{6}$$

As before, we see here the results of a *sorting* process in which the forecaster 1) tries to place as many cases as possible in categories in which the relative frequency of occurrence departs strongly from the control value [to

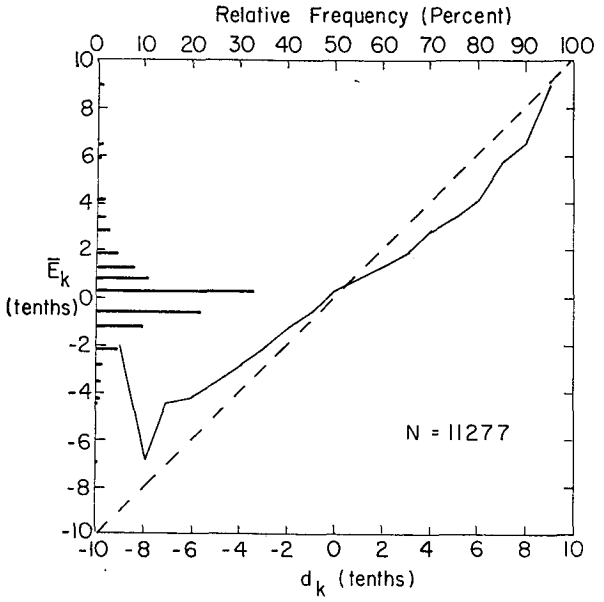


FIG. 2. Analysis of instructor's forecasts in relation to climatological control, 1955–56.

maximize $\bar{E}_k^2$], and 2) tries to put realistic *labels* on each category [to minimize $(d_k-\bar{E}_k)^2$].

By way of illustration, the forecasts in Fig. 1 were sorted according to forecast category of departure from climatological probability of occurrence and were re-evaluated. The results appear in Fig. 2. Note that few of the forecasts strayed far from the climatological value. Those that did suffered from overconfidence. That is, the magnitude of the forecast departure was systematically larger than the magnitude of the departure of the observed relative frequency from the climatological likelihood of occurrence. These forecasts were produced out of the exuberence of a decade of categorical forecasting (in which overconfidence is a sort of occupational disease) and an ignorance of what the quantitative climatological control value was. Six years later, and after an interval of underconfidence, we find the author to be rather better calibrated in a sample of some 4000 synoptic laboratory forecasts. Note in Fig. 3 that bias has been almost completely removed, except for a few apparently irresistible cases out near the fringes of the distribution. The sorting gain is definitely greater than in the preceding sample, but the forecast questions had been changed and were probably more amenable to skillful answer.

We are still left with $C-F$, a number in limbo. If this is divided by $C$ we have the percent reduction of the variance of the control score. For the 1962 sample this value is 22.4%, so we may say with respect to these particular forecast questions that we are not quite one-quarter of the way from no skill over the control to perfect skill.

## 4. Further discussion

A number of interesting aspects of forecasting can be illustrated by an analysis of a recent departmental forecast derby operated at MIT from January to September 1963. The events forecast were below-normal minimum temperatures and 0.01 inch or more precipitation at Boston for the periods 0–24, 24–48, 48–72 and 72–96 hr. Forecasts were evaluated for two forecasters, A and B, who stood near the top of the cumulative standing, for two others, C and D, who were in the middle of the pack, and for the consensus (obtained by

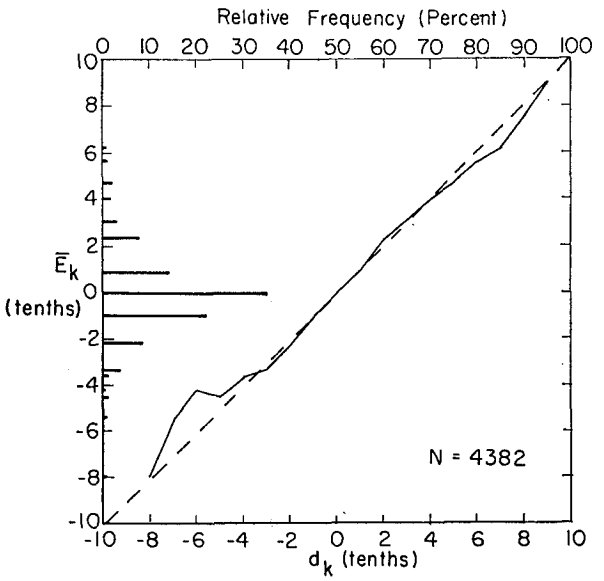


FIG. 3. Analysis of instructor's forecasts in relation to climatological control, 1962.

