What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting

ALLAN H. MURPHY

College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon

(Manuscript received 11 August 1992, in final form 20 January 1993)

ABSTRACT

Differences of opinion exist among forecasters—and between forecasters and users—regarding the meaning of the phrase "good (bad) weather forecasts." These differences of opinion are fueled by a lack of clarity and/or understanding concerning the nature of goodness in weather forecasting. This lack of clarity and understanding complicates the processes of formulating and evaluating weather forecasts and undermines their ultimate usefulness.

Three distinct types of goodness are identified in this paper: 1) the correspondence between forecasters' judgments and their forecasts (type 1 goodness, or *consistency*), 2) the correspondence between the forecasts and the matching observations (type 2 goodness, or *quality*), and 3) the incremental economic and/or other benefits realized by decision makers through the use of the forecasts (type 3 goodness, or *value*). Each type of goodness is defined and described in some detail. In addition, issues related to the measurement of consistency, quality, and value are discussed.

Relationships among the three types of goodness are also considered. It is shown by example that the level of consistency directly impacts the levels of both quality and value. Moreover, recent studies of quality/value relationships have revealed that these relationships are inherently nonlinear and may not be monotonic unless the multifaceted nature of quality is respected. Some implications of these considerations for various practices related to operational forecasting are discussed. Changes in these practices that could enhance the goodness of weather forecasts in one or more respects are identified.

1. Introduction

Statements such as "that was a good forecast" or "that was a bad forecast" are heard quite frequently, both in the meteorological community and in the community of potential users of weather forecasts. Despite their familiar ring, the meaning of such statements is seldom entirely clear. In addition to practical issues such as the way in which goodness is (or should be) evaluated and the reliability of individual perceptions of goodness, considerable ambiguity exists about what constitutes a good or bad forecast in the first place. From the forecaster's point of view, the goodness of a forecast is generally related—in one way or another—to the degree of similarity between the forecast conditions and the observed conditions. On the other hand, users are primarily concerned with whether or not a forecast leads to beneficial outcomes in the context of their respective decision-making problems. Moreover, goodness evidently possesses many different shades of meaning within each of these two communities.

Although the impacts of this lack of clarity and/or ambiguity are not well documented, they appear to be substantial and pervasive. For example, it is difficult to establish well-defined goals for any project designed to enhance forecasting performance without an unambiguous definition of what constitutes a good forecast. Moreover, it is essential that forecasters who formulate forecasts on an operational basis possess a clear understanding of the nature of goodness in weather forecasting. Otherwise, the efficiency of the forecasting process may be compromised, the effectiveness of the practice of forecast verification may be undermined, and the usefulness of the forecasts themselves may be adversely affected. For these and other reasons, clarification of the nature of goodness in this context appears to be a very worthwhile objective.

In this expository paper we identify three distinct ways in which a forecast can be good (bad). These three types of goodness can be described briefly as follows: (a) a forecast is good in the type 1 sense if it corresponds to the forecaster's best judgment derived from her (forecasters are assumed to be feminine in this paper) knowledge base; (b) a forecast is good in the type 2 sense if the forecast conditions correspond closely to the observed conditions at (or during) the valid time of the forecast; and (c) a forecast is good in the type 3 sense if the forecast, when employed by one or more users as an input into their decision-making

Corresponding author address: Dr. Allan H. Murphy, 3115 NW McKinley Dr., Corvallis, OR 97330-1139.

processes, results in incremental economic and/or other benefits. Type 2 goodness and, to a lesser extent, type 3 goodness are familiar concepts to most weather forecasters, at least in their broad outlines. However, many forecasters may not be familiar with the concept of consistency or with the nature of the relationships that exist among these three types of goodness.

The primary purposes of this paper are to describe the three types of goodness, to clarify the nature of good and bad forecasts in each sense, and to discuss the relationships among these types of goodness. Because type 1 goodness is not a familiar concept, particular attention is devoted to the development of this concept and to a discussion of the relationships between type 1 goodness and the other two types of goodness. Sections 2, 3, and 4 describe type 1, type 2, and type 3 goodness, respectively. The relationships among the three types of goodness are considered in section 5. Section 6 discusses the implications of these concepts for practices related to operational weather forecasting and identifies possible beneficial changes in these practices. This section also contains some concluding remarks.

2. Type 1 goodness: Consistency

It is assumed here that forecasters derive their forecasts concerning future weather conditions from a knowledge base. This knowledge base consists of various sources or types of information. The latter include observations and analyses of many different types; numerical, statistical, and conceptual models; the output of these models; previous forecasting experience; and feedback regarding prior forecasting performance.

Moreover, we assume that the forecasting process, as performed by a forecaster, culminates in the formulation of judgments regarding future values of weather variables or the occurrence/nonoccurrence of future weather events. These judgments are based on the forecaster's rational distillation of the information contained in her knowledge base. (As a result, they are sometimes referred to here as the forecaster's "best judgments.") The judgments are internal to the forecaster in the sense that they are by definition recorded only in the forecaster's mind. To distinguish between these internal assessments and the forecaster's external (i.e., spoken or written) statements regarding future weather conditions, the former are referred to as *judgments* and the latter are referred to as *forecasts.*

Even though the forecaster's judgments are not available for evaluation, it is reasonable to postulate that they must by their very nature satisfy certain conditions. For example, the judgments must be consistent with the current state of the art of weather forecasting, as well as with the forecaster's knowledge base on those occasions on which they are formulated. Among other things, such considerations determine the spatial and temporal specificity of the judgments. Moreover, the fact that the forecaster's knowledge base is necessarily incomplete—and imperfect in other respects—implies that the forecasting process contains an inherent element of uncertainty. The forecaster's judgments should reflect this uncertainty, which generally varies from occasion to occasion, event to event, location to location, etc. (On the other hand, we need not be concerned here with the way in which this uncertainty is characterized in the judgments. For example, it could be described qualitatively or quantitatively.) In summary, a forecaster's judgments on a particular occasion are assumed here to contain *all* of the information in her knowledge base on this occasion that relates to the future weather conditions of interest.

Since a forecaster's judgments are the result of a rational process of assimilation and distillation of the information contained in her knowledge base, it may seem reasonable to require that the forecasts—which represent the external manifestation of the judgments—correspond to the judgments. However, users of forecasts—both individually and collectively—may not require all of the information contained in the judgments. With these considerations in mind, it is useful here to introduce the concept of a *requisite forecast.* A requisite forecast contains all of the information that potential users require to act optimally in the context of their respective decision-making problems. Hereafter, we focus our attention on requisite forecasts, and the term "forecasts" should be understood to be synonymous with "requisite forecasts."

What is the appropriate relationship between a (requisite) forecast and a forecaster's judgment? Here this relationship is expressed in the form of a basic maxim of forecasting; namely, *a (requisite) forecast should always correspond to a forecaster's best judgment.* Of course, some information needed by one or more users may not be included in the forecaster's judgment (e.g., it may not be possible for a forecaster to produce a particular kind of information given the current state of the art of weather forecasting). Nevertheless, the forecast should be consistent with the information that *is* contained in the judgment. Otherwise, such a forecast would neither properly reflect the forecaster's true state of knowledge nor completely satisfy users' needs. This maxim seems quite reasonable, in the sense that the overall goal of forecasting systems presumably is to provide the best and most appropriate information available to potential users of weather forecasts.

