A Coherent Method of Stratification within a General Framework for Forecast Verification

ALLAN H. MURPHY

Prediction and Evaluation Systems, Corvallis, Oregon

1 April 1994 and 18 October 1994

ABSTRACT

The general framework for forecast verification described by Murphy and Winkler embodies a statistical approach to the problem of assessing the quality of forecasts. This framework is based on the joint distribution of forecasts and observations, together with conditional and marginal distributions derived from decompositions of the underlying joint distribution. An augmented version of the original framework is outlined in this paper. The extended framework provides a coherent method of addressing the problem of stratification in this context and it can be used to assess forecast quality—and its various aspects—under specific meteorological conditions. Conceptual examples are presented to illustrate potential applications of this methodological framework. Some issues concerning the extended framework and its application to real-world verification problems are discussed briefly.

1. Introduction

A general framework for forecast verification has been described by Murphy and Winkler (1987, hereafter MW87). The foundation of this framework is the joint distribution of forecasts and observations that (assuming statistical stationarity) contains all of the nontime-dependent information relevant to the problem of assessing the quality of the forecasts of interest. To obtain insight into basic characteristics (or aspects) of forecast quality, it is useful to decompose this joint distribution into conditional and marginal distributions. Consideration of the problem of forecast verification from the perspective provided by these joint, conditional, and marginal distributions leads to the identification of a distributions-oriented body of verification methods. Applications of these methods to short-range weather forecasts have been reported by Murphy et al. (1989) and Murphy and Winkler (1992).

Distributions-oriented verification methods, as described in the published literature, have focused to date on the *statistical* characteristics of the forecasts, the observations, and their relationship, as embodied in the underlying joint, conditional, and marginal distributions. Not surprisingly, however, various questions of a *meteorological* nature arise when these methods (or any verification methods) are applied. For example, what differences exist between the various aspects of forecast quality under different weather regimes? Are forecasts more reliable (accurate, skillful, . . .) under some weather conditions than under other weather conditions, in one season than in another season, etc.? Are forecasts of changes in the weather more successful given some initial conditions than given other initial conditions? To answer these and other similar questions, it is necessary to stratify the results of verification programs (studies, exercises, etc.) on the basis of a relevant set of meteorological conditions. The following basic question arises in this context: Can the general framework for forecast verification be extended in a way that such stratifications can be accommodated?

The primary purpose of this paper is to outline—and illustrate conceptually—an extended version of the original framework that allows the requisite stratifications to be performed in a natural and coherent manner. This extension is based on a fundamental concept in elementary probability theory, which is known as *extending the argument* (e.g., O'Hagan 1988). In this context, extending the argument involves (a) introducing a variable (or covariate) that describes completely and unambiguously the meteorological conditions of interest and (b) conditioning the underlying joint distribution on the basis of the values of this covariate. As a result, the overall (i.e., unconditional) joint distribution is decomposed into a set of conditional joint distributions, each of which describes forecast quality completely under specific meteorological conditions, and a univariate distribution specifying the probabilities of these mutually exclusive and collectively exhaustive conditions.

To obtain insight into the basic aspects of forecast quality under such conditions, each conditional joint

Corresponding author address: Dr. Allan H. Murphy, Prediction and Evaluation Systems, 3115 NW McKinley Drive, Corvallis, OR 97330-1139.

distribution can be decomposed into univariate distributions of two types: (a) univariate distributions conditional on the meteorological conditions *and* either a distinct forecast or a distinct observation and (b) univariate distributions conditional only on the meteorological conditions. This second level of decomposition is analogous to that described in MW87, except that the joint distributions of interest already are conditioned on the values of a meteorological covariate. Thus, the concept of extending the argument provides a coherent (i.e., logically complete) approach to the problem of stratifying the results of distributions-oriented verification programs on the basis of meteorological considerations.

Section 2 introduces the concept of extending the argument using simple examples. The general framework for forecast verification and its extended version—which embodies stratification—are described in section 3. Potential applications of the extended framework are illustrated using conceptual examples in section 4. Section 5 contains a discussion of various issues concerning the framework itself as well as its application to real-world verification problems, and section 6 consists of a brief summary and some concluding remarks.

