NOTES AND CORRESPONDENCE

Comments on “On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables”

ALLAN H. MURPHY *

UCAR Visiting Scientist Program, National Meteorological Center, National Weather Service, NOAA, Washington, D.C.

2 February 1991 and 22 March 1991

I would like to offer some comments on the interesting paper by Doswell et al. (1990) (hereafter DDK90) that appeared recently in Weather and Forecasting. The authors discuss various verification measures related to the $2 \times 2$ problem (i.e., categorical forecasts and dichotomous observations) and describe extensions and applications of these measures to problems involving probabilistic or dichotomous forecasts. My general comments are primarily concerned with issues related to these extensions and applications. In addition, I have several short comments about specific statements made—or issues raised—in the paper.

The authors state (p. 583) that “the distinction between dichotomous (i.e., ‘categorical’) and polychotomous (e.g., ‘probabilistic’) situations is illusory.” This statement is quite misleading. As DDK90 demonstrate in section 3, it is possible to verify probabilistic forecasts using methods designed for categorical forecasts. However, basic information regarding the quality of the probabilistic forecasts is neglected when such an approach is followed. For example, evaluation of probabilistic forecasts in terms of “POD-like,” “FAR-like,” etc., quantities provides little if any information regarding the reliability of the forecasts (i.e., the correspondence between forecast probabilities and observed relative frequencies). Moreover, this approach ignores the sharpness of the forecasts, as reflected by the frequency-of-use distribution over the various probability values.

It may be of interest here to note that verification problems can be characterized in terms of their dimensionality (Murphy 1991). The dimensionality ($D$) of a verification problem is the number of quantities that must be specified in order to reconstruct the joint distribution of forecasts and observations (this distribution contains all of the nontime-dependent information relevant to forecast verification in a particular situation; see Murphy and Winkler 1987). Note that $D = 3 (=2 \times 2 - 1)$ in a dichotomous situation involving categorical forecasts, whereas $D = 21 (=11 \times 2 - 1)$ in a dichotomous situation involving probabilistic forecasts with 11 distinct probability values. That is, it takes 3 numbers to describe forecasting performance completely in the former situation and 21 numbers to describe forecasting performance completely in the latter situation. Thus, although it may be reasonable to describe the quality of categorical forecasts in terms of two or three numbers (POD, FAR, etc.), this small set of numbers can provide only limited insight into the quality of probabilistic forecasts. In effect, the use of such quantities to evaluate probabilistic forecasts arbitrarily reduces the dimensionality of the underlying verification problem and necessarily results in a substantial loss of information.

DDK90 use a linear regression model to determine POD-like, FAR-like, etc., quantities for the probabilistic forecasts. They then compare these quantities with POD, FAR, etc., values calculated for categorical forecasts formulated at the same time as the probabilistic forecasts. The authors suggest that the two types of forecasts are “inconsistent” (p. 580) because the regression lines—and the values of POD, FAR, etc.—are different. Two comments are in order here. First, the meaning of the term “inconsistent” in this context is not clear, especially since the categorical forecasts evidently were not derived from the probabilistic forecasts. Second, the regression lines for the categorical and probabilistic forecasts generally would not be expected to coincide, except in the case in which the probabilistic forecasts are completely reliable and the categorical forecasts are perfect (in which case, all four lines would coincide with the 45° line). Specifically, many different sets of probabilistic forecasts (and regression lines based on such forecasts) are consistent

* Permanent affiliation: Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331.

Corresponding author address: Allan H. Murphy, Climate Analysis Center, NMC/NWS/NOAA, WWB, Room 800, Washington D.C. 20233.

© 1991 American Meteorological Society
with the same set of categorical forecasts. (Here, the term “consistent” refers to the use of a specific threshold probability value to convert the former into the latter.)

Moreover, the authors suggest (p. 581) that regression models should be used to determine POD-like, FAR-like, etc., quantities in polychotomous situations. In effect, this approach reduces a $k \times k$ ($k > 2$) verification problem to a $2 \times 2$ verification problem. Reducing the dimensionality of such problems is a worthy objective (see Murphy 1991), and the authors’ proposal warrants careful consideration from both theoretical and practical points of view. However, the dimensionality of $2 \times 2$ and $k \times k$ problems is generally quite different, and the loss of information associated with reductions in the dimensionality of $k \times k$ problems may be important from both scientific (i.e., meteorological) and economic (i.e., user-related) perspectives.