It should be noted that the conditions that determine what constitutes a requisite forecast generally vary from user to user. Thus, a requisite forecast provided to multiple users must satisfy the union of their information requirements. To design such forecasts in a rational manner, it is necessary to obtain detailed information about the users and uses of the forecasts. Unfortunately, such information is seldom if ever readily available to forecasters or others in the operational meteorological community. For further discussion of this and other issues related to requisite forecasts, as well as trade-offs related to the content and mode of expression of forecasts, see section 6.

The concept of type 1 goodness, as set forth in this paper, is derived from the aforementioned maxim. A requisite forecast is good in the type 1 sense if the forecast corresponds to the relevant judgment, and we use the term *consistency* to describe this characteristic of forecasts. For the convenience of the reader, a short definition of type 1 goodness (and the other two types of goodness) is included in Table 1.

A (requisite) forecast can be inconsistent with the underlying judgment in several different ways. For example, it may contain more or less spatial or temporal specificity than the judgment. We will focus our attention here primarily on one particular type of inconsistency: namely, the inconsistency that arises when the uncertainty inherent in forecasters' judgments is not properly reflected in their forecasts. Since forecasters' judgments necessarily contain an element of uncertainty, their forecasts must reflect this uncertainty accurately in order to satisfy the basic maxim of forecasting. In general, then, forecasts must be expressed in probabilistic terms. (It is not necessary here to address or resolve the thorny—and sometimes controversial—issue as to whether words or numbers should be used to describe this uncertainty.) However, simply expressing a forecast in probabilistic terms does not by itself guarantee that the highest level of type 1 goodness has been achieved. In addition, the degree of uncertainty expressed in the forecast must correspond with that embodied in the relevant judgment.

For example, suppose that a forecaster's best judgment concerning the likelihood of occurrence of precipitation on a particular occasion, based on her knowledge base, is 0.2. (For concreteness, the individual is assumed to "record" her judgment in terms of a numerical probability.) If the forecaster reports a probability of 0.2, then her forecast has attained the highest level of type 1 goodness. On the other hand, suppose that the forecaster reports a probability of 0.4, because she perceives (incorrectly) that this forecast will maximize or minimize, whichever is appropriate, the value of the verification measure used to evaluate the forecast, or because of the perceived adverse impacts of issuing a forecast that would imply the continuation of a prolonged dry spell. Clearly, consistency has not attained its highest possible level in this case. Inconsistencies of a spatial or temporal nature could arise in this context if a forecaster reported a single overall precipitation probability on an occasion on which her judgment indicated that the likelihood of occurrence of precipitation varied significantly over the local forecast area or during the valid period of the forecast. In any case, it should now be obvious that expressing forecasts in a nonprobabilistic (i.e., categorical) format generally is a decidedly inferior strategy, in terms of achieving high levels of consistency. Relationships between the level of type 1 goodness and the levels of the other two types of goodness are discussed in section 5.

TABLE 1. Names and short definitions of three types of goodness.

Type	Name	Definition
1	Consistency	Correspondence between forecasts and judgments
2	Quality	Correspondence between forecasts and observations
3	Value	Incremental benefits of forecasts to users

Since a forecaster's judgments are, by definition, internal to the forecaster and unavailable for explicit evaluation (see Winkler and Murphy 1968), the degree of correspondence between judgments and forecasts cannot be assessed directly. However, various devices such as lotteries involving hypothetical bets and proper scoring rules can be used to encourage a high level of type 1 goodness, at least in the sense that the uncertainty inherent in the judgments is accurately reflected in the forecasts (Winkler and Murphy 1968). For example, strictly proper scoring rules are defined in such a way that forecasters are rewarded with the best (expected) scores if and only if their forecasts correspond with their judgments (see Murphy and Winkler 1971; Winkler and Murphy 1968). The Brier score (Brier 1950) and the ranked probability score (Epstein 1969; Murphy 1971), two common measures of the accuracy of probabilistic forecasts, are strictly proper scoring rules and thus also serve the purpose of encouraging high levels of type 1 goodness. An example of the way in which a strictly proper scoring rule promotes a high level of type 1 goodness is included in section 5a.

It is important to recognize that type 1 goodness is largely under the control of the forecaster (except for any constraints that may be imposed on the format, length, etc., of the forecasts—see section 5a). Thus, it is generally possible for forecasters to achieve very high levels of consistency simply by making their forecasts correspond with their judgments. In this sense, type 1 goodness differs from the other two types of goodness.

3. Type 2 goodness: Quality

Goodness in the type 2 sense relates to the degree of correspondence between forecasts and observations. Here, we refer to this type of goodness as *quality* (see Table 1). Thus, forecasts of high quality exhibit a close correspondence with the observations. To fully appreciate the nature of type 2 goodness and the problems associated with its measurement, it is necessary to describe briefly the current status of forecast verification, the process by which forecast quality is evaluated.

Traditionally, forecast verification has consisted of the computation of measures of the overall correspondence between forecasts and observations (e.g., see Murphy and Daan 1985; Stanski et al. 1989). Prominent examples of such measures include the mean absolute error, the mean-square error, and various skill scores. This traditional, measures-oriented approach has tended to focus on one or two overall aspects of forecast quality, such as accuracy and skill.

It is useful here to contrast the measures-oriented approach with the recently developed distributions-oriented approach. The latter is based on the notion that the joint distribution of forecasts (denoted by f) and observations (denoted by x), $p(f, x)$, contains all of the non-time-dependent information relevant to evaluating forecast quality (see Murphy and Winkler 1987). Moreover, the information contained in the joint distribution becomes more accessible when $p(f, x)$ is factored into conditional and marginal distributions. These distributions include the conditional distributions of the observations given the forecasts [$p(x\,|f)$—a conditional distribution exists for each value of f], the conditional distributions of the forecasts given the observations [$p(f\,|\,x)$—a conditional distribution exists for each value of x], the marginal distribution of the forecasts [$p(f)$], and the marginal distribution of the observations [$p(x)$]. It is the totality of the information contained in these distributions that constitutes forecast quality in its fullest sense.

The perspective provided by the distributions-oriented approach reveals that forecast quality is inherently multifaceted in nature. For example, aspects of quality generally referred to as *reliability* and *resolution* can be assessed by examining the conditional distributions $p(x\,|f)$ and the marginal distribution $p(f)$. Reliability relates to the correspondence between the mean of the observations associated with a particular forecast ($\bar{x}_f$) and that forecast (f), averaged over all forecasts. Evaluation of reliability can provide answers to the following questions: Does the mean observed temperature on those occasions on which the forecast temperature is 80°F correspond to this forecast? Is the relative frequency of precipitation on those occasions on which the precipitation probability forecast is 0.3 equal to this probability? Clearly, small differences between $\bar{x}_f$ and f are preferred to large differences. To aid the reader, short definitions of reliability and the other aspects of quality considered here are included in Table 2. This table also identifies the basic distribution(s) associated with each aspect of quality.

Resolution relates to the difference between this same conditional mean observation ($\bar{x}_f$) and the overall unconditional mean observation ($\bar{x}$), again averaged over all forecasts (see Table 2). A relevant question here might be as follows: To what extent do the conditional means of the observations corresponding to temperature forecasts of 60° and 80°F differ from each other and from the overall mean observation? In this case, large differences are preferred to small differences, since the latter indicate that, on average, different forecasts are followed by different observations.

TABLE 2. Short definitions and relevant distributions for various aspects of forecast quality.

