J1.5

Measuring the Decision Support Value of Probabilistic Forecasts

F. Wesley Wilson*

National Center for Atmospheric Research, Boulder, CO

1. INTRODUCTION

It is widely accepted that probabilistic weather forecasts have the potential to provide more useful information for decision support than deterministic forecasts, Hallenbeck, 1920, Sanders, 1963, Katz and Murphy, 1997. Finding metrics to quantify this comparison has proven somewhat elusive, Wilks, 2001, Jolliffe and Stephenson, 2003. More complicated still is the issue of comparing the skill and value of a probabilistic forecast system, with the value and skill of a deterministic forecast system that it is intended to replace, Gringorton, 1958. How do we quantify the improvement?

This presentation is primarily a tutorial in which several standard concepts are presented in a way that provides some insight into these issues. Standard skill measures, the Brier Score (BS) and the Peirce Skill Score (PSS) are discussed in the probabilistic and deterministic contexts. The PSS is examined from the viewpoint of Decision Theory, as a measure of skill for both deterministic and probabilistic forecast systems. Geometric interpretations are provided.

2. SKILL MEASURES

A deterministic forecast predicts the occurrence of an event, while a probabilistic forecast provides the probability that the event will occur. There is a wealth of skill scores for deterministic forecasts; listings can be found in Murphy and Daan, 1985, and Wilks, 1995.

*Corresponding author address: Wesley Wilson, NCAR, Boulder, CO 80307 email: wes@ucar.edu

An interesting history of the evolution of these score is found in Murphy, 1996. The most widely used skill score for probability forecasts is the Brier Score, Brier, 1950. The area under the ROC curve has recently received considerable attention, Mason and Graham, 2002, though more as a measure of the skill of application of the probability forecast, than of its intrinsic merit. All of this material is reviewed in Jolliffe and Stephenson, 2003.

Sanders (1963) noted that if the probability forecast is confident (issues a probability of 0 or 1) then it is effectively a deterministic forecast. Stated otherwise, a deterministic forecast is a special case of a probabilistic forecast, and can be viewed as a limiting case of probability forecasts as they become more confident. What is the behavior of a skill score for a probability forecast during this limiting process?

By wide agreement, the worst and best of the skill scores for deterministic forecasts are, respectively, the Finley Hit Rate (FHR) and the PSS, Murphy and Daan, 1985 and Murphy, 1996. The FHR was an minor issue is an otherwise significant paper by Finley, 1884. Testament that the Peirce Skill Score, Peirce, 1884, is a major contribution to forecast verification is that it has been twice rediscovered by men of substantial statistical reputation, Kuipers in 1965 and Flueck in 1987, the latter declaring it to be the True Skill Statistic.

We shall briefly review the definitions of these scores, since the formulas will be needed in the subsequent discussions. Consider a deterministic forecast (f) of a binary event, with a history of N trials. In some trials the event occurs (T) and in some it fails (F). In some trials the forecast is for occurrence (Y) and in some for non-

occurrence (N). We count the cases: A = #(Y&T), B = #(Y&F), C = #(N&T), D = #(N&F). In this notation, FHR = (A+D)/N, the sample probability of a correct forecast, both of occurrence and of non-occurrence. The PSS is most simply expressed in terms of sample conditional probabilities. We note that P(T) = (A+C)/N and P(F) = 1 – P(T). By Bayes relation, P(Y|T) = P(Y&T)/P(T) = A/(A+C) and P(Y|F) = B/(B+D). In this notation, PSS = P(Y|T) – P(Y|F), the probability of a correct forecast minus the probability of a false alarm. The PSS is a statistic if we define it intrinsically in terms of these conditional probabilities, and view the sample probability formulas as merely providing a sample statistic.

For probability forecasts, the Brier Score is defined as a quadratic expression of the events. For each trial, let f_i denote the forecasted probability and let o_i denote the observed outcome, which takes value 1 for observed occurrence and 0 for non-occurrence. Then BS = (1/N) Σ $(f_i - o_i)^2$. There is a rich documentation of the utility of the application of BS to the measure of probabilistic forecasting success. It also has a legitimate mathematical explanation: if f is viewed as a probability measure on the space of forecast trials, and o is the atomic probability measure of observed outcomes, then BS is the L_2-norm of the difference of these measures.

There is a deficiency in the interpretation of BS as the probability forecast f becomes confident. In this case, each f_i has value 0 or 1 and so the terms $(f_i - o_i)^2$ take value 1 or 0, depending on whether of not the forecast and the observation agree. Therefore, BS = (B+C)/N = 1 – FHR. If we compare the skill of a probability forecast and a deterministic forecast of the same situation, by means of BS, then we are evaluating skill by a poor measure of skill for deterministic forecasts.