2. The concept of extending the argument

The concept of extending the argument is simply the concept of conditional probability in a different guise and with a specific purpose in mind. O'Hagan (1988, 45–51), for example, describes this concept succinctly in the context of subjective probability measurement or assessment. Difficulties are sometimes encountered in assessing the *unconditional* probability of an event directly in this context. In such situations, it is frequently useful to introduce a covariate (i.e., a variable related to the event of interest) and to assess the conditional probability of the event given each value (or range of values) of the covariate and the unconditional probabilities of the values of the covariate. These conditional and unconditional probabilities can then be combined according to a basic probability law to obtain the desired probability.

To make this discussion more concrete, consider a situation in which E denotes an event for which the probability must be determined. Further, let Z denote a covariate, which (for simplicity) can take on only two distinct values z_1 and z_2. Then the probability of E, $\Pr(E)$, can be expressed as follows:

$$\Pr(E) = \Pr(E|z_1)\Pr(z_1) + \Pr(E|z_2)\Pr(z_2), \quad (1)$$

where $\Pr(E|z_1)$ is the conditional probability of E given $Z = z_1$, $\Pr(E|z_2)$ is the conditional probability of E given $Z = z_2$, $\Pr(z_1)$ is the unconditional (or marginal) probability that $Z = z_1$, and $\Pr(z_2)$ [$=1 - \Pr(z_1)$] is the unconditional (or marginal) probability that $Z = z_2$.

Consider a simple meteorological example involving the application of this concept. Let E denote the event "measurable precipitation in Corvallis tomorrow afternoon." Rather than assessing $\Pr(E)$ directly, a forecaster might find it easier to assess the conditional probability of E given that weather regime z_j prevails tomorrow afternoon and the unconditional probability that regime z_j will indeed prevail ($j = 1, 2$). The probability $\Pr(E)$ can then be reconstructed from these conditional and unconditional probabilities by means of (1), which represents a two-term version of the law of total probability.

The decomposition in (1) can be readily extended to a situation in which the covariate Z is defined in terms of an m-fold partition. In this case,

$$\Pr(E) = \sum_j \Pr(E|z_j)\Pr(z_j), \quad (j = 1, \cdots, m), \quad (2)$$

where $\Pr(E|z_j)$ is the conditional probability of E given $Z = z_j$, and $\Pr(z_j)$ is the unconditional probability that $Z = z_j$. Here the m values of the covariate Z might represent a set of m mutually exclusive and collectively exhaustive (m.e.c.e.) weather regimes.

Analogous decompositions can be defined in situations involving probability distributions. Suppose that instead of the probability of an event E, a probability distribution for a (discrete or continuous) variable Y, $g(y)$, must be determined. Then, for a covariate Z defined in terms of an m-fold partition, it follows that

$$g(y) = \sum_j g(y|z_j)\Pr(z_j), \quad (j = 1, \cdots, m), \quad (3)$$

where $g(y|z_j)$ is the conditional distribution of Y given $Z = z_j$, and $\Pr(z_j)$ is once again the unconditional probability that $Z = z_j$. Similar decompositions can be defined in situations in which the distribution of interest is a bivariate or multivariate distribution.

3. Extended version of general framework

a. Basic framework

As noted in section 1, the Murphy–Winkler general framework for forecast verification is based on the joint distribution of forecasts and observations (see MW87). If F denotes the forecast and X denotes the observation, then this distribution can be written as $p(f, x)$, where f represents a generic forecast and x represents a generic observation. This distribution specifies the probability that $F = f$ *and* $X = x$ for all possible combinations of f and x and, in practice, it would be estimated by means of the joint relative frequencies of the various combinations of forecasts and observations in a verification data sample.

To provide insight into basic aspects of forecast quality, it is useful to decompose $p(f, x)$ into conditional and marginal distributions. Two such decompositions can be defined:

$$p(f, x) = p(x|f)p(f) \quad (4)$$

and

$$p(f, x) = p(f|x)p(x) \quad (5)$$

(see MW87). Here $p(x|f)$ represents the conditional distribution of X given $F = f$, $p(f|x)$ represents the conditional distribution of F given $X = x$, $p(f)$ represents the marginal distribution of F, and $p(x)$ represents the marginal distribution of X. It should be noted that a conditional distribution $p(x|f)$ in (4) and $p(f|x)$ in (5) exists for each possible forecast f and observation x, respectively. These conditional distributions play particularly important roles in the distributions-oriented approach to forecast verification because they characterize the relationship between the forecasts and observations.

A "complete" approach to forecast verification can be based on $p(f, x)$, on $p(x|f)$ and $p(f)$, or on $p(f|x)$ and $p(x)$. Nevertheless, it is useful in general to examine all of these distributions because they provide insight into different aspects of forecast quality (see MW87). Moreover, misleading or erroneous conclusions may be drawn concerning the absolute and/or relative quality (and value) of forecasts if an approach is adopted that fails to respect the full dimensionality of a verification problem, as defined by the underlying joint, conditional, and marginal distributions (see Murphy 1991; Murphy and Ehrendorfer 1994).

With these distributions in mind, it is possible to assemble a distributions-oriented body of verification methods that constitutes a structured approach to the problem of assessing absolute or relative forecast quality. Three classes of distributions-oriented methods can be identified: 1) the underlying joint, conditional, and marginal distributions themselves; 2) summary measures of these distributions (means, variances, etc.); and 3) performance measures (i.e., measures of various aspects of the relationship between F and X). Applications of tailored versions of this body of methods to verification problems involving nonprobabilistic temperature forecasts and precipitation probability forecasts have been described in detail by Murphy et al. (1989) and Murphy and Winkler (1992), respectively.

b. *Extended framework*

The augmented version of the original framework, based on the extending-the-argument concept, is formulated here by introducing a discrete covariate Z. Specifically, the variable Z is assumed to possess m m.e.c.e. values or states. The overall joint distribution of F and X, $p(f, x)$, can then be decomposed, by appealing to the extending-the-argument concept, as follows:

$$p(f, x) = \sum_j p(f, x|z_j) \Pr(z_j), \quad (j = 1, \cdots, m), \quad (6)$$

where $p(f, x|z_j)$ is the conditional joint distribution of F and X given $Z = z_j$, and $\Pr(z_j)$ (as before) is the unconditional probability that $Z = z_j$. Since the distribution $p(f, x|z_j)$ contains all of the information relevant to forecast quality under the condition $Z = z_j$ (assuming statistical stationarity), it follows that the decomposition in (6) provides a coherent approach to the problem of assessing forecast quality under *all* possible conditions, as defined by the values of the covariate Z.

Note that the introduction of the covariate Z leads to the stratification of the overall verification data sample into m subsamples, where the jth subsample consists of all pairs of forecasts and observations given $Z = z_j$. The distribution $p(f, x|z_j)$ would be estimated on the basis of the joint relative frequencies of the various combinations of forecasts and observations in this subsample. Some issues related to this estimation problem are discussed briefly in section 5.

To obtain information concerning the various basic aspects of forecast quality given the condition $Z = z_j$, it is necessary to decompose the joint distribution $p(f, x|z_j)$ into univariate distributions. As in the case of the overall joint distribution [see (4) and (5)], two such decompositions can be identified:

$$p(f, x|z_j) = p(x|f, z_j)p(f|z_j) \quad (7)$$

and

$$p(f, x|z_j) = p(f|x, z_j)p(x|z_j). \quad (8)$$

In (7), $p(x|f, z_j)$ represents the conditional distributions of the observations given the forecasts and the condition $Z = z_j$, and $p(f|z_j)$ represents the marginal distribution of the forecasts under this condition. Likewise, in (8), $p(f|x, z_j)$ represents the conditional distributions of the forecasts given the observations and the condition $Z = z_j$, and $p(x|z_j)$ represents the marginal distribution of the observations under this condition. To obtain diagnostic information of potential interest under all conditions, it is necessary to perform this decomposition for all m values of the covariate Z.

A coherent approach to the problem of assessing forecast quality—and its various aspects—under the m.e.c.e. set of meteorological conditions described by the covariate Z requires a body of verification methods analogous to that identified in the basic (i.e., unconditional) situation. Given $Z = z_j$, these three classes of methods are (a) the basic distributions $p(f, x|z_j)$, $p(x|f, z_j)$, $p(f|x, z_j)$, $p(f|z_j)$, and $p(x|z_j)$; (b) summary measures of these distributions; and (c) performance measures describing various aspects of the relationship between F and X under this condition. These classes of methods must be evaluated for all m values of Z to obtain a complete assessment.