Aspect	Definition	Relevant distribution(s)
Bias	Correspondence between mean forecast and mean observation	$p(f)$ and $p(x)$
Association	Overall strength of linear relationship between individual pairs of forecasts and observations	$p(f, x)$
Accuracy	Average correspondence between individual pairs of forecasts and observations	$p(f, x)$
Skill	Accuracy of forecasts of interest relative to accuracy of forecasts produced by standard of reference	$p(f, x)$
Reliability	Correspondence between conditional mean observation and conditioning forecast, averaged over all forecasts	$p(x\,|f)$ and $p(f)$
Resolution	Difference between conditional mean observation and unconditional mean observation, averaged over all forecasts	$p(x\,|f)$ and $p(f)$
Sharpness	Variability of forecasts as described by distribution of forecasts	$p(f)$
Discrimination 1	Correspondence between conditional mean forecast and conditioning observation, averaged over all observations	$p(f\,|x)$ and $p(x)$
Discrimination 2	Difference between conditional mean forecast and unconditional mean forecast, averaged over all observations	$p(f\,|x)$ and $p(x)$
Uncertainty	Variability of observations as described by distribution of observations	$p(x)$

The conditional distributions $p(f\,|\,x)$ provide insight into other aspects of quality, which are collectively referred to under the label *discrimination.* Roughly speaking, these aspects of quality relate to the ability of the forecasts to discriminate among the observations (see Table 2). In the case of precipitation probability forecasts, for example, discrimination is relatively strong if high probabilities are used on most occasions when precipitation occurs ($x = 1$), and low probabilities are used on most occasions when precipitation does not occur ($x = 0$). Weak discrimination would be represented by a situation in which these two conditional distributions, $p(f\,|\,x = 1)$ and $p(f\,|\,x = 0)$, largely coincide with each other. Measures of discrimination are concerned with the correspondence between the mean forecast for a particular observation ($\bar{f}_x$) and that observation (x), as well as with the difference between this same conditional mean forecast ($\bar{f}_x$) and the overall unconditional mean forecast ($\bar{f}$). Good discrimination is represented by small differences between $\bar{f}_x$ and x and by large differences between $\bar{f}_x$ and $\bar{f}$.

The marginal distribution $p(f)$, by itself, relates to the *sharpness* of the forecasts (see Table 2). In the case of precipitation probability forecasts, the forecasts are relatively sharp if forecast probabilities near zero and one are used on most occasions. On the other hand, forecast probabilities equal to the climatological probability are completely lacking in sharpness. Sharpness and resolution become identical aspects of quality when the forecasts of interest are completely reliable (i.e., when $\bar{x}_f = f$ for all f).

The marginal distribution $p(x)$ relates to the *uncertainty* associated with the forecasting situation (see Table 2). A situation in which the events are approximately equally likely is indicative of relatively high uncertainty, whereas a situation in which one or two events predominate is indicative of relatively low uncertainty. Although this aspect relates to the forecasting situation rather than to the forecasts, the level of uncertainty can have a substantial impact on other aspects of quality (e.g., skill). In this sense, uncertainty can be viewed as closely related to the concept of forecast difficulty. For further discussion of the various aspects of quality, see Murphy and Winkler (1987).

A distributions-oriented approach avoids many of the pitfalls inherent in the measures-oriented approach and provides a coherent framework for the verification process. Recently, efforts have been made to assemble a diagnostic body of methods consistent with the distributions-oriented approach to forecast verification and the multifaceted nature of forecast quality. These methods have been applied to different types of weather forecasts in two recent studies (see Murphy et al. 1989; Murphy and Winkler 1992).

To this point, we have focused on the problem of assessing the level of type 2 goodness of a single set of forecasts. The deficiencies associated with the measures-oriented approach become even more evident when the problem of comparing the quality of two (or more) sets of forecasts is confronted. Traditionally, relative forecasting performance has been assessed by comparing the magnitudes of one or two overall measures of accuracy or skill. However, it is relatively easy to show that this approach neither guarantees that the forecasts with the better score(s) are of higher quality—in all aspects of its multifaceted nature—nor ensures that the better forecasts are of greater value to all users (e.g., see Murphy and Ehrendorfer 1987).

To avoid these shortcomings, it is necessary to perform comparative evaluation within an appropriate framework. In the case of two sets of forecasts (denoted here by f and g), the basic elements of such a framework are the joint distributions $p(f, x)$ and $p(g, x)$. (For simplicity, it has been assumed here that both sets of forecasts are made for the same variable or event on the same set of forecasting occasions.) Thus, we are concerned in this context with the conditional and marginal distributions that can be obtained from factorizations of $p(f, x)$ and $p(g, x)$, as well as with the comparison of the various aspects of quality associated with these joint, conditional, and marginal distributions. Clearly, comparative evaluation is a complex problem, and it is beyond the scope of this paper to pursue these complexities in detail.

However, we might ask what general conditions must be satisfied to ensure that the forecasts f are better in all respects than the forecasts g. These conditions are embodied in a statistical relationship referred to as the *sufficiency relation,* which depends on the characteristics of $p(f \mid x)$ and $p(g \mid x)$ (see Ehrendorfer and Murphy 1988). When f's forecasts can be shown to be sufficient for g's forecasts, according to this relation, the former exhibit greater type 2 goodness than the latter in all relevant respects. Moreover, under these conditions, f's forecasts also possess greater type 3 goodness than g's forecasts. That is, all users regardless of the nature of their decision-making problems will find f's forecasts more useful than g's forecasts. Thus, the sufficiency relation can produce very powerful results. Unfortunately, it is not always possible to show that f's forecasts are sufficient for g's forecasts, or vice versa (i.e., application of the sufficiency relation may indicate that f's forecasts are not sufficient for g's forecasts and that g's forecasts are not sufficient for f's forecasts). The practical utility of this relation as a means of comparing the quality—and value—of different sets of forecasts is currently under investigation in various meteorological contexts (e.g., see Ehrendorfer and Murphy 1992; Krzysztofowicz 1992; Krzysztofowicz and Long 1991; Murphy and Ye 1990).

Unlike type 1 goodness, type 2 goodness is not entirely under the control of the forecaster. The forecaster can decide what to forecast and how to express the forecast, but the observations with which the forecasts are compared cannot be controlled. The forecaster's best strategy is to make effective use of the information contained in her knowledge base and to express the forecasts in a manner consistent with her judgments. Under these conditions, the forecasts should represent the forecaster's best possible assessments of the likelihood of occurrence of the future observations.

4. Type 3 goodness: Value

The goodness of forecasts in the type 3 sense relates to the benefits realized—or expenses incurred—by individuals or organizations who use the forecasts to guide their choices among alternative courses of action. Type 3 goodness is referred to here as *value* (see Table

1). In this section we consider the nature of forecast value and briefly discuss some basic issues related to its measurement.

First, it should be understood that forecasts possess no intrinsic value. They acquire value through their ability to influence the decisions made by users of the forecasts. Various methods are available to estimate forecast value—these methods include descriptive analyses involving studies of the behavior of weather-information-sensitive users and prescriptive analyses based on decision-analytic and/or econometric models (see Katz and Murphy 1993). Moreover, forecast value may be measured in a variety of different units. For example, it may be measured in terms of monetary benefits or expenses or in terms of nonmonetary gains or losses (e.g., lives saved or lost). For the purposes of this discussion, we will assume that forecast value is measured in (or translated into) monetary units.

In examining and describing the results of forecast-value studies, it is important to distinguish between ex post and ex ante approaches to value-of-information assessment. The ex post approach consists of determining the actual value of the forecasts after the forecasts and observations have become available. In this approach the forecasts are taken at face value, in the sense that users are assumed to base their decisions on the information as specified in the forecasts. Thus, ex post forecast-value estimates relate to the (actual) value of a set of forecasts that have been made in the past.