DDK90’s extension (pp. 578–579) of the TSS to polychotomous (i.e., $k \times k$) situations—and the comparison of TSS and S—warrants some comments. The TSS, originally defined for individual categories in the $2 \times 2$ problem, is the difference between POD and POFD (POFD is the probability of false detection). Although POD is readily generalized to $k \times k$ problems, it is not completely clear how to generalize POFD (POFD involves one element in the $2 \times 2$ matrix, but it involves $k - 1$ elements in the $k \times k$ matrix). Perhaps it is not surprising, then, that the authors take a different approach.

The generalization of the TSS described in the paper involves the trace of a matrix that consists of the differences between the elements in the basic verification matrix and their expected values based on random (i.e., chance) forecasts. This quantity is then divided by the trace of a matrix that consists of the differences between the elements in the matrix corresponding to perfect forecasts and their expected values. But this ratio (i.e., the ratio of the two traces) is essentially identical to the definition of S in the $k \times k$ situation. Thus, it appears that the authors’ extension of the TSS is really the polychotomous version of S in a slightly different guise.

The following specific comments address particular issues raised by the authors:

1. DDK90 identified eight measures of forecasting performance based on $2 \times 2$ contingency tables (pp. 577–578). These measures are the eight distinct conditional probabilities that can be identified when the four entries in the body of the table, representing the joint distribution of forecasts and observations, are divided by the marginal totals representing the unconditional distributions of forecasts and observations, respectively (Murphy and Winkler 1987).

2. The authors’ geometrical interpretation of the eight basic performance measures (i.e., conditional probabilities) in terms of the intercepts of the two regression lines with the unit square (p. 580) is indeed a very interesting result. It is important to recognize, however, that this interpretation is strictly applicable only in the $2 \times 2$ problem. In polychotomous situations, as well as in situations involving probabilistic forecasts, more than eight conditional probabilities can be identified. It is also of interest to note that the slopes of both regression lines correspond to specific verification measures (see comment 3 below).

3. DDK90 state (p. 580) that “we do not know of any summary statistic comparable to the TSS that is associated with $b_{st+f}$” ($b_{st+f}$ is the slope of the regression line when the observations are regressed on the forecasts). A measure identical to $b_{st+f}$ was defined by Clayton (1927) and identified as the skill percentage.

4. The authors’ brief discussion (p. 581) of the weights that should be assigned to different combinations of forecast and observed categories is interesting. It is important to recognize that all verification measures either explicitly or implicitly employ such weights. For example, in the case of the measure of accuracy underlying TSS and S, unit weight is assigned to all correct forecasts and zero weight is assigned to all incorrect forecasts. The appropriateness of these weights—and the weights associated with other skill scores—is a topic worthy of further study.

5. DDK90’s statements regarding proper scoring rules (p. 583) may engender some confusion. A strictly proper scoring rule encourages forecasters to make their forecasts correspond exactly to their judgments, by rewarding the forecasters with the best possible expected score only under this condition (Murphy and Epstein 1967; Winkler and Murphy 1968; see also Murphy and Daan 1985; Murphy and Winkler 1971). Thus, strictly proper scoring rules discourage hedging. Since the concept of strictly proper scoring rules is based on the presumption that forecasters’ true judgments are inherently probabilistic, only scoring rules defined for probabilistic forecasts can be strictly proper. Therefore, neither the TSS nor S are strictly proper.

The statement quoted by the authors (p. 576) to the effect that the TSS is a “proper formulation of a skill score” (Murphy and Daan 1985, p. 409) refers to the fact that the TSS satisfies certain desirable criteria for such skill scores. The use of the term “proper” here is not related to the use of this term in conjunction with strictly proper scoring rules. It might have been better to use a term such as “appropriate” rather than the term “proper” in this context.

6. I am in complete agreement with the authors’ statements that “no single measure of forecasting success can give a complete picture” and “many verification efforts put too much emphasis on a single score . . .” (p. 583). These statements are consistent with the perspective provided by the general framework for forecast verification described by Murphy and Winkler (1987). This perspective has led to the development of diagnostic verification (Murphy et al. 1989; Murphy
and Winkler 1991), in which particular emphasis is placed on assessing the basic characteristics of forecasting performance. Since verification problems generally involve many such characteristics, one or two overall measures are seldom adequate to describe forecasting performance in anything but the most superficial manner.

Acknowledgment. This work was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8714108.

REFERENCES