For some summary measures of the underlying distributions and some measures of aspects of quality, a simple relationship exists between the condition-dependent values of these measures and their overall val-

ues. For these measures, the overall values are simply the weighted averages of the condition-dependent values, where the weights are the probabilities of the various conditions. Let M_j denote the value of such a measure given that condition $Z = z_j$ prevails ($j = 1, \cdots, m$). If M denotes the value of this measure for the verification data sample as a whole, it follows that

$$M = \sum_j M_j \Pr(z_j), \quad (j = 1, \cdots, m). \qquad (9)$$

Thus, if the overall values of such measures are required in addition to their condition-dependent values, then it is necessary to determine the probabilities of the respective conditions—that is, the $\Pr(z_j)$ ($j = 1, \cdots, m$). Measures for which (9) holds include (for example) the means of the underlying conditional and marginal distributions and the mean square error. On the other hand, many other measures, such as the variances of the underlying conditional and marginal distributions and the correlation coefficient, do not satisfy the relationship embodied in (9) (unless the condition-dependent, or subsample, means are all equal).

4. Applications: Conceptual examples

In this section we consider conceptual examples of the application of the extended framework described in section 3b. Each example involves the decomposition of the original joint distribution of forecasts and observations into conditional joint distributions on the basis of a m.e.c.e. set of values of a meteorological covariate. In terms of verification data, the decomposition process stratifies the overall data sample into subsamples, which can then be used to estimate these conditional joint distributions.

a. Example 1—Stratification by weather regime

Assessing (and/or comparing) the quality of forecasts under different weather regimes is a familiar problem. Such regimes can be defined in a variety of ways, ranging from large-scale weather patterns to regional weather types or local weather conditions (e.g., easterly or westerly surface wind, surface pressure rising or falling). The extended framework appears to provide a natural and coherent means of assessing forecast quality—and its various aspects—in such situations.

In applying the extended framework in this context, the first step is to define a relevant set of m.e.c.e. regimes $\{R_j;\ j = 1, \cdots, m\}$. Then, using the notation introduced in section 3, the quality of the forecasts given regime R_j is characterized fully (assuming statistical stationarity) by the conditional joint distribution $p(f, x|R_j)$. Moreover, the conditional distributions $p(x|f, R_j)$, $p(f|x, R_j)$, $p(f|R_j)$, and $p(x|R_j)$, obtained from decompositions of $p(f, x|R_j)$, provide insight into basic aspects of forecast quality (e.g., reliability, resolution, sharpness, discrimination) under regime R_j.

As indicated in section 3b, these basic distributions—that is, $p(f, x|R_j)$, $p(x|f, R_j)$, $p(f|x, R_j)$, $p(f|R_j)$, and $p(x|R_j)$—represent the first (i.e., primary) class of distributions-oriented verification methods. The other two classes of methods consist of summary measures (e.g., means, variances) of these basic distributions and performance measures describing specific aspects of the relationship between the forecasts and observations. A full assessment of forecast quality for regime R_j would involve the examination of results based on the application of all three classes of verification methods. Comprehensive assessment of forecast quality for all m regimes necessarily requires the examination of the results of applying these methods to all conditional joint distributions $p(f, x|R_j)$ ($j = 1, \cdots, m$). To calculate the overall values of summary measures or measures of aspects of quality that satisfy the relationship embodied in (9), the regime probabilities $\Pr(R_j)$ also must be determined.

This conceptual example is also a generic example in the sense that the form of the extended framework outlined here presumably applies (with relatively modest changes) to stratifications defined in terms of a wide variety of other meteorological conditions. Such conditions might include the presence or absence of some particular feature(s) in an observed or analyzed two-dimensional field or a m.e.c.e. set of initial conditions (i.e., the set of possible initial conditions that may prevail when the forecasts are made). Application of the extended framework to problems involving stratifications defined in terms of initial conditions might provide an insightful approach to problems of assessing/comparing the quality of forecasts of *changes* in the weather.

b. Example 2—Stratification by forecast difficulty

Another class of verification problems in which the extended framework may be useful are situations characterized by various levels of forecast difficulty. These problems are perhaps best exemplified by situations involving forecasts of rare and/or severe events (e.g., tornadoes, severe thunderstorms). In these situations, many forecasts are relatively "easy" in the sense that it is quite evident on these occasions that the rare or severe events are very unlikely to occur. Since the events of interest generally do not occur on most such occasions, the verification data sample (in situations involving "yes–no" forecasts and observations) often is dominated by cases involving the "no–no" combination (i.e., the combination involving cases in which the events are neither forecast nor observed to occur). In such situations, traditional verification methods may be relatively insensitive to changes or differences in quality for the remaining "difficult"—and usually more important—forecasts.