The ex ante approach, on the other hand, is concerned with determining the expected value of the forecasts before the forecasts and observations have become available. This approach, which is consistent with decision-analytic methods of analyzing decision-making problems (e.g., see Winkler and Murphy 1985), involves recalibration of the forecasts on the basis of the observations. That is, the decision maker is assumed to base his choice of an optimal course of action on the conditional distributions of the observations given the possible forecasts. Thus, ex ante forecast-value estimates relate to the (expected) value of a set of forecasts that may be made in the future.

From the perspective of this paper, perhaps the most important practical consequence of the differences between these two approaches relates to the value-of-information estimates themselves. In the ex post approach, forecast value can be positive or negative, with forecasts of very high quality generally realizing positive value and forecasts of very low quality possibly realizing negative value. However, in the ex ante approach, the process of recalibration transforms low-quality forecasts into high-quality forecasts. (In two-event situations, for example, forecasts that are always incorrect are translated into forecasts that are always correct.) As a result, forecast value in the ex ante approach is always nonnegative. For further discussion of the ex post and ex ante approaches to value-of-information assessment, see Murphy (1985).

The ex post approach was adopted in many early studies of the value of weather forecasts, most of which were undertaken in the context of the cost–loss ratio situation (e.g., see Thompson 1952; Thompson and Brier 1955). These studies demonstrated (inter alia) that forecasts based solely on climatological probabilities can be of greater value than relatively inaccurate nonprobabilistic forecasts. Subsequent investigations based on this approach revealed that reliable probabilistic forecasts generally are of greater value than nonprobabilistic forecasts (Murphy 1977; see also Thompson 1962).

Recently, most studies of the value of weather and/or climate forecasts have been based on an ex ante approach (see Sonka et al. 1986; Winkler and Murphy 1985; Winkler et al. 1983). Four determinants of type 3 goodness have been identified in this context (Hilton 1981): (a) the courses of action available to the decision maker, (b) the payoff structure (e.g., benefits or expenses) associated with the decision-making problem, (c) the quality of the information used as a basis for decision making in the absence of the forecasts, and (d) the quality of the forecasts. Determinants (a) and (b) relate to characteristics of the decision-making problem (and/or the decision maker). Thus, forecast value generally varies from problem to problem and from user to user within a specific problem. For example, forecast-value estimates generally differ among users who, although they rely on the same forecasts, are faced with decision-making problems that exhibit different characteristics (i.e., different sets of actions and/or different payoff structures).

The fact that forecast quality is a determinant of forecast value is hardly a surprise. (The complex relationship between quality and value is discussed in section 5c.) However, determinants (c) and (d), taken together, highlight an important but sometimes overlooked feature of forecast-value estimates. These estimates represent the *incremental* benefits realized by users when their decisions are made with the aid of forecasts. In the case of users whose payoff functions are linear in monetary benefit (or expense), these incremental benefits are measured as the difference between the users' expected payoffs when decisions are made with and without the forecasts. Thus, a single set of forecasts can lead to quite different value estimates, even in the case of two individuals faced with the same decision-making problem, if these individuals have access to different types of information in the absence of the forecasts. In effect, the availability of different nonforecast information sources means that the decision makers' forecast-value scales possess different zero points.

Several prescriptive studies of the ex ante value of weather and/or climate forecasts have been conducted in recent years. These studies have involved prototypical decision-making problems such as the cost–loss ratio situation (e.g., Katz and Murphy 1987, 1990), as well as real-world decision-making problems such as the fruit frost, fallowing-planting, corn production, and choice-of-crop situations (e.g., Brown et al. 1986; Katz et al. 1982; Mjelde et al. 1988; Wilks and Murphy 1986). Both single-stage (static) and multiple-stage (dynamic) models have been employed in these studies, and quantitative estimates of the value of short-range and long-range forecasts have been obtained in various contexts. Nevertheless, these studies have only scratched the surface of the extensive body of actual and potential users of such forecasts. Moreover, prescriptive studies of this type should be accompanied, whenever possible, by descriptive analyses (e.g., see Stewart et al. 1984), in which the information-processing and decision-making procedures of individual users are monitored and evaluated in the field. Among other things, descriptive analyses provide information that can be used to evaluate the models and assumptions on which prescriptive studies—and ex ante forecast-value estimates—are based.

Clearly, the level of type 3 goodness is not under the forecaster's control. The determinants of forecast value reveal that it is influenced by various characteristics of users' decision-making problems (i.e., courses of action, payoff structure, information available in the absence of forecasts) as well as by the level of type 2 goodness (quality). Thus, a forecaster can do no better than provide the best possible forecasts consistent with her knowledge base and judgments. Such forecasts may attain relatively high levels of type 3 goodness in the cases of some users, but other users (because of the characteristics of their decision-making problems) may find such forecasts of little or no value.

5. Relationships among consistency, quality, and value

Relationships exist among all three types of goodness. Specifically, consistency directly influences both quality and value. In addition, as already noted in section 4, forecast quality is a determinant of forecast value. The nature of these relationships is explored in this section.

a. Consistency and quality

The nature of the relationship between consistency and quality can be described by means of an example. Consider a forecaster who has arrived at a judgment p ($0 \leqslant p \leqslant 1$) regarding the occurrence of an event in a dichotomous situation (e.g., precipitation/no precipitation, frost/no frost). Here, the judgment is assumed to describe the uncertainty inherent in the forecasting process, in the sense that p represents the forecaster's best assessment regarding the likelihood of occurrence of this event, as characterized by her knowledge base. Further, suppose that the overall quality of the forecast is to be evaluated in terms of the half Brier score, BS, where

$$\mathrm{BS} = (f - x)^2, \tag{1}$$

in which f denotes the forecast ($0 \leqslant f \leqslant 1$) and x denotes the observation ($x = 1$ if the event occurs, $x = 0$ if the event does not occur). What forecast f should be given by the forecaster in order to minimize her expected score? (Recall that the BS has a negative orientation, in the sense that smaller scores generally are better. It is necessary to consider expected scores, rather than actual scores, in this context because the decision regarding the value of f must be made before it is known whether or not the event has occurred.)

If the forecaster reports a forecast f, then she will receive a score $\mathrm{BS}_1 = (f - 1)^2$ if the event occurs ($x = 1$) and a score $\mathrm{BS}_0 = f^2$ if the event does not occur ($x = 0$). Thus, the forecaster's expected score is EBS, where

$$\mathrm{EBS} = p\mathrm{BS}_1 + (1 - p)\mathrm{BS}_0, \tag{2}$$

since p and $1 - p$ represent her best judgments regarding the likelihood of obtaining the scores BS_1 and BS_0, respectively. Substituting the expressions for BS_1 and BS_0 into (2) yields

$$\mathrm{EBS} = p(f - 1)^2 + (1 - p)f^2. \tag{3}$$

What choice of f minimizes the expression for the EBS in (3)? When p^2 is added and subtracted from the right-hand side of (3), this expression can be rewritten as

$$\mathrm{EBS} = p(1 - p) + (f - p)^2. \tag{4}$$

It is clear that the EBS in (4) is minimized by choosing $f = p$; that is, by making the forecast correspond exactly with the judgment. Moreover, any choice of f for which $f \neq p$ increases the value of the EBS—a result that is undesirable from the forecaster's point of view. The fact that the choice $f = p$ minimizes the EBS demonstrates that the BS is a *strictly proper scoring rule* (see Murphy and Daan 1985; Winkler and Murphy 1968).

