
STATISTICAL NOTE

COMMENTS ON SOME ASPECTS AND IMPACTS OF FORECAST VERIFICATION

Allan H. Murphy

Department of Atmospheric Sciences
Oregon State University
Corvallis, Oregon 97331

I would like to offer some comments on two recent publications in the *National Weather Digest (NWD)*: (1) a paper entitled "Improving Your Weather Forecasts through a Better Knowledge of Skill Scores" by R. L. Vislocky and G. S. Young (*NWD*, 13, No. 3, pp. 15–17) (hereafter simply VY) and (2) a Letter to the Editor concerning the VY paper by R. E. Rieck (*NWD*, 14, No. 2, p. 5). These publications address various issues related to forecast verification, including the impact of verification methods on the process of translating probabilistic forecasts into categorical forecasts and the appropriateness of a single overall measure of forecasting performance. In particular, they contain some statements that are potentially misleading and/or erroneous.

The VY paper is concerned with "how to translate the event probabilities into a categorical weather forecast in such a way that the forecaster can optimize a skill score" (p. 15). This objective, when considered in conjunction with the title of the paper, may lead some readers to conclude that the probabilistic-to-categorical translation process improves forecasts from the user's point of view. Nothing could be further from the truth. This process actually destroys information—readily available and potentially useful information that could be communicated to users simply by expressing the forecasts in terms of probabilities. Failure to communicate this information adversely affects forecast credibility, which in turn amplifies existing tendencies on the part of many users to ignore the forecasts or to respond to them in an inappropriate manner. Moreover, it is relatively easy to show that reliable probabilistic forecasts are generally of greater value in an economic sense than categorical forecasts (e.g., Murphy, 1). Only those users whose cost-loss ratios correspond to the threshold probability used in the translation process attain the same economic benefits from categorical and probabilistic forecasts.

It is important to recognize that the translation process explicitly or implicitly involves assumptions about the "payoff structure" (i.e., costs and losses) of users of the forecasts. The choice of a threshold probability may be based on a typical user's cost-loss ratio—if the forecaster possesses sufficient knowledge of the primary users and uses of the forecasts (a rare situation indeed). Alternatively, and more commonly, the threshold probability is chosen to maximize (or minimize, whichever is appropriate) a verification measure whose properties are believed to be similar in some respects to the user's payoff structure; this latter rationale is exemplified by the translation procedures described by VY in their paper. Whatever rationale is adopted and whatever threshold is chosen, only those users whose cost-loss ratios are equal to the threshold probability value are well served by this translation process; others can generally expect to achieve suboptimal results.

The difficulties inherent in choosing an appropriate translation procedure are illustrated by VY's discussion of whether a precipitation probability between 30% and 50% (inclusive) should be converted into a categorical forecast of "rain" or "no rain" (pp. 16–17). In the absence of detailed knowledge of each user's payoff structure and lacking the resources to communicate suitably tailored categorical forecasts to individual users, only one rational solution exists; namely, the forecaster should communicate the best possible forecast—necessarily, a probabilistic forecast—and the user should make the best possible decision based on this forecast. When forecasters translate basic probabilities into categorical forecasts they are in effect assuming the role of user or decision maker, a role for which they are generally ill-prepared.

Verification of forecasts is actually simplified when the forecasts are expressed in a probabilistic format. The use of strictly proper scoring rules (e.g., Murphy and Daan, 2), such as the Brier score or ranked probability score, encourages forecasters to make their probabilistic forecasts correspond to their best judgments. That is, a probabilistic format provides a means of greatly reducing if not entirely eliminating the problem of "hedging." Hedging is unavoidable when forecasts are expressed in a categorical format.

VY refer to the verification measures considered in their paper as skill scores. In reality, these measures—absolute error, squared error, percent correct, and threat score—are measures of accuracy (accuracy represents the degree of correspondence between individual forecasts and observations). Skill scores are concerned with *relative* accuracy; that is, with the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by some reference forecasting procedure such as climatology or persistence (Murphy and Daan, 2).

It is disturbing to see the statement "50% probabilities . . . represent a cop out" appear in VY's paper (especially without any supplemental explanation or discussion). This statement perpetuates a serious misconception held by some forecasters (and others). In this context, the only probability that could possibly be viewed as a "cop out" (if this phrase is interpreted to mean a forecast that contains no useful information) is a forecast equal to the climatological probability. Since the climatological probability of measurable precipitation in 12-hr periods in most locations in the U.S. is considerably less than 0.50, a 50% PoP forecast generally contains potentially useful information. In any case, without prior knowledge of the nature and quality of the information on which users base their decisions in the absence of the forecasts, no PoP forecast (even a forecast equal to the climatological probability) can be said to be without value.

It is certainly true, as noted by both VY and Rieck, that forecasters should attempt to maximize (or minimize) their

verification scores. The "trick" is to design the verification system in such a way that, by following this maxim, forecasters provide information in the form of weather forecasts that faithfully reflects their best judgments and maximizes the value of the forecasts to users. In the typical situation in which a single forecast is used for many different purposes, these two objectives can be realized only by expressing the forecasts in terms of probabilities.

Rieck's call for a single overall measure of performance based on "subjective weightings" of scores associated with forecasts of individual weather elements, which appears to echo ideas set forth by Gulezian (3), is misguided. In view of forecasters' (and most meteorologists') extremely limited knowledge of the myriad of uses that are made of forecasts of different weather elements, to say nothing of their lack of quantitative information concerning the magnitudes of the costs and losses associated with these decision-making problems, how will the subjective weights be determined? Will they reflect a compromise among many different uses, thereby representing a fictitious—and possibly nonexistent—user? Is a precipitation forecast twice as important as a temperature forecast or only half as important? In fact, doesn't the relative importance of these two types of forecasts vary from user to user?

An approach such as that advocated by Rieck (and Gulezian) necessarily represents an unknown blend of the scientific and economic aspects of forecast evaluation. Forecast verification (i.e., the scientific aspects of forecast evaluation) should focus on the assessment of the various dimensions of forecast quality (e.g., accuracy, bias, reliability, etc.) for each weather element separately. Special studies could then be conducted to investigate the relationship between (various aspects of) forecast quality and forecast value for individual or truly "representative" users (e.g., Katz *et al.*, 4).

As indicated by Murphy and Winkler (5), the appropriate framework for forecast verification involving a single weather element at a specific location is the bivariate distribution of forecasts and observations for that location. If forecasters are concerned about the quality of forecasts of (for example) two weather elements simultaneously, then they must consider the multivariate probability distribution involving both types of forecasts and their respective matching observations. To date, these multivariate distributions have seldom if ever been investigated.

Rieck states that verification systems that focus on evaluating forecasts separately by weather element cause "meteorological cancer" (Snellman, 6). Little if any evidence exists to support this point of view. On the contrary, evaluation of operational PoP forecasts is based on the Brier score, a strictly proper scoring rule that encourages forecasters to make their forecasts correspond to their true judgments (see Murphy and Winkler, 7). In any case, the complete lack of "with and without" studies—that is, studies with and without the imposition of particular measures of performance (or with and without guidance forecasts)—makes it impossible to attrib-

ute this "illness" (if it exists) to any specific cause or set of causes. Moreover, it is unclear how an approach based on a single overall performance measure, whose effect on forecasts of specific weather elements—either individually or collectively—would be difficult if not impossible to fathom, could be expected to cure this (or any such) illness.

The simplest and most straightforward solution to many of the problems raised, either explicitly or implicitly, in these publications is to express weather forecasts in terms of probabilities and to verify these probabilistic forecasts using strictly proper scoring rules. This approach would encourage forecasters to make the best possible forecasts consistent with their current state of knowledge and subjective judgments. Moreover, it would eliminate the need to perform the probabilistic-to-categorical translation process, with its many tenuous assumptions and arbitrary decisions, and would return the forecasting and decision-making tasks to the individuals who possess the relevant knowledge and experience—the forecaster and the user, respectively.

In conclusion, forecast verification is an essential component of the forecasting process. If properly designed and implemented, verification systems can make a positive contribution to this process and to the post-forecasting activities of model refinement and forecast improvement. The latter activities necessarily require a coherent and diagnostic approach to forecast verification—an approach (for example) that focuses on identifying the basic strengths and weakness in forecasts of individual weather elements (e.g., Murphy *et al.*, 8). Information forthcoming from such in-depth analyses can be of significant benefit to modelers, forecasters, and users.

REFERENCES

1. Murphy, A. H., 1977: *The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation*. *Mon. Wea. Rev.*, 105, 803–816.
2. 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, Editors). Boulder, Colorado, Westview Press, pp. 379–437.
3. Gulezian, D. P., 1981: *A new verification score for public forecasts*. *Mon. Wea. Rev.*, 109, 313–323.
4. Katz, R. W., A. H. Murphy, and R. L. Winkler, 1982: *Assessing the value of frost forecasts to orchardists: a dynamic decision-making approach*. *J. of Appl. Meteor.*, 21, 518–531.
5. Murphy, A. H., and R. L. Winkler, 1987: *A general framework for forecast verification*. *Mon. Wea. Rev.*, 115, 1330–1338.
6. Snellman, L. W., 1977: *Operational forecasting using automated guidance*. *Bull. of the Amer. Meteor. Soc.* 58, 1036–1044.
7. Murphy, A. H., and R. L. Winkler, 1971: *Forecasters and probability forecasts: some current problems*. *Bull. of the Amer. Meteor. Soc.* 52, 239–247.
8. Murphy, A. H., B. G. Brown, and Y.-S. Chen, 1989: *Diagnostic verification of temperature forecasts*. *Weather and Forecasting*, 4, 485–501.