To illustrate the application of the extended framework in this context, consider a simple situation in

which a covariate W is used to stratify (a priori) the forecasting situations into easy cases ($W = w_1$) and difficult cases ($W = w_2$). Such a stratification might be accomplished by comparing the actual value of W with a critical value w_c, and by setting $W = w_1$ if $w < w_c$ and $W = w_2$ if $w \geq w_c$. Then, given the values of the covariate W, the overall joint distribution $p(f, x)$ can be decomposed as follows:

$$p(f, x) = p(f, x|w_1) \Pr(w_1) + p(f, x|w_2) \Pr(w_2), \tag{10}$$

where $p(f, x|w_1)$ represents the joint distribution of F and X for the easy cases ($W = w_1$), $p(f, x|w_2)$ represents the joint distribution of F and X for the difficult cases ($W = w_2$), $\Pr(w_1)$ represents the probability that $W = w_1$, and $\Pr(w_2)$ [$=1 - \Pr(w_1)$] represents the probability that $W = w_2$.

Suppose further that the forecasts and observations are binary variables; that is, the rare or severe events are either forecast to occur ($F = 1$) or forecast not to occur ($F = 0$), and they are subsequently either observed to occur ($X = 1$) or observed not to occur ($X = 0$). Under these conditions, the components of the conditional joint distributions $p(f, x|w_j)$ ($j = 1, 2$) correspond to the elements of 2×2 verification matrices. Let $\mathbf{P}_{w_1}$ and $\mathbf{P}_{w_2}$ denote the matrices containing the components of the joint distributions $p(f, x|w_1)$ and $p(f, x|w_2)$, respectively. Then the quality of the easy and difficult forecasts can be assessed separately by examining the elements of $\mathbf{P}_{w_1}$ and $\mathbf{P}_{w_2}$, respectively, together with the components of the respective univariate conditional distributions derived from decompositions of the joint conditional distributions [see (7) and (8)].

This approach may be particularly useful in the context of comparative verification. Suppose that the quality of two rare- or severe-event forecasting systems, A and B say, must be compared. In such a situation it seems reasonable to require that 1) A and B both adopt the same critical value w_c to determine whether $W = w_1$ or $W = w_2$ and 2) they both follow the same forecast strategy for the easy cases in which $W = w_1$; namely, they always forecast that the event(s) of interest will not occur (i.e., $F = 0$). Under these assumptions $\mathbf{P}_{w_1}(A) = \mathbf{P}_{w_1}(B)$, so that the easy cases are "quality neutral" with respect to the comparison of systems A and B. Thus, comparative verification of A and B can be "reduced" to the comparison of $\mathbf{P}_{w_2}(A)$ and $\mathbf{P}_{w_2}(B)$, and the univariate conditional and marginal distributions that can be derived from decompositions of these conditional joint distributions. Since the no–no [i.e., $\Pr(F = 0, X = 0)$] element in the matrix $\mathbf{P}_{w_2}$ presumably no longer plays the extremely dominant role that it did in the overall matrix $\mathbf{P}$ (the 2×2 matrix for the overall verification data sample), any differences between $\mathbf{P}_{w_2}(A)$ and $\mathbf{P}_{w_2}(B)$ should be more readily apparent and/or easier to detect.

The notion of forecast difficulty (or degree of predictability) also arises in the context of ensemble forecasting (e.g., Tracton and Kalnay 1993). Here, the decomposition associated with the extended framework might be defined in terms of a covariate representing a measure of ensemble dispersion. Application of the augmented framework in this context could provide a means of obtaining a coherent assessment of various aspects of forecasting performance as a function of such a dispersion measure.

5. Discussion

The extended framework for forecast verification embodies a coherent approach to the problem of stratifying verification data samples into subsamples on the basis of a specified set of meteorological conditions. Within this framework, the joint distribution of forecasts and observations for a particular subsample characterizes forecast quality completely (assuming statistical stationarity) under the corresponding meteorological condition. Several questions arise regarding the extended version of the general framework itself, as well as its application to verification problems, and some of these questions are addressed briefly in this section.