To provide further insight into the behavior of the EBS, it is plotted as a function of f for selected values of p in Fig. 1. As noted in the previous paragraph, the EBS attains its minimum value for a particular judgment p when the forecast f equals p. Larger—and less desirable—expected scores are obtained as the difference between f and p increases. The largest values of the EBS for $p < (>)$ ½ arise when $f = 1$ (0). Thus, forecasts that ignore the uncertainty inherent in the judgments—so-called nonprobabilistic (or categorical) forecasts—yield the largest and least desirable expected scores.

This example, which can be readily generalized to situations involving $n > 2$ events, provides a graphic illustration of the nature of the relationship between consistency and quality. Clearly, failure to maintain a high level of consistency, in the sense of ensuring that the forecasts accurately reflect the uncertainty inherent

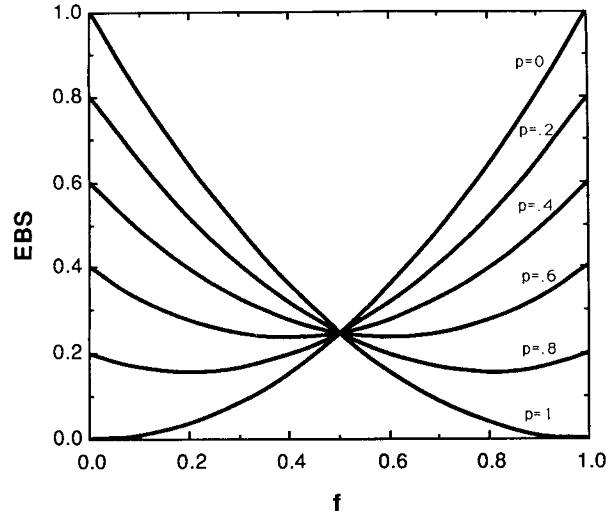

FIG. 1. The expected half Brier score EBS plotted as a function of the forecast probability f, for selected values of the judgmental probability p.

in the judgments, can adversely affect (expected) forecast quality. Assessment of the impacts of other types of inconsistencies on forecast quality, such as inconsistencies in spatial and/or temporal specificity, must await the development of procedures designed to accommodate and evaluate these inconsistencies. In any case, it should be evident that placing arbitrary restrictions on the content, format, etc., of forecasts may introduce inconsistencies that detract from their quality. The pros and cons of such restrictions need to be weighed very carefully (see section 6).

b. Consistency and value

As in the case of the relationship between type 1 and type 2 goodness, the nature of the relationship between consistency and value can be described by means of an example. Consider a decision maker in the *cost–loss ratio situation* (Thompson 1962; Murphy 1977) who must decide whether or not to protect an activity or operation in the face of uncertainty regarding the occurrence of adverse weather. If the decision maker protects, a cost of protection C is incurred and the activity is completely protected. On the other hand, if protective action is not taken and adverse weather occurs, then the decision maker suffers a loss L ($L > C$). If protective action is not taken and adverse weather does not occur, no expense (cost or loss) is incurred. Let r denote the decision maker's a priori probability of adverse weather (in which case $1 - r$ denotes the a priori probability of no adverse weather). If protective action is taken, the decision maker's expected expense is EE(protect) $= rC + (1 - r)C = C$. On the other hand, if protective action is not taken, the decision maker's expected expense is EE (do not protect) $= rL + (1 - r)0 = rL$. It is then easy to see that a decision maker who wants to minimize expected expense will protect if $r > C/L$ and will not protect if $r < C/L$ (hence the name "cost–loss ratio situation").

Now suppose that 1) a forecaster provides the decision maker with forecasts related to the occurrence/nonoccurrence of adverse weather; 2) the decision maker adopts—and uses—the forecasts as a basis for choosing between the two possible actions; and 3) attention is focused on the decision maker's expected expense from the perspective of the forecaster. (In effect, the latter amounts to an assumption that the decision maker's expected expense can be interpreted as an expected score assigned to the forecast.) Once again, we will assume that the forecaster's best judgment is denoted by p and that her forecast is denoted by f. (In this case, p and f refer to the likelihood of occurrence of adverse weather.) What value of f should be selected by the forecaster?

Suppose that $p < C/L$, and assume for the moment that the forecaster chooses to make her forecast correspond to her judgment (i.e., $f = p$). Then the expected expense is $\mathrm{EE}(f) = \mathrm{EE}(p)$, where

$$\mathrm{EE}(p) = pL + (1 - p)0 = pL. \qquad (5)$$

In fact, any choice of $f < C/L$ leads to this same expected expense. (The choice of the optimal action is based on the forecast f, but the likelihood of occurrence of the events given that action is specified by the judgment p.) In all such cases, $\mathrm{EE}(f) = \mathrm{EE}(p) = pL$. However, suppose that the forecaster chooses $f > C/L$. Then,

$$\mathrm{EE}(f) = pC + (1 - p)C = C. \qquad (6)$$

Since $p < C/L$, it follows from a comparison of (5) and (6) that $\mathrm{EE}(f) > \mathrm{EE}(p)$. That is, if the forecaster reports a value of $f > C/L$ when $p < C/L$, then the decision maker's expected expense is larger than it would have been if the forecaster had reported a value of $f < C/L$. From the forecaster's perspective, this difference in expected expense translates directly into an inferior expected score.

An analogous situation arises when $p > C/L$, and it yields similar results. In this situation, $\mathrm{EE}(f) = \mathrm{EE}(p) = C$ if $f > C/L$, but $\mathrm{EE}(f) = pL > \mathrm{EE}(p) = C$ if $f < C/L$. It is now evident that the expected expenses, interpreted as expected scores, define a *proper* (but not strictly proper) *scoring rule.* That is, the choice $f = p$ achieves the minimal expected score, but some $f \neq p$ also obtain this same expected score. These results are illustrated in a schematic diagram in Fig. 2, in which expected expense is plotted against the forecast f when $p = 0.1 < C/L = 0.3$ (Fig. 2a) and when $p = 0.5 > C/L = 0.3$ (Fig. 2b). In Fig. 2a, any choice of $f < C/L = 0.3$ leads to the same expected expense per unit loss—namely, $\mathrm{EE}^*(f) = \mathrm{EE}(f)/L = p = 0.1$—as the choice $f = p$. However, choosing $f > C/L = 0.3$ leads to a greater expected expense per unit loss—namely, $\mathrm{EE}^*(f) = C/L = 0.3$. Analogous results can