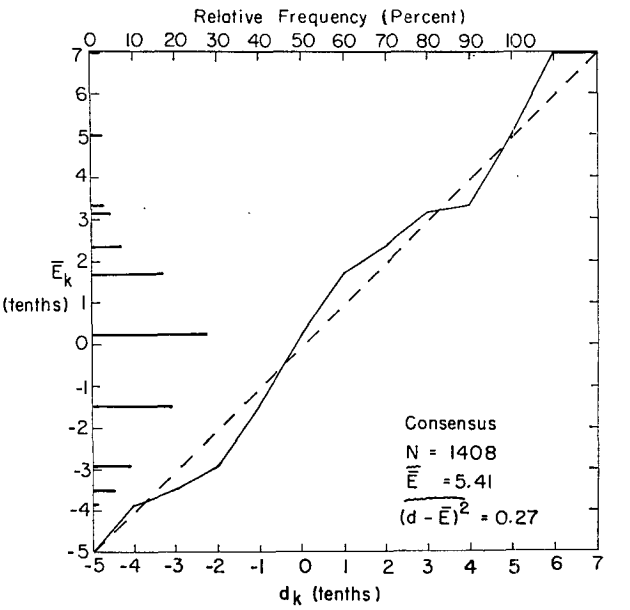


FIG. 4. Analysis of consensus forecasts in relation to climatological control, 1963.

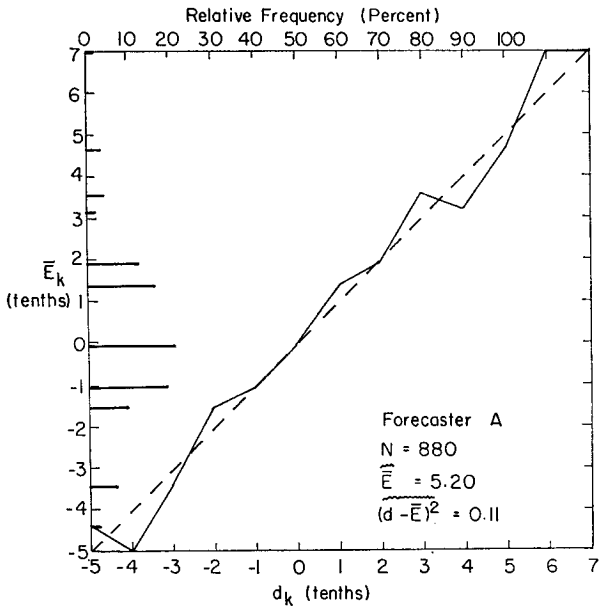


FIG. 5. Analysis of A forecasts in relation to climatological control, 1963.

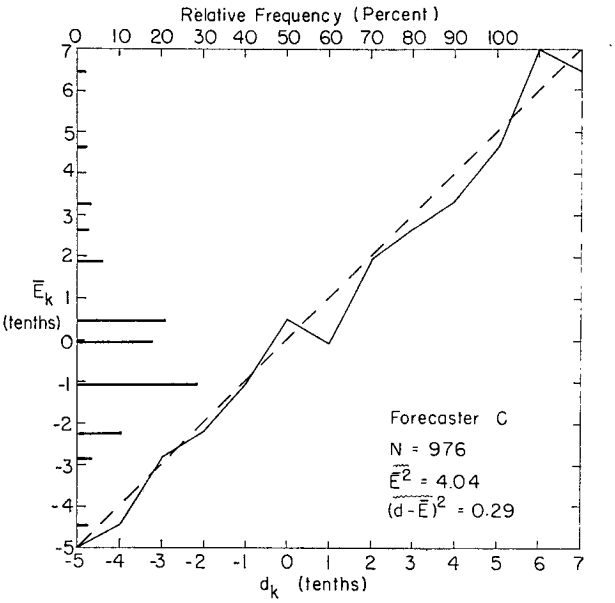


FIG. 7. Analysis of C forecasts in relation to climatological control, 1963.

averaging $f_i$, to the nearest number of chances in 10, over all participating forecasters each day). Overall results are given in Figs. 4–8 and some detail is added in Table 1. Note, for example, that consensus was the best forecaster, by virtue of superior sorting ability rather than less bias penalty. Consensus tended to be underconfident. Note that A and B, who were relatively experienced forecasters, differed more from C and D in sorting gain than in bias penalty, though both were involved. Note that A and B gained more over C and D at 72 and 96 hr than in the earlier periods, though the amounts of skill were small. Note that the bias penalty in general tended to decrease as the sample size increased and that more skill was obtained in forecasting temperature than in forecasting precipitation. It would be interesting to know what aspects of forecast performance are specific human and what aspects are common to both objective and subjective forecasts.