A basic question relates to the need for—and benefits of—such a framework. In this regard, modelers, forecasters, and others have conducted assessments of forecasting performance under specific meteorological conditions for many years. In most studies of this type, however, only a relatively small subset of the overall set of forecasting occasions has been investigated. Thus, although these case studies may have provided some useful insights into the behavior of models—and the characteristics of forecasting performance—in particular situations, their significance in terms of the contribution of subset forecasting performance to overall forecasting performance has seldom been very clear. Moreover, such studies generally have not taken full advantage of the diagnostic statistical methods associated with the distributions-oriented approach to verification problems.

In a practical sense, the extended framework provides a formal means of decomposing sample forecast quality into subsample forecast quality. Moreover, within this framework, the contribution of the latter to the former can be seen to depend on two factors; namely, the subsample quality itself and the probability (or relative frequency) of occurrence of the meteorological conditions that define the subsample. Thus, subsamples for which the latter factor is small generally will make relatively modest contributions (in a positive or negative sense) to overall forecast quality.

The extended framework also provides some insight into the nature of both meteorological case studies and traditional verification exercises, including their relationship. In particular, these two types of studies can

be seen to represent the extremes of a broad spectrum of possible forecast-quality assessments, with case studies usually focusing on some characteristics of a conditional joint distribution $p(f, x|z_j)$ (associated with the meteorological condition $Z = z_j$) and traditional verification exercises focusing on characteristics of the overall joint distribution $p(f, x)$. Clearly, a potentially rich "middle ground" for forecast-quality studies exists between these two extremes.

Diagnostic verification, as exemplified by the studies of Murphy et al. (1989) and Murphy and Winkler (1992), defines a particular class of statistical forecast-quality assessments in this middle ground. Within this class of assessments, the conditions of interest are represented by—and limited to—the values of the forecast F and the observation X. On the other hand, the extended framework provides a formal means of decomposing overall forecast quality into contributions associated with an essentially *unlimited* variety of meteorological conditions (through the introduction of the covariate Z). Moreover, when the problem of improving forecasting performance is considered from the perspective of this augmented framework, it may be possible to identify new ways of enhancing the diagnosticity and usefulness of future studies of forecast quality, regardless of whether these studies are primarily of a meteorological or statistical nature.

In this paper descriptions of the extended framework have been concerned principally with situations involving one set of forecasts (i.e., absolute verification). Comparative verification, which necessarily involves the relative quality of two (or more) sets of forecasts, is obviously also of interest in this context. Within the original framework, comparison of F's and G's forecasts (for example) would be based on their respective joint distributions, $p(f, x)$ and $q(g, x)$, and on the conditional and marginal distributions derived from decompositions of these underlying joint distributions. Application of the extended framework in the context of comparative verification appears to be relatively straightforward, since it presumably would involve a common stratification scheme applied to both sets of distributions. In addition to the use of diagnostic verification methods, the application of the sufficiency relation (Ehrendorfer and Murphy 1988; Murphy and Ehrendorfer 1994) to the respective subsamples of forecasts and observations defined by this stratification scheme might be explored as well. Although F's and G's forecasts may be insufficient for each other in an overall sense, such an investigation might reveal that F's forecasts are sufficient for (i.e., unambiguously superior to) G's forecasts—or vice versa—under some conditions defined by the meteorological covariate.

The utility of the extended framework as a means of obtaining condition-dependent assessments (or comparisons) of forecast quality depends on the availability of verification data samples—and subsamples—of adequate size. Moreover, it should be kept in mind that the second level of decomposition required to obtain diagnostic insight into basic aspects of forecast quality places additional requirements on sample or subsample size. Thus, only relatively large verification data samples can support a full condition-dependent, diagnostic assessment of forecast quality. As a result, the desirability of adopting relatively "narrow" definitions of meteorological conditions in order to focus on similar situations of particular interest must be weighed against the likelihood that estimates of forecast quality—and its various aspects—may be relatively unreliable under such conditions.

As described in this paper, the covariates characterizing the meteorological conditions of interest generally have been assumed to be univariate in nature. The weather regimes considered in section 4a are an exception since they usually are defined in terms of multivariate (or multidimensional) weather patterns or weather types. In any case, the extended framework described in section 3b places no restriction on the dimensionality of the covariates. Two or more different variables (e.g., temperature and precipitation) could be used to define the relevant conditions, or they could be defined in terms of the presence or absence of particular features in two-dimensional fields involving a single variable (e.g., a surface pressure field, a geopotential height field).