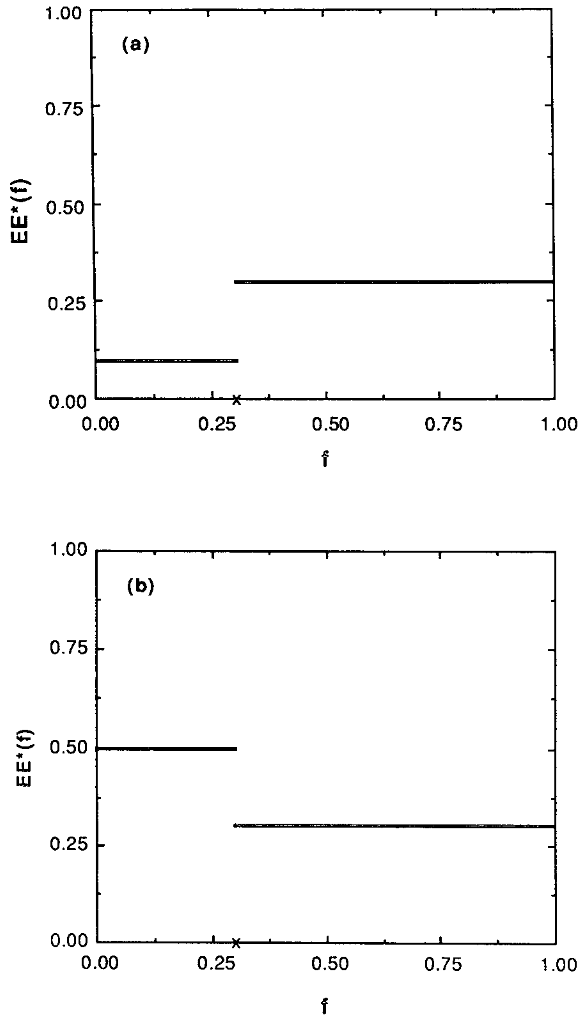

FIG. 2. Expected expense per unit loss, $EE^*(f) = EE(f)/L$, in the cost–loss ratio situation plotted as a function of the forecast probability f: (a) $p < C/L$, with $p = 0.1$, $C/L = 0.3$, $EE^*(f) = p$ if $f < C/L$, $EE^*(f) = C/L$ if $f > C/L$, and $EE^*(p) = p = 0.1$; (b) $p > C/L$, with $p = 0.5$, $C/L = 0.3$, $EE^*(f) = p$ if $f < C/L$, $EE^*(f) = C/L$ if $f > C/L$, and $EE^*(p) = C/L = 0.3$.

be gleaned from Fig. 2b; any choice of $f > C/L = 0.3$ leads to the same expected expense [$EE^*(f) = C/L = 0.3$] as the choice $f = p$, but choosing $f < C/L = 0.3$ leads to a greater expected expense [$EE^*(f) = p = 0.5$].

What do these results tell us about the relationship between consistency and value? They seem to imply that this relationship is weaker than that between consistency and quality. To minimize expected expense it is necessary only that the forecast f fall on the same side of the cost–loss ratio as the judgment p (i.e., either $f < C/L$ when $p < C/L$, or $f > C/L$ when $p > C/L$). Evidently, it is not necessary that the forecast correspond exactly to the judgment to achieve the best expected score. However, it is important to recognize that these results relate to a particular decision maker with a known cost–loss ratio. In the real world, forecasters generally possess little if any first-hand knowledge of decision makers' cost–loss ratios. Moreover, individual forecasts are frequently used by many decision makers who presumably possess different cost–loss ratios. In fact, in the absence of information to the contrary, it must be assumed that decision makers exist for all values of the cost–loss ratio ($0 < C/L < 1$). Under these circumstances, the only forecast that ensures that *no* decision maker will experience an unnecessary increase in expected expense—and that the forecaster will realize the best expected score—is the forecast $f = p$.

It is also of interest in this context to consider the consequences, for prospective users of forecasts, when a forecaster chooses a forecast $f \neq p$. The expected expenses of decision makers whose values of C/L fall outside of the interval between f and p will not be affected, since both the forecast and the judgment lead to the same decision. However, decision makers whose cost–loss ratios fall in this interval will experience increases in expected expense, because their decisions based on f would be different than their decisions based on p. This situation is illustrated in Fig. 3, in which the difference in expected expense per unit loss—namely, $\Delta EE^* = EE^*(f) - EE^*(p)$—is plotted against C/L. Note that $\Delta EE^* = 0$ outside of the interval between f and p but that $\Delta EE^* > 0$ inside this interval. The larger the difference between f and p, the more users are adversely affected and the greater the magnitude of the adverse impact. Choosing $f = 0$ or $f = 1$—extreme choices in which the uncertainty in the judgment p is ignored when the forecast is reported—maximizes the number of users who will experience increases in their expected expenses.

c. Quality and value

Since forecast quality is a determinant of forecast value (see section 4), the fact that a relationship exists between type 2 and type 3 goodness is hardly surprising. However, the nature of the concepts of quality and value dictates that the quality/value relationship is both complex and user specific. The discussion of this relationship here is limited to a brief overview of some recent results, including an example intended to illustrate that quality/value "reversals" can occur when relevant aspects of quality are overlooked.

Quality/value relationships have been investigated in detail in both prototypical and real-world situations using an ex ante approach (e.g., see Brown et al. 1986; Katz and Murphy 1990; Katz et al. 1982). These studies have shown (inter alia) that this relationship is inherently nonlinear and generally differs from situation to situation (e.g., from decision-making problem to decision-making problem, and from user to user within a specific decision-making problem). Specifically, a quality threshold exists in many situations, with the

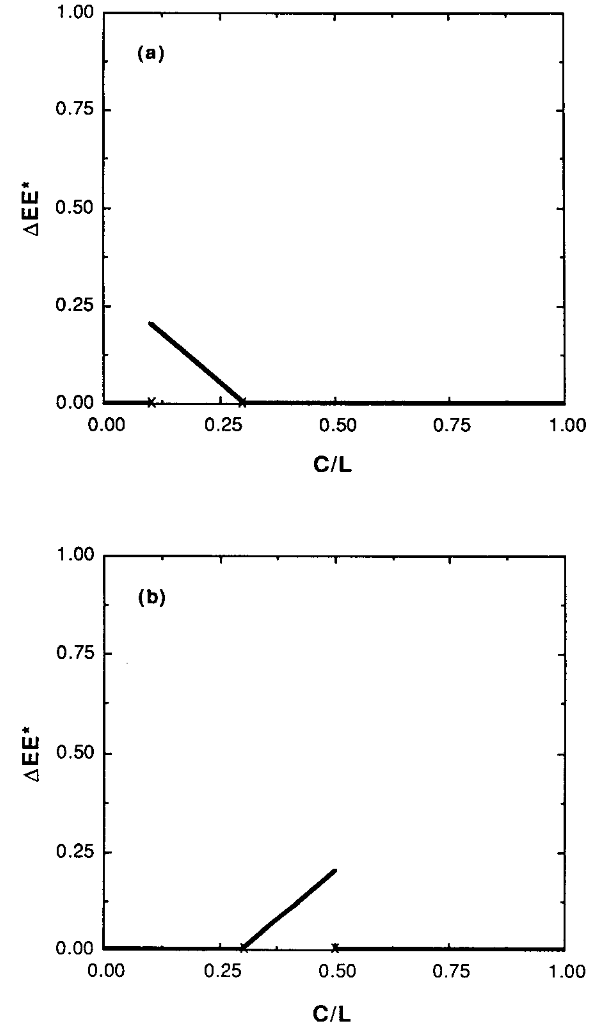

FIG. 3. Difference in expected expense per unit loss, $\Delta EE^* = EE^*(f) - EE^*(p)$, plotted as a function of the cost-loss ratio C/L: (a) $f < p$, with $f = 0.1$, $p = 0.3$, $\Delta EE^* = p - C/L$ for $0.1 < C/L < 0.3$, and $\Delta EE^* = 0$ otherwise; (b) $f > p$, with $f = 0.5$, $p = 0.3$, $\Delta EE^* = C/L - p$ for $0.3 < C/L < 0.5$, and $\Delta EE^* = 0$ otherwise.

forecasts of interest possessing no value until this level of quality is exceeded. Above this threshold, value usually increases nonlinearly as quality increases.