But we have tacitly assumed that the sorting and labeling abilities are necessary, sufficient and satis-

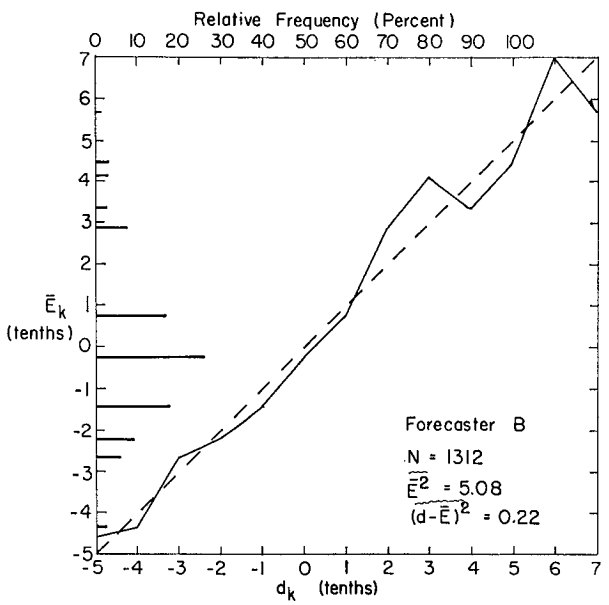


FIG. 6. Analysis of B forecasts in relation to climatological control, 1963.

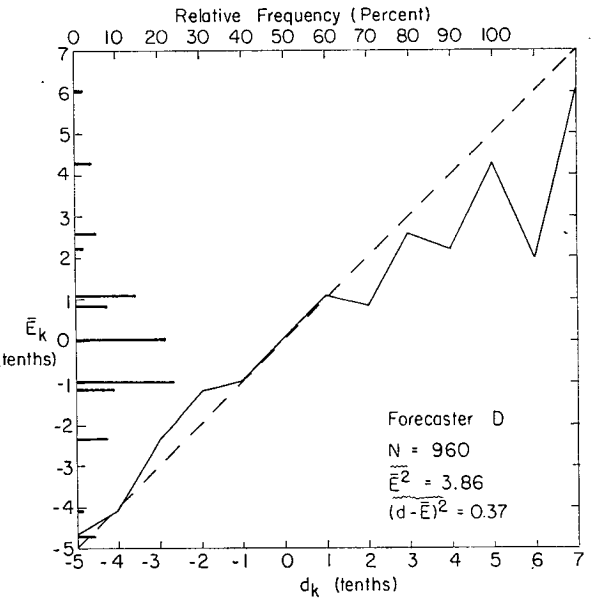


FIG. 8. Analysis of D forecasts in relation to climatological control, 1963.

TABLE 1. Summary of results of MIT probability forecasting program, January–September 1963.

| | Consensus | | | | Forecaster A | | | | Forecaster B | | | | Forecaster C | | | | Forecaster D | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | 1 | 2 | 1–2 | *N* | 1 | 2 | 1–2 | *N* | 1 | 2 | 1–2 | *N* | 1 | 2 | 1–2 | *N* | 1 | 2 | 1–2 |
| All forecasts | 1408 | 5.41 | 0.27 | +5.41 | 880 | 5.20 | 0.11 | +5.09 | 1312 | 5.08 | 0.22 | +4.86 | 976 | 4.04 | 0.29 | +3.75 | 960 | 3.86 | 0.37 | +3.49 |
| All temperatures forecasts | 704 | 6.61 | 0.47 | +6.14 | 440 | 6.20 | 0.33 | +5.87 | 656 | 6.44 | 0.36 | +6.08 | 488 | 5.48 | 0.34 | +5.14 | 480 | 4.46 | 0.59 | +3.87 |
| All precipitation forecasts | 704 | 4.66 | 0.44 | +4.22 | 440 | 4.68 | 0.44 | +4.24 | 656 | 4.02 | 0.41 | +3.61 | 488 | 2.97 | 0.64 | +2.33 | 480 | 3.51 | 0.40 | +2.91 |
| All 0–24 hr forecasts | 352 | 13.26 | 0.81 | +12.45 | 220 | 13.16 | 0.92 | +12.24 | 328 | 12.16 | 0.88 | +11.28 | 244 | 11.80 | 0.67 | +11.13 | 240 | 11.52 | 0.64 | +10.88 |
| All 24–48 hr forecasts | 352 | 6.37 | 0.52 | +5.85 | 220 | 5.83 | 0.25 | +5.58 | 328 | 5.32 | 0.54 | +4.78 | 244 | 3.86 | 1.15 | +2.71 | 240 | 4.38 | 0.96 | +3.42 |
| All 48–72 hr forecasts | 352 | 2.38 | 0.60 | +1.78 | 220 | 2.44 | 0.33 | +2.11 | 328 | 2.51 | 0.49 | +2.02 | 244 | 0.94 | 0.04 | +0.90 | 240 | 1.42 | 1.74 | −0.32 |
| All 72–96 hr forecasts | 352 | 1.08 | 0.58 | +0.50 | 220 | 1.37 | 0.65 | +0.72 | 328 | 1.07 | 0.19 | +0.88 | 244 | 0.54 | 1.31 | −0.77 | 240 | 0.70 | 0.64 | +0.06 |