We propose to resolve this dilemma by expanding the definition of the PSS to probabilistic forecasts. For deterministic forecasts, we can view the count A as the sum of the f_i when T (the event does occur): some with value 1 (forecast occurrence) and some with value 0 (forecast non-occurrence); C is the sum of contrary values. Note that A/(A+C) is the sample estimate for the average value forecasted when the event occurs, i.e. the expectation E(f|T). Similarly, B/(B+D) is a sample estimate for E(f|F). But then PSS = E(f|T) - E(f|F), and this definition applies to any probabilistic forecast. Moreover, as the forecasts become confident, the value of PSS converges to the standard skill statistic for deterministic forecasts. We propose that PSS should be considered as the common metric by which probabilistic forecasts are compared with competing deterministic forecasts. This process permits us to extend the definition of any contingency-based skill score to probabilistic forecasts.

3. DECSION THEORY

Classical Decision Theory, Van Trees, 1968, provides graphical interpretations of the roles that the conditional probabilities play in measuring the utility of forecasts for decision support. In particular, there is an illustrative interpretation of the extension of the PSS to probabilistic forecasts. The illustration is provided in Figure 1, in which two probability distributions are indicated. On the right, is the distribution of the forecasted probabilities, when the event occurred, and on the left is the distribution of probabilities when the event failed to occur. The means of these distributions are indicate by vertical dashed lines. The PSS is the distance between these means. A frequent application of probabilistic forecasts is to choose a separating threshold (τ) and to anticipate an event to occur when the forecasted probability exceeds τ. The failures of this process are measured by the erroneous tails of the distributions. When the separation (PSS) is large, we can expect a successful separation.

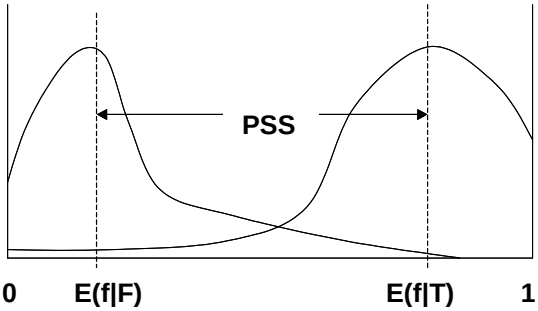

Figure 1. An illustration of PSS measuring the separation between the T and F probability distributions.

4. CONCLUSIONS

There is a need for measures that can be used to compare the skill of probabilistic forecasts with the skill of the deterministic forecasts that they are intended to replace. There is little merit in replacing a skillful deterministic forecast with a less skilled probabilistic forecast. The Brier Score has questionable value for measuring the skill of deterministic forecasts, since it provides an inferior measure of skill for deterministic forecasts. We have presented an extension of the Peirce Skill Statistic that applies consistently to probabilistic and to deterministic forecasts, and which may prove useful for making probabilistic-deterministic comparisons.

5. REFERENCES

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

Finley, J.P., 1884: Tornado Predictions, Amer. Meteor. J., **1**, 85-88.

Flueck, J.A., 1987: A study of some measures of forecast verification, 10th Conf. on Prob. And Stat. in the Atmos. Sci., Edmonton.

Gringorton, I.I.1958: On the comparison of one or more sets of probabilistic forecasts, J. Meterol , **15**, 283-287.

Hallenbeck, C. 1920: Forecasting precipitatioin in percentageds of probability, Monthly Wea. Rev., **34**, 274-275.

Jolliffe, I.T. and D.B. Stephenson, 2003: Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley, West Sussex.

Katz, R.W. and A.H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*, Cambridge University Press, Cambridge.

Mason, S.J. and N.E. Graham, 2002: Areas beneath the relative operating characteristic (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, Q> J> R. Meteorol. Soc., **128**, 2145-2166.

Murphy, A.H. and H. Daan, 1985: Forecast Evaluation, *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A.H. Murphy and R.W. Katz, ed., Westview Press, Boulder.

Murphy, A.H. 1996: The Finley affair: A signal event in the history of forecast verification, Wea. and forecasting, **11**, 3-20.

Peirce, C.S., 1884: The numerical measure of the success of predictions, Science, **4**, 453-454.

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

Van Trees, H.L., 1968: *Detection, Estimation, and Modulation Theory,* Part I, Wiley & Sons, New York.

Wilks, D.S. 1995: Statistical Methods in the Atmospheric Sciences, Academic Press, New York.

Wilks, D.S. 2001: A skill score based on economic value for probability forecasts, Meteorol. Appl., **8,** 209-219.