However, if aspects of forecast quality that influence a decision maker's choice of an optimal action are ignored in the process of measuring type 2 goodness, then even the monotonic relationship that generally is assumed to exist between quality and value may be violated. Murphy and Ehrendorfer (1987) investigated the relationship between forecast accuracy and forecast value in the context of the cost-loss ratio situation with this particular problem in mind. In their study forecast quality was measured by the expected half Brier score (EBS) and forecast value was measured by the difference in expected expense when decisions were made with and without the aid of the forecasts (VF). (In the latter case, decisions were based on climatological probabilities.)

The accuracy/value relationship in this context for a particular set of values of the basic parameters is depicted in Fig. 4. In this case, the climatological probability (π) is 0.3 and the cost-loss ratio (C/L) is 0.2. Note that the quality/value relationship is multivalued, in the sense that a range of values of VF exists for a particular value of the EBS (and vice versa). It is multivalued because the EBS, which measures forecast accuracy, does not tell the whole story. Specifically, VF depends on two aspects of quality in this situation. These aspects of quality could be measured by two conditional probabilities or by one conditional probability and one marginal probability (see section 3). However, a single measure of forecasting performance—for example, the EBS—cannot uniquely characterize two such basic aspects of quality. In other words, a multivalued relationship between the EBS and VF exists because a particular value of the EBS corresponds to many different possible combinations of values of two such probabilities and these various combinations of probability values lead to different values of VF.

The existence of a multivalued relationship allows for the possibility of quality/value reversals. To illustrate this possibility, consider two forecasting systems A and B and suppose that A's forecasts possess values of the EBS and VF equal to 0.175 and 0.030, respectively. Further, suppose that the EBS for B's forecasts is 0.150. In terms of the EBS, B's forecasts are more accurate than A's forecasts. However, VF for B's forecasts ranges from 0.000 to approximately 0.080 (see Fig. 4). Thus, depending on the characteristics of B's forecasts (as described by the conditional and/or marginal probabilities), these forecasts could be either more valuable or less valuable than A's forecasts. Moreover,

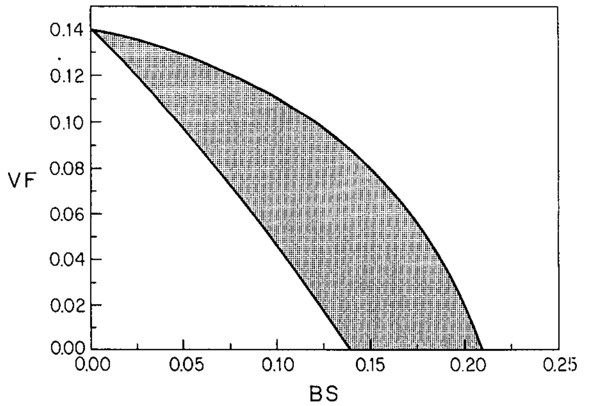

FIG. 4. Relationship between forecast accuracy and forecast value in the cost-loss ratio situation, with climatological probability $\pi = 0.3$ and cost-loss ratio $C/L = 0.2$ (taken from Murphy and Ehrendorfer 1987).

situations involving other combinations of values of π and C/L would yield similar results. In summary, it is necessary to measure forecast quality in its full dimensionality—two dimensions in this context—to ensure the existence of a monotonic quality/value relationship.

As noted in section 3, the general conditions that relative quality must satisfy to guarantee that a monotonic relationship exists between forecast quality and forecast value are embodied in the sufficiency relation (see Ehrendorfer and Murphy 1988; Krzysztofowicz and Long 1991). It should now be evident that the use of forecast quality as a surrogate for forecast value, a common practice in the meteorological community, is not justified unless these (or other equivalent) conditions are respected. Clearly, a need exists to explore quality/value relationships for weather forecasts in greater detail, both in general and in the context of specific decision-making problems.

6. Discussion and conclusion

The arguments set forth in the preceding sections of this paper clearly indicate that a forecaster's—and a user's—interests are best served by striving to attain high levels of all three types of goodness. To realize such goals, renewed attention must be given to several activities related to the forecasting process, including the formulation, evaluation, and communication of weather forecasts. In this section we briefly discuss some implications of these arguments, describe several serious deficiencies in current operational practices, and identify some potentially beneficial changes in these practices.

It seems quite evident that the concept of consistency possesses important implications for the ways in which weather forecasts are formulated and communicated to users. First and foremost, a basic maxim of forecasting is violated whenever forecasts do not correspond to judgments (see section 2). Second, the examples introduced in sections 5a and 5b demonstrate that the failure to maintain high levels of consistency leads directly to reductions in the levels of both (expected) forecast quality and forecast value. Moreover, it was shown that the widespread practice of ignoring uncertainty when formulating and communicating forecasts represents an extreme form of inconsistency and generally results in the largest possible reductions in quality and value.

Obviously, renewed efforts must be made to ensure that the uncertainty inherent in judgments is properly reflected in forecasts. Forecasts and judgments may be inconsistent in other respects as well—for example, in terms of their spatial and/or temporal specificity. In view of their possible adverse impacts on quality and value, the presence and implications of such inconsistencies warrant further investigation. In this regard, it is important to recognize that type 1 goodness is a relatively new and undeveloped concept. It will take some time before the full implications of this concept are understood and appreciated. Moreover, the concept itself may require further refinement, and methods must be developed to discourage inconsistency in all of its real-world manifestations.

Nevertheless, the nature of type 1 goodness and its impacts on type 2 and type 3 goodness raise several questions of considerable importance in an operational forecasting context. For example: What information generally contained in forecasters' judgments is not included—or is inadequately reflected—in their forecasts? What reasons are given for treating the information in this way? (To make these questions more meaningful, it is relatively easy to imagine the existence of trade-offs between consistency—or completeness—and conciseness in the process of translating judgments into forecasts.) In view of the benefits of including such information (in terms of enhanced levels of type 2 and type 3 goodness), is its current treatment justified? If not, what practical steps can be taken to facilitate the incorporation of this information into operational weather forecasts? Although these questions deserve careful consideration, they are clearly beyond the scope of the present paper.

In the case of forecast verification (the scientific aspects of forecast evaluation), the discussion of forecast quality in section 3 reveals that quality is inherently multifaceted in nature. As traditionally practiced, however, forecast verification has tended to focus on one or two aspects of overall forecasting performance such as accuracy and skill. In particular, fundamental aspects of quality that relate to the conditional distributions of the observations given forecasts and the conditional distributions of the forecasts given observations generally have been overlooked.

These practices are deficient in two important respects. First, they fail to consider all of the potentially relevant aspects of forecast quality. For example, information describing conditional aspects of quality such as reliability, resolution, and discrimination may be quite useful to forecasters who want to identify basic strengths and weaknesses in their forecasts. Identification of such basic characteristics of forecasting performance is an essential first step in the process of model refinement and forecast improvement.

Second, as noted in section 5c, various aspects of forecast quality are frequently used as surrogates for forecast value. For example, it is often assumed that increases in forecast accuracy necessarily imply increases in forecast value. However, a monotonic relationship between quality and value exists only when the multifaceted nature of quality is respected in the measurement process. To ensure that all of the potentially relevant aspects of forecast quality are considered, it is necessary to take a distributions-oriented approach to forecast verification. (In effect, whatever body of methods is used to assess forecast quality, they must be equivalent to the methodology underlying the sufficiency relation.) In some cases, it may be possible to reduce the complexity and dimensionality of the most general distributions-oriented approach (see Murphy 1991). However, arbitrary reductions in complexity and dimensionality, such as those explicitly or implicitly made in conjunction with current verification practices, frequently lead to incomplete and/or misleading results. To make rational choices regarding reductions in the complexity or dimensionality of verification problems, detailed information is needed about the users of the forecasts and about their decision-making problems.