$1=\overline{\overline{E}^2}$ (hundredths).
$2=\overline{\overline{(d_k-\bar{E}_k)^2}}$ (hundredths).
1–2 = Average amount of improvement per forecast over the climatological control score (hundredths).
The average climatological score per forecast for temperature forecasts is 25 hundredths, that for precipitation forecasts is approximately 21 hundredths, and that for combinations of temperature and precipitation forecasts is approximately 23 hundredths.

factory measures of scientific skill in forecasting. Satisfaction being a subjective concept, the author can only state that eight years of experience with the Brier score have left a decidedly favorable impression. If the Brier score is satisfactory, then the sorting gain and the mislabeling penalty are sufficient, since they together comprise the Brier score. We may then ask are both aspects necessary? Few would disagree with the importance of sorting ability, which, in fact, assumes the dominant position in our verification results. This leaves the bias penalty, which can be attacked on two grounds: first, is it really part of the concept of scientific accuracy, and second, does its presence inhibit forecasters from displaying their maximum sorting ability? As to the first objection, the author believes that the bias penalty is necessary as a sort of nonsense control. For example, in the Brier score as used there is never any gain nor loss compared to the control score for the forecast category of zero departure from the control probability. That is, the sorting gain *in this category* is exactly erased by the bias penalty, which is as it should be. If the penalty were removed the forecaster would receive the unintended sorting gain in this category, which he doesn't seem to deserve. In an extreme and somewhat unlikely example, if the bias penalty were removed the forecaster would get great credit for always being wrong, a rather perverse form of skill. The second objection is based upon the observation that a mere forecast of the control probability would very nearly eliminate bias, and upon the inference that the forecaster in attempting to minimize his bias penalty will therefore stick too close to the control probability. That is, he will be underconfident. This would seem then to be a mere matter of labeling. But once the sorting has been accomplished, the forecaster is hurt no less by a given amount of underconfidence than by the same amount of overconfidence. Once the forecaster understands this circumstance it is difficult to see why his probability statements should be systematically biased if his aim is to maximize the amount of gain over the control score. But perhaps the inhibition occurs during the sorting process, even though no bias is present. Perhaps the categories representing small departures from the control value contain some marginal cases which should properly have been placed in higher departure categories. But then the departure of relative frequency of occurrence of the event would probably drop in all departure categories and bias would appear. Thus, it seems that the second objection cannot be maintained.

## 5. Some possible psychological effects

In our results the forecaster near the top of the standings tended to be slightly underconfident while those below tended to be slightly overconfident. The author feels that this is a psychological effect in which the front-runners are subconsiously maximizing their likeli-

hood of staying ahead in the immediate future while those behind are similarly maximizing their likelihood of overtaking the next fellow up in the standings at the earliest possible moment. The strategies of accomplishing these aims would lead to the observed biases, and are detrimental to the score.

Finally, another source of possible bias arises due to the particular nature of our forecast questions. For example, four forecasts are made for the minimum temperature Friday morning, a 96-hr forecast made Monday morning, a 72-hr forecast made Tuesday, and so forth. Now it happens not infrequently that a change or unanticipated development on, say, Wednesday makes Monday's and Tuesday's forecasts look extremely unappetizing. In this circumstance there is a psychological pressure to recoup by reversing the direction of the forecast with a vengence. It seems not unlikely that these forecasts representing "agonizing reappraisals" are biased opposite to the direction of the earlier forecasts. Still, individuals vary, and with some in these same circumstances the subconscious desire seems to be to go down with the sinking ship, all guns blazing.

In any case the resulting overall bias seems quite small indeed and does not seriously undermine the effectiveness of the Brier score.

## 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*, Boston, Amer. Meteor. Soc., 841–848.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.

Hunt, Joseph A., 1963: Decision theory and subjective probability in meteorological forecasts. MS Thesis, Dept. of Electrical Engineering, MIT.