15.5 How Traditional NWS Verification Encourages "Hedging", and a Possible Remedy

Leslie R. Colin*
NOAA/NWS, Boise, ID

1. Introduction

Forecast verification is the process of measuring a forecast against observations of the future conditions. The verification "score" can take many forms: MAE, Brier score (Brier, 1950), percent "correct", etc. But as soon as different forecast methods come to exist, the issue of comparison among the scores arises. For example, is method A better than some simple control, or is method B? The simple way is to compare the scores of A and B and choose the better score. However, the way in which the comparison is made must be as carefully constructed as the score itself. It is one thing to forecast accurately, and another to compete against a rival method. When competition is seen as the more important issue, forecast accuracy can suffer. This is especially true when the comparison method is biased or addresses issues not considered in the original verification, such as lead-time. Some of the biases are quite subtle, and unintentional. This article shows what can happen when Brier scores, or the related rms scores, are compared among rival methods.

2. Bias in Brier-Score Comparison

Brier score is usually applied to forecasts of Probability of Precipitation (PoP). Forecast PoP (f) can take on any value from zero to 100, but observed PoP (o) can only be zero or 100. In NWS practice, Brier score (BS) is usually computed as:

$$BS = (f - o)^2 / 100 \qquad (1)$$

The goal is to minimize BS, which, in turn, implies forecasting a PoP as close to zero or 100 as possible. But since it is not completely known in advance if it will rain or not, forecasting intermediate PoPs has meaning. The point of squaring differences in the Brier formula is to emphasize the larger differences, and especially to discourage large busts when forecasting PoP.

Now consider two rival PoP forecasters. One may be human and the other may be MOS, or a model, or another human. The objectivity of MOS makes it good as a "control" rival when considering the "value" of the human forecast. In some sense, the "value" of the human forecast can be equated to how effectively it departs correctly from the MOS PoP. MOS always concedes an error if it forecasts any number other than zero or 100. Can humans do better? When MOS forecasts a PoP of 50 it concedes 25 Brier points whether it rains or not. Suppose the human forecasts a PoP of zero. If it doesn't rain, the human Brier score is zero, or 25 better than MOS. If it does rain, the human Brier score is 100, or 75 worse than MOS. Note how the penalty for a 50 PoP wrong departure is much larger than the reward for a 50 PoP correct departure.

Under these circumstances the human

* Leslie R. Colin, WFO, 3833 S.Development Ave. Bldg 3807, Boise, ID 83705-5354;
e-mail: les.colin@noaa.gov

forecaster would be foolish to forecast either zero or 100 PoP, even if he strongly felt it was the right (most accurate) forecast. By weighting the penalty more heavily than the reward, the comparison method tends to make the human hedge his own forecast.

Now consider two forecasters, A and B, competing against MOS PoP forecasts. In one case, MOS forecasts 30 PoP and forecaster A forecasts 10 PoP. It doesn't rain, so MOS gets a Brier score of 9, and forecaster A gets a Brier score of 1. Forecaster A is 8 points better than MOS. In a second case MOS makes another forecast (possibly for the very same event but using a newer model run, for example). MOS now forecasts 50 PoP and forecaster B forecasts 40 PoP. It doesn't rain, so MOS gets a Brier score of 25, and forecaster B gets a Brier score of 16. Forecaster B is 9 points better than MOS. Since it didn't rain, forecaster B's 40 PoP verifies worse than forecaster A's 10 PoP, and their respective Brier scores of 16 and 1 correctly reflect this. However, if we compare improvements vs MOS, we see that forecaster B (+9 vs MOS) is <u>better</u> than forecaster A (+8 vs MOS), even though forecaster B deviates <u>less</u> from MOS than forecaster A does. This is almost certainly not the intended message.

Over many attempts, both forecasters will learn to hedge somewhere between MOS and their true beliefs, developing their cost-benefit calculation skills instead of their forecasting skills. How can the comparison scheme be modified to correct this problem?

3. A Remedy

Consider again the previous example, but make the comparison differently. When MOS forecasts a 30 PoP it offers forecaster A an <u>opportunity</u> to score 9 Brier points better if rain doesn't occur. But forecaster A doesn't "know" this in advance. Forecaster A should only concern himself with the forecast, not the consequences of the comparison method. He chooses 10 PoP and gets a 4 point improvement over MOS (20 PoP difference squared).

When MOS forecasts 50 PoP it offers forecaster B an opportunity of 25 points whether it rains or not. Forecaster B chooses 40 PoP, only 1 point better than MOS (10 PoP difference squared), although 25 are available.

Forecaster A therefore makes the better forecast than forecaster B. Forecaster A not only verifies better than forecaster B (10 PoP vs 40 PoP), but also departs further from MOS than forecaster B and in the right sense.

What happens here is that departures from MOS, correct or incorrect, are weighted equally. The size of the penalty or reward is still squared, just as the errors themselves are squared in the Brier calculations. Squaring emphasizes the big departures, both good and bad. But the disproportionate penalty is removed, allowing the forecaster to concentrate more on the forecast and less on the comparison method.

Finally, squaring departures from MOS brings out the forecaster's confidence in his decisions. These cases are the ones where the human can add the most value to automated forecasts. Squaring also discourages the practice of accumulating points by shading MOS slightly one way or the other, which over many forecasts can obscure the real value of the fewer but significant big departures.

4. References

Brier, G.W., 1950: "Verification of Forecasts Expressed in Terms of Probability", *Mon. Wea. Rev.*,78, 1-3.