Moreover, no formal restriction exists on the types of forecasts (and observations) whose quality can be assessed and decomposed using either the basic or extended framework. For example, these frameworks can be applied to probabilistic, as well as nonprobabilistic, forecasts and to forecasts defined in terms of two-dimensional fields, as well as forecasts for specific points (e.g., stations, grid points). It should be noted that in some specific cases (e.g., probabilistic forecasts for polychotomous events) the number of distinct forecasts can impose relatively severe sample-size requirements for diagnostic verification (see Murphy 1991), requirements that may seldom be satisfied in practice.

6. Conclusions

As described in MW87, the general framework for forecast verification makes no explicit allowance for the inclusion of meteorological considerations. Such considerations are exemplified by the problems of assessing and/or comparing the quality of forecasts under specific meteorological conditions (e.g., weather regimes). An extended version of the general framework has been outlined here that addresses the problem of stratification in a coherent manner. The augmented framework has been formulated by appealing to the concept of extending the argument, a basic concept in elementary probability theory.

Application of the extending-the-argument concept in this paper has consisted of introducing a covariate whose values define a complete set of relevant mete-

orological conditions. These covariate values are then used to decompose the overall joint distribution of forecasts and observations into conditional joint distributions, each of which describes forecast quality completely under specific conditions. From a practical point of view, this process stratifies the overall verification data sample into subsamples, each of which provides a basis for estimating the conditional joint distribution associated with the corresponding meteorological conditions. The structure of the extended framework reveals that the contribution of subsample quality to sample quality depends on two factors; namely, the subsample quality itself and the relative frequency of the meteorological conditions associated with that subsample.

Decomposition of these conditional joint distributions into univariate distributions, in a manner analogous to that described for the overall joint distribution in MW87, facilitates insights into basic aspects of forecast quality under the various meteorological conditions. A body of diagnostic verification methods, essentially identical to that assembled in connection with the basic framework, is available to assess the various aspects of quality under these conditions. The methodological approach to forecast verification embodied in the extended framework thus appears to be quite powerful and flexible.

Conceptual examples were considered to illustrate the potential utility of the extended framework. These examples suggest that this framework may be useful in providing a structured approach to the problems of assessing and/or comparing forecast quality under a wide variety of meteorological conditions, including different weather regimes, various initial weather conditions, and different levels of forecast difficulty. Its use as a means of separating easy and difficult situations in the context of rare- or severe-event forecasting, thereby allowing attention to be focused on the absolute and/or relative quality of forecasts of significant events, appears to offer particular promise.

Some issues related to the generality and applicability of the extended framework were discussed briefly. These issues included insights provided by the framework itself into the nature of the relationship between meteorological case studies and traditional verification exercises, requirements concerning sample (and subsample) size imposed by application of the proposed framework, and use of the framework in comparative verification as a means of investigating the conditional sufficiency of alternative forecasting systems. It also was noted that the extended framework can be applied in situations involving multivariate covariates, as well as in situations involving forecasts expressed in probabilistic formats.

Firm conclusions regarding the utility of the extended version of the general framework for forecast verification introduced in this paper must await the application of this framework in a variety of real-world situations and the assessment of its impact on verification practices. At a minimum, however, the augmented framework appears to provide a reasonably general and potentially useful structural setting within which alternative verification or forecast-quality studies can be designed and/or their results evaluated. Moreover, application of the extended framework, including the decompositions identified with the phrase "diagnostic verification" (see section 3b), should lead to more insightful and useful meteorological case studies, whatever strategies are followed in the design and conduct of these studies.

Acknowledgments. In the early stages of this work, the author benefitted from discussions with Edward Epstein and Klaus Fraedrich. The helpful comments of an anonymous reviewer are greatly appreciated. The support of the University of Hamburg (Meteorological Institute), Max Planck Institute for Meteorology, and Swedish Meteorological and Hydrological Institute during the course of this research is gratefully acknowledged.

REFERENCES

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

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

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

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

——, and M. Ehrendorfer, 1994: Evaluation of forecasts. *Predictability and Nonlinear Modelling in Natural Sciences and Economics,* J. Grasman and G. van Straten, Eds., Kluwer Academic Publishers, 11–28.

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

O'Hagan, A., 1988: *Probability: Methods and Measurement.* Chapman and Hall, 291 pp.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.