Forecast Verification: Its Complexity and Dimensionality

ALLAN H. MURPHY*

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

(Manuscript received 14 September 1990, in final form 24 December 1990)

ABSTRACT

Two fundamental characteristics of forecast verification problems—*complexity* and *dimensionality*—are described. To develop quantitative definitions of these characteristics, a general framework for the problem of absolute verification (AV) is extended to the problem of comparative verification (CV). Absolute verification focuses on the performance of individual forecasting systems (or forecasters), and it is based on the bivariate distribution of forecasts and observations and its two possible factorizations into conditional and marginal distributions.

Comparative verification compares the performance of two or more forecasting systems, which may produce forecasts under 1) identical conditions or 2) different conditions. The first type of CV is matched comparative verification, and it is based on a 3-variable distribution with 6 possible factorizations. The second and more complicated type of CV is unmatched comparative verification, and it is based on a 4-variable distribution with 24 possible factorizations.

Complexity can be defined in terms of the number of factorizations, the number of basic factors (conditional and marginal distributions) in each factorization, or the total number of basic factors associated with the respective frameworks. These definitions provide quantitative insight into basic differences in complexity among AV and CV problems. Verification problems involving probabilistic and nonprobabilistic forecasts are of equal complexity.

Dimensionality is defined as the number of probabilities that must be specified to reconstruct the basic distribution of forecasts and observations. It is one less than the total number of distinct combinations of forecasts and observations. Thus, CV problems are of higher dimensionality than AV problems, and problems involving probabilistic forecasts or multivalued nonprobabilistic forecasts exhibit particularly high dimensionality.

Issues related to the implications of these concepts for verification procedures and practices are discussed, including the reduction of complexity and/or dimensionality. Comparative verification problems can be reduced in complexity by making forecasts under identical conditions or by assuming conditional or unconditional independence when warranted. Dimensionality can be reduced by parametric statistical modeling of the distributions of forecasts and/or observations.

Failure to take account of the complexity and dimensionality of verification problems may lead to an incomplete and inefficient body of verification methodology and, thereby, to erroneous conclusions regarding the absolute and relative quality and/or value of forecasting systems.

1. Introduction

In recent years, fundamental concepts and issues related to the problem of forecast verification have received some attention in the meteorological literature. For example, Murphy and Winkler (1987) (hereafter MW87) described a general framework for (absolute) forecast verification. This framework is based on the joint distribution of forecasts and observations, and it provides the basis for a coherent approach to verification procedures and practices. In particular, it led to the development of *diagnostic verification*, an approach that provides detailed insight into the basic characteristics of forecasting performance (Murphy et al. 1989). The extension of the framework for absolute verification to the more complicated problem of comparative verification has been sketched by Murphy (1989, pp. 94–95).

With regard to basic characteristics of verification problems, it has long been recognized that some verification problems are more complicated than others. For example, the problem of comparing the performance of two or more forecasting systems (or forecasters) is inherently more "complex" than the problem of evaluating the performance of an individual system (forecaster). Moreover, it is intuitively understood that verification of multicategory forecasts is necessarily more difficult than verification of 2-category forecasts. The "dimensionality" of the former, in terms of the number of different combinations of forecasts and observations, is higher than that of the latter. Cur-

* *Permanent affiliation:* Departments of Atmospheric Sciences and Statistics, Oregon State University, Corvallis, Oregon 97331.

Corresponding author address: Professor Allan H. Murphy, c/o Climate Analysis Center, NOAA/NWS/NMC, W/NMC5, WWB, Room 606, Washington, D.C.

rently, however, complexity and dimensionality are ill-defined concepts in the context of forecast verification, and the implications of these concepts for verification procedures and practices remain unclear.

The primary purposes of the present paper are 1) to provide quantitative definitions of the concepts of complexity and dimensionality in this context and 2) to discuss their implications for verification procedures and practices. To accomplish these objectives the framework for absolute verification is extended to the problem of comparative verification, and the framework for the latter is described in some detail. Explicit recognition of the complexity and dimensionality of verification problems is an important first step in the process of developing a coherent, adequate (or complete), and insightful body of verification methodology.

Section 2 describes the complexity of several verification problems. To provide a rationale for the quantitative definitions of complexity proposed here, frameworks for both absolute and comparative verification are outlined in this section. The dimensionality of verification problems is defined in section 3, and examples of verification problems are presented to illustrate the concept of dimensionality. Issues related to the implications of these concepts for verification procedures and practices are discussed in section 4. These issues include the dangers of ignoring complexity and dimensionality, possible ways of reducing the complexity and dimensionality of verification problems, and the implications of the concepts for verification datasets. Section 5 consists of a summary and some concluding remarks.

2. Complexity of verification problems

a. Absolute verification

In the case of *absolute verification* (AV) we are concerned with the performance of an individual forecasting system (or forecaster). A general framework for the AV problem is described in MW87. This framework is based on the joint distribution of forecasts and observations $p(f, x)$, where f denotes the forecasts and x denotes the observations. The distribution $p(f, x)$ contains all of the nontime-dependent information relevant to forecast verification (i.e., all of the information—except the time order of the forecast-observation pairs—that is required to determine the statistical characteristics of the forecasts, the observations, and their relationship).

As described in MW87, the bivariate distribution $p(f, x)$ can be factored into conditional and marginal distributions in two ways:

$$p(f, x) = p(x \mid f)p(f) \quad (1)$$

and

$$p(f, x) = p(f \mid x)p(x), \quad (2)$$

where $p(x \mid f)$ represents the conditional distributions of the observations given the forecasts, $p(f \mid x)$ represents the conditional distributions of the forecasts given the observations, $p(f)$ represents the marginal distribution of the forecasts, and $p(x)$ represents the marginal distribution of the observations. These conditional and marginal distributions provide access to the information contained in the joint distribution. The expressions in (1) and (2) generally are referred to as the *calibration-refinement* (CR) and *likelihood-base rate* (LBR) factorizations, respectively, of the distribution $p(f, x)$ (see MW87).

The practice of AV can be said to be adequate if it is based on a body of methodology that permits reconstruction of the bivariate distribution $p(f, x)$. For example, in the case of probabilistic forecasts for a 2-category variable (e.g., precipitation/no precipitation), AV based on an examination of $p(x \mid f)$—the calibration (or reliability) function—and $p(f)$—the refinement (or sharpness) function—is adequate (see MW87), since $p(f, x)$ can be reconstructed from these conditional and marginal distributions. On the other hand, AV based solely on a measure of overall forecast accuracy such as the Brier score (BS) (Brier 1950) is inadequate, since knowledge of the BS generally is *not* adequate to reconstruct the joint distribution.

Since verification practices based on either factorization can be adequate, these factorizations could be viewed as the bases of alternative approaches to AV. However, the two factorizations involve factors (i.e., distributions) that provide information regarding different characteristics of the forecasts, the observations, and their relationship (see MW87). Thus, it is more appropriate to view the CR and LBR factorizations as the bases of complementary approaches to AV. Of course, since the two factorizations are derived from the same bivariate distribution, the respective sets of factors are related, and complete knowledge of one set allows reconstruction of the other set (via the joint distribution). Nevertheless, in order to obtain a *complete* (as opposed to an adequate) assessment of forecasting performance, it is necessary to examine the distributions associated with both factorizations (as well as the bivariate distