Comparative Evaluation of Categorical and Probabilistic Forecasts: Two Alternatives to the Traditional Approach

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331

1 January 1985 and 3 August 1985

ABSTRACT

Situations sometimes arise in which it is necessary to evaluate and compare the performance of categorical and probabilistic forecasts. The traditional approach to this problem involves the transformation of the probabilistic forecasts into categorical forecasts and the comparison of the two sets of forecasts in a categorical framework. This approach suffers from several serious deficiencies. Alternative approaches are proposed here that consist in (i) treating the categorical forecasts as probabilistic forecasts or (ii) replacing the categorical forecasts with primitive probabilistic forecasts. These approaches permit the sets of forecasts to be compared in a probabilistic framework and offer several important advantages vis-a-vis the traditional approach. The proposed approaches are compared and some issues related to these approaches and the overall problem itself are discussed.

1. Introduction

Comparative evaluation refers to the process of evaluating and comparing two (or more) forecasting systems or forecasters. Situations sometimes arise in this context in which it is necessary to compare the performance of categorical forecasts and probabilistic forecasts (e.g., see Bryan and Enger, 1967; Glahn, 1974; Miller and Best, 1979). The traditional approach to this problem has been to transform the probabilistic forecasts into categorical forecasts and then to evaluate both sets of forecasts in a categorical framework (i.e., using categorical verification measures). This approach suffers from several serious deficiencies, and the primary purpose of the present paper is to describe alternative approaches to the problem that offer definite advantages over the traditional approach.

The traditional approach is described in Section 2, and its principal deficiencies are summarized. Section 3 contains a description of the two alternative approaches and their perceived advantages. Further discussion of issues related to the proposed approaches and some concluding remarks appear in Section 4.

2. The traditional approach

The traditional approach to the problem of comparative evaluation of categorical and probabilistic forecasts has generally consisted of two steps: (i) transformation of the probabilistic forecasts into categorical forecasts by adhering to some procedure that identifies the "optimal" categorical forecast associated with each probabilistic forecast, and then (ii) evaluation and comparison of the quality of the two sets of categorical forecasts using verification measures and techniques appropriate for such forecasts. For convenience, these steps are referred to here as the *transformation step* and the *evaluation step*, respectively. Several different procedures have been proposed and/or used to perform the transformation step within the framework of the traditional approach. Neither the specific nature nor relative merits of these procedures are of particular concern in this paper. Nevertheless, it should be noted that several investigators (e.g., Bryan and Enger, 1967; Mason, 1979) have recognized the importance of using procedures in the transformation step that are consistent with the measures employed in the evaluation step. In particular, if the traditional approach is followed, then it would appear to be desirable to perform the transformation step in such a way as to optimize the primary measure subsequently used to evaluate the forecasts.

Clearly, the most serious deficiency in the traditional approach results from the fact that important information contained in the probabilistic forecasts is destroyed in the process of transforming these forecasts into categorical forecasts. This information consists of quantitative expressions describing the uncertainties associated with the occurrence of the relevant events. Thus, an essential feature—and inherent advantage—of probabilistic forecasts vis-a-vis categorical forecasts is necessarily completely ignored in the evaluation step when the traditional approach is adopted.

This primary deficiency can be interpreted in a different (but related) manner by recognizing that the transformation step in the traditional approach, in effect, "tailors" the probabilistic forecasts to a particular user. Thus, in the evaluation step, these forecasts are judged solely from the perspective of this individual. On the other hand, evaluation of such forecasts in a

probabilistic framework would provide a *general* assessment of their quality.

Another important deficiency in the traditional approach relates to the fact that the process of tailoring (or formulating) the two sets of categorical forecasts may not have been accomplished in the same manner. As noted above, the categorical forecasts derived from the probabilistic forecasts are obtained by following an explicit and presumably consistent procedure. However, the forecasts originally expressed in categorical terms are derived by forecasters (at least in the case of subjective forecasts) from their probabilistic judgments by an intuitive process that may vary from forecaster to forecaster, and even from situation to situation for a particular forecaster. Specifically, these forecasts may have been formulated in a manner designed to optimize a quite different quantity than that optimized in the process of transforming probabilistic forecasts into categorical forecasts. Thus, the traditional approach may, in effect, involve a comparison of forecasts tailored to different users, which in turn raises serious questions concerning the interpretation of the results of the comparative evaluation. That is, any differences in quality may be due to differences in performance, differences in transformation procedures, or a combination of these and other factors (examination of specific characteristics of the forecasts—for example, bias—may help to identify any substantial differences in transformation procedures).[1] It should be noted that this problem arises in all evaluation studies involving the comparison of the performance of two (or more) sets of categorical forecasts.

3. Two alternative approaches

a. *Treatment of categorical forecasts as probabilistic forecasts*

The key to the first of the two alternative approaches (hereafter, Alternative 1 or A1) is to recognize that a categorical forecast is simply a special case of a probabilistic forecast. Specifically, a categorical forecast assigns a probability of one to a particular event or value of the relevant variable and a probability of zero to all other events or values. For example, a categorical forecast of the event "measurable precipitation" specifies that the probability of this event is one and that the probability of the complementary event "no measurable precipitation" is zero. Recognition of this fact eliminates the need for a transformation step in the comparative evaluation process.

Since categorical forecasts represent special probabilistic forecasts, the evaluation step in the problem of concern here can be performed within a probabilistic framework (i.e., using probabilistic verification measures). For example, in the case of forecasts of the occurrence of measurable precipitation, the Brier score (Brier, 1950) could be used to verify both the categorical and probabilistic forecasts. When the forecasting systems or forecasters produce forecasts for ordinal variables involving more than two events, the ranked probability score (Epstein, 1969; Murphy, 1971) rather than the Brier score should be used to evaluate and compare the forecasts.

An obvious advantage of Alternative 1 is that it does not involve a transformation of either the probabilistic or categorical forecasts. Thus, the two types of forecasts are evaluated and compared in their original form. As a result, this approach offers an important advantage over any approach that requires a transformation of probabilistic forecasts into categorical forecasts or vice versa.

Perhaps the only disadvantage of Alternative 1 is the fact that the performance of the categorical forecasts, when evaluated using probabilistic measures of performance such as the Brier score or the ranked probability score, may appear to be relatively unsatisfactory vis-a-vis the performance of the probabilistic forecasts. This "problem" arises because the categorical forecasts are limited to two probability values, whereas the probabilistic forecasts generally involve a considerably larger number of probability values. Of course, the difference in the number of permissible probability values associated with the two types of forecasts only serves to emphasize a basic deficiency in categorical forecasts; namely, the language of categorical forecasting is extremely limited, whereas the language of probability forecasting is essentially unlimited. Nevertheless, it should be recognized that differences between the Brier (or ranked probability) scores for the categorical and probabilistic forecasts may be due in part simply to the fact that the number of possible categorical forecasts is generally considerably smaller than the number of possible probabilistic forecasts. In this regard, it may be of interest to note that the Brier score for categorical forecasts, BS_{CAT}, can be expressed as follows:

$$BS_{CAT} = 2(1 - FC),$$

where FC is the fraction of correct forecasts. Thus, categorical forecasts that are correct 85 percent of the time (i.e., for which FC = 0.85) would receive a BS_{CAT} value of 0.30.[2]

[1] It also may be possible to use evaluation techniques based on signal detection theory to distinguish between differences in transformation procedures (see Mason, 1982).

[2] The ranked probability score for categorical forecasts, RPS_{CAT}, can be expressed as follows:

$$RPS_{CAT} = \sum_{n=1}^{N} nFI_n,$$

where FI_n is the fraction of incorrect forecasts for which the absolute value of the difference between the forecast and observed *events* or values is equal to n.

b. *Transformation of categorical forecasts into probabilistic forecasts*

The second of the two alternative approaches (hereafter, Alternative 2 or A2) involves two steps: (i) transformation of the categorical forecasts into probabilistic forecasts and (ii) evaluation and comparison of the two sets of probabilistic forecasts using appropriate measures of performance. The key to Alternative 2 is to recognize that, after the categorical forecasts have been made and the corresponding observations have been obtained, it is possible to construct a contingency or verification table that summarizes the empirical joint probability distribution of the forecast and observed events in the sample (e.g., see Murphy and Daan, 1985). In particular, the verification table can be used to derive the empirical conditional probability distributions of observed events given forecast events.[3] Then each of the categorical forecasts can be replaced by the appropriate conditional probability distribution. To illustrate the application of this approach, a simple example is presented.

Consider a situation involving categorical precipitation/no precipitation forecasts. Let z_i ($i = 1, 2$) represent the two possible forecasts—$z_1 = (1, 0)$ denotes a forecast of precipitation, and $z_2 = (0, 1)$ denotes a forecast of no precipitation—and let θ_j ($j = 1, 2$) represent the two possible observations—$\theta_1 = (1, 0)$ denotes an observation of precipitation, and $\theta_2 = (0, 1)$ denotes an observation of no precipitation. Suppose that 100 forecasts and matching observations are available and that the two-by-two verification table (in frequency form) summarizing these data is depicted in Table 1. Such a table can be converted into the empirical joint probability distribution $f(z_i, \theta_j)$ ($i, j = 1, 2$) by dividing the joint frequencies by the total number of forecasts ($n = 100$), and this probability distribution is presented in Table 2. Finally, the empirical conditional probability distributions $g(\theta_j|z_i)$ can be derived from Table 2 simply by dividing $f(z_i, \theta_j)$ by the marginal or predictive probabilities $h(z_i)$, where

$$h(z_i) = \sum_{j=1}^{2} f(z_i, \theta_j) \quad (i = 1, 2).$$

These conditional distributions are presented in Table 3, and they specify the probabilities of observing the events in question given categorical forecasts of the respective events.

The empirical conditional probability distributions provide the information needed to transform the categorical forecasts into "primitive" probabilistic forecasts. Specifically, a sample of probabilistic forecasts corresponding (on a one-to-one basis) with the sample of categorical forecasts can be derived by replacing z_i with $g(\theta_j|z_i)$ ($i, j = 1, 2$). For the sample of forecasts and observations considered here, each categorical forecast $z_1 = (1, 0)$ would be replaced by $[g(\theta_1|z_1), g(\theta_2|z_1)] = (0.625, 0.375)$ and each categorical forecast $z_2 = (0, 1)$ would be replaced by $[g(\theta_1|z_2), g(\theta_2|z_2)] = (0.167, 0.833)$. The entire transformation process can, of course, be extended to include situations involving more than two events.

TABLE 1. The verification table, in the form of joint and marginal frequencies, for a hypothetical sample of categorical precipitation/no precipitation forecasts.

		Observation θ_1	Observation θ_2	Row totals
Forecast	z_1	25	15	40
	z_2	10	50	60
	Column totals	35	65	100 ($=n$)

After the categorical forecasts have been replaced by the primitive probabilistic forecasts, the two sets of forecasts can be evaluated and compared using a probabilistic measure of performance. As noted in Section 3a, the Brier score is a suitable evaluation measure for two-event probabilistic forecasts. It may be of interest to note that the Brier score for the primitive probabilistic forecasts, BS_{PP} say, can be written in terms of the empirical joint and conditional probabilities as follows:

$$BS_{PP} = \sum_{i=1}^{2} \sum_{j=1}^{2} f(z_i, \theta_j) \sum_{k=1}^{2} [g(\theta_k|z_i) - \delta_k]^2,$$

where $\delta_k = 1$ if $k = j$ and $\delta_k = 0$ otherwise. For the sample of data considered here, $BS_{PP} = 0.354$. In situations involving ordinal predictands consisting of more than two events, the ranked probability score should be used instead of the Brier score. Of course, if the two sets of forecasts have been formulated for different locations and/or time periods, then it would be necessary to use a relative rather than an absolute measure of quality in the evaluation step (in order to take any differences in climatological probabilities into account).

Alternative 2 offers several advantages vis-a-vis the traditional approach. First and foremost, it leads to an evaluation of the two forecasting systems (or forecasters) in a probabilistic framework, thereby avoiding the need to apply an information-destroying, probabilistic-to-categorical transformation procedure. In a related vein, this approach provides a general evaluation of the relevant forecasts, as opposed to an evaluation that focuses on a specific user (namely, the user characterized by the transformation procedure). In addition, al-

[3] It has been recognized for some time that verification tables can be used to derive such conditional probabilities (e.g., see Leight, 1953). Moreover, Roberts (1965) discussed the use of such tables to derive probabilistic forecasts from categorical forecasts. However, these treatments were not advanced within the framework of comparative evaluation.

TABLE 2. The empirical joint probability distribution $f(z_i, \theta_j)$ and the associated marginal distributions $h(z_i)$ and $h'(\theta_j)$ $(i, j = 1, 2)$ for the data presented in Table 1.

		Observation		
		θ_1	θ_2	$h(z_i)$
Forecast	z_1	0.25	0.15	0.40
	z_2	0.10	0.50	0.60
	$h'(\theta_j)$	0.35	0.65	

though uncertainty is ignored when forecasts are reported in categorical terms, it should be recognized that uncertainty is present nevertheless. Thus, Alternative 2 provides a means of assigning to individual categorical forecasts an *overall* measure of uncertainty based on the quality of the sample of forecasts as a whole. This process necessarily leads to an upgrading of the categorical forecasts (see Section 4).

4. Discussion and conclusion

This paper has described two alternative approaches to the problem of comparative evaluation of categorical and probabilistic forecasts, and these approaches both appear to offer important advantages over the traditional approach. The relative merits of the various approaches are briefly reviewed in this section. Moreover, several issues related to the proposed approaches and the overall problem itself are discussed.

Alternative 1 (the approach involving the treatment of categorical forecasts as probabilistic forecasts) appears relatively attractive vis-a-vis both the traditional approach and Alternative 2 (the approach involving the transformation of categorical forecasts into probabilistic forecasts). The attractiveness of A1 is due primarily to the fact that it eliminates the need for a transformation step in the comparative evaluation process. Thus, the categorical and probabilistic forecasts can be evaluated and compared in their original form within the framework of this approach. As noted in Section 3a, this fact alone appears to be sufficient to give Alternative 1 a substantial advantage over any approach that involves a transformation of either of the two types of forecasts.

Despite the strong arguments in favor of Alternative 1 set forth in the previous paragraph, some individuals may feel that this approach is "unfair" to the categorical forecasts because it evaluates them as probabilistic forecasts (see Section 3a).[4] Such individuals may be forced to choose between Alternative 2 and the traditional approach, and the relative merits of these two approaches were discussed in Section 3b. In brief, A2 involves a transformation that upgrades categorical forecasts into primitive probabilistic forecasts, whereas the traditional approach involves a transformation that downgrades probabilistic forecasts into categorical forecasts. Since the latter destroys information and the former creates and adds information (information that is implicit in categorical forecasts but remains unexpressed), Alternative 2 appears to be more attractive than the traditional approach.

Two additional issues related to Alternative 2 require some discussion. First, the primitive probabilistic forecasts introduced in Section 3b consist of conditional probabilities obtained by evaluating the sample of categorical forecasts of interest. It might be argued that the conditional probabilities should be based on a prior sample of forecasts and observations. However, since these forecasts are *expressed* in categorical terms, the need for the conditional probabilities to be available prior to the formulation of any of the forecasts is debatable. Moreover, in quantifying the uncertainties associated with the categorical forecasts, it seems more appropriate to consider the quality of the "current" sample rather than the quality of a prior sample. As a result, it appears unnecessary to postulate (or require) the existence of such a sample of forecasts and observations when Alternative 2 is adopted.

The second issue relates to the fact that A2 involves the calibration of the categorical forecasts on the basis of their performance over a sample of forecasting occasions. Thus, it might then be argued that the set of probabilistic forecasts should also be calibrated before they are evaluated and compared with the (calibrated) categorical forecasts. Calibration of the probabilistic forecasts would involve determining the empirical conditional distribution of observed events for each distinct probabilistic forecast (in a manner analogous to that described in Section 3b for the categorical forecasts). If this calibration is performed and the uncalibrated forecasts are replaced by the calibrated forecasts, then the Brier score for the latter (BS_{PC} say) will necessarily be less than or equal to the Brier score for the former (BS_P say)—that is, $BS_{PC} \leq BS_P$. Moreover, equality holds in this relationship if and only if the original forecasts are perfectly calibrated. From the point of view of consistency, it would indeed appear to be desirable to calibrate the probabilistic forecasts as well as the categorical forecasts prior to performing the evaluation step in this approach.

In conclusion, the alternative involving the treat-

TABLE 3. The empirical conditional probability distributions $g(\theta_j|z_i)$ $(i, j = 1, 2)$ for the data presented in Table 1.

		Observation	
		θ_1	θ_2
Forecast	z_1	0.625	0.375
	z_2	0.167	0.833

[4] In this regard, it is indeed true that $BS_{CAT} \geq BS_{PP}$. For example, $BS_{CAT} = 0.500$ and $BS_{PP} = 0.354$ for the data presented in Table 1.

ment of categorical forecasts as probabilistic forecasts appears to offer an (almost) ideal solution to the problem of comparative evaluation of the two types of forecasts. Specifically, this approach avoids the necessity of transforming probabilistic forecasts into categorical forecasts (or vice versa), and it allows the forecasts to be evaluated and compared using probabilistic measures of performance. Moreover, even the alternative involving the transformation of categorical forecasts into primitive probabilistic forecasts appears to offer important advantages over the traditional approach. With these two alternatives available, it should no longer be necessary to use an information-destroying procedure (such as that inherent in the traditional approach) in the process of comparative evaluation of categorical and probabilistic forecasts.

Acknowledgments. The author would like to express his appreciation to B. G. Brown, H. R. Glahn, R. W. Katz, I. B. Mason, and R. L. Winkler for comments on earlier versions of this paper. This research was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8507495.

REFERENCES

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

Bryan, J. G., and I. Enger, 1967: Use of probability forecasts to maximize various skill scores. *J. Appl. Meteor.*, **6**, 762–769.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Glahn, H. R., 1974: Problems in the use of probability forecasts. *Preprints, Fifth Conf. on Weather Forecasting and Analysis*, St. Louis, Amer. Meteor. Soc., 32–35.

Leight, W. G., 1953: The use of probability statements in extended forecasting. *Mon. Wea. Rev.*, **81**, 349–356.

Mason, I. B., 1979: On reducing probability forecasts to yes/no forecasts. *Mon. Wea. Rev.*, **107**, 207–211.

——, 1982: A model for assessment of weather forecasts. *Austral. Meteor. Mag.*, **30**, 291–303.

Miller, R. G., and D. L. Best, 1979: A model for converting probability forecasts to categorical forecasts. *Preprints, Sixth Conf. on Probability and Statistics in Atmospheric Sciences*, Banff, Amer. Meteor. Soc., 98–102.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156.

——, and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, Boulder, 379–437.

Roberts, C. F., 1965: On the use of probability statements in weather forecasts. Tech. Note 8-FCST-1, Washington, DC, ESSA, Weather Bureau, 15 pp.