In this regard, it is important to recognize that forecasters—and others in the operational meteorological community—routinely choose among alternative forecasting methods or models, alternative modes of presentation of forecast information, alternative channels of communication of forecasts, etc. (as well as among alternative methods of measuring forecasting performance). Despite the fact that these choices directly affect the content and quality of forecasts, the information requirements and decision-making problems of potential users are seldom considered explicitly in making such choices. In particular, little if any effort is generally expended to acquire specific user-related information regarding the nature of these requirements and problems, information that is essential if forecasters are to make rational decisions regarding forecast formulation and communication. Clearly, no weather forecasting system can achieve the highest possible levels of type 3 goodness unless forecasters acquire such user-specific information and then make judicious use of the information to guide their choices among alternative methods of formulating, evaluating, and communicating the relevant forecasts.

This paper was motivated in part by the differences of opinion that exist among forecasters—and between forecasters and users—regarding the meaning of the phrase "good (bad) weather forecasts" (see section 1). Now that the three basic types of goodness have been identified and described, it seems appropriate to ask the following question: To what extent are such differences of opinion warranted? In the case of the forecasters, these differences of opinion appear to be largely unwarranted. For the most part, they exist because forecasters (and others in the meteorological community) do not fully appreciate the multifaceted nature of forecast quality. Attention is generally focused on one or two aspects of quality such as accuracy and skill. (The arguments that do occur usually relate to which measures of these aspects are appropriate rather than which aspects of quality should be measured.) However, as the discussion in section 3 indicates, measures of accuracy or skill do not and cannot tell the whole story regarding forecasting performance. This story can be told only by the joint distribution of forecasts and observations, or equivalently by conditional and marginal distributions derived from this joint distribution. Hopefully, differences of opinion among forecasters regarding forecasting performance will decrease over time as the holistic distributions-oriented approach gains new adherents among members of the operational meteorological community.

With regard to differences of opinion between forecasters and users (and among users), it should be evident from the discussion of forecast value in section 4 that such differences of opinion are quite likely to be warranted in most instances. For example, forecasts of relatively high quality (according to a distributions-oriented evaluation of forecasting performance) may be of little or no value to some users, because of the nature of their decision-making problems or because of the relatively high quality of their prior information. Moreover, users facing different decision-making problems may find the same set of forecasts of quite different value. In effect, forecasters and users generally use fundamentally different methods of evaluating forecasts. Nevertheless, the differences between their respective assessments of goodness in weather forecasting would be less pronounced if forecasters adopted a distributions-oriented approach to forecast evaluation. Moreover, it would be beneficial for members of both communities to gain a greater understanding and familiarity with the ways in which their "opposite" numbers evaluate forecasting performance.

Finally, since consistency may be a relatively new concept to many forecasters, the links between consistency, quality, and value deserve special emphasis here. The arguments and examples set forth in this paper have demonstrated that type 1 goodness directly impacts both type 2 and type 3 goodness. Thus, at any point in time, the highest possible levels of forecast quality and forecast value cannot be achieved without attaining the highest possible level of consistency. In the context of subjective weather forecasting, it would seem that the correspondence between forecasts and judgments is an important concept that warrants considerably more attention than it has received heretofore.

Acknowledgments. Edward Epstein provided insightful comments that materially strengthened the discussion of type 1 goodness, and his comments are gratefully acknowledged. Barbara Brown, Robert Clemen, Charles Doswell, Frederick Sanders, Daniel Wilks, Robert Winkler, and two anonymous referees provided helpful comments on earlier versions of the manuscript. This work was supported in part by the National Science Foundation under Grant SES-9106440.

REFERENCES

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

Brown, B. G., R. W. Katz, and A. H. Murphy, 1986: On the economic value of seasonal-precipitation forecasts: The fallowing/planting problem. *Bull. Amer. Meteor. Soc.*, **67**, 833–841.

Ehrendorfer, M., and A. H. Murphy, 1988: Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Mon. Wea. Rev.,* **116,** 1757–1770.

——, and ——, 1992: Evaluation of prototypical climate forecasts: The sufficiency relation. *J. Climate,* **5,** 876–887.

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

Hilton, R. W., 1981: The determinants of information value: Synthesizing some general results. *Manage. Sci.,* **27,** 57–64.

Katz, R. W., and A. H. Murphy, 1987: Quality/value relationship for imperfect information in the umbrella problem. *Am. Stat.,* **41,** 187–189.

——, and ——, 1990: Quality/value relationships for imperfect weather forecasts in a prototype multistage decision-making model. *J. Forecasting,* **9,** 75–86.

——, and ——, 1993: *Economic Value of Weather and Climate Forecasts.* Cambridge University Press, in press.

——, ——, and R. L. Winkler, 1982: Assessing the value of frost forecasts to orchardists: A dynamic decision-making approach. *J. Appl. Meteor.,* **21,** 518–531.

Krzysztofowicz, R., 1992: Bayesian correlation score: A utilitarian measure of forecast skill. *Mon. Wea. Rev.,* **120,** 208–219.

——, and D. Long, 1991: Forecast sufficiency characteristic: Construction and application. *Int. J. Forecasting,* **7,** 39–45.

Mjelde, J. W., B. L. Dixon, S. T. Sonka, and P. J. Lamb, 1988: Valuing forecast characteristics in a dynamic agricultural production system. *Amer. J. Agric. Econ.,* **70,** 674–684.

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

——, 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.,* **105,** 803–816.

——, 1985: Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Mon. Wea. Rev.,* **113,** 362–369.

——, 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

——, and R. L. Winkler, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.,* **52,** 239–247.

——, 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, 379–437.

——, and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Wea. Forecasting,* **2,** 243–251.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

——, and Q. Ye, 1990: Comparison of objective and subjective precipitation probability forecasts: The sufficiency relation. *Mon. Wea. Rev.,* **118,** 1783–1792.

——, and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

——, B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4,** 485–501.

Sonka, S. T., P. J. Lamb, S. E. Hollinger, and J. W. Mjelde, 1986: Economic use of weather and climate information: Concepts and an agricultural example. *J. Climatol.,* **6,** 447–457.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification measures in meteorology. Downsview, Ontario, Canada, Atmospheric Environment Service, Research Report No. 89-5, 113 pp.

Stewart, T. R., R. W. Katz, and A. H. Murphy, 1984: Value of weather information: A descriptive study of the fruit-frost problem. *Bull. Amer. Meteor. Soc.,* **65,** 126–137.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.,* **33,** 223–226.

——, 1962: Economic gains from scientific advances and operational improvements in meteorological prediction. *J. Appl. Meteor.,* **1,** 13–17.

——, and G. W. Brier, 1955: The economic utility of weather forecasts. *Mon. Wea. Rev.,* **83,** 249–254.

Wilks, D. S., and A. H. Murphy, 1986: A decision-analytic study of the joint value of seasonal precipitation and temperature forecasts in a choice-of-crop problem. *Atmos.-Ocean,* **24,** 353–368.

Winkler, R. L., and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.,* **7,** 751–758.

——, and ——, 1985: Decision analysis. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 493–524.

——, A. H. Murphy, and R. W. Katz, 1983: The value of climate information: A decision-analytic approach. *J. Climatol.,* **3,** 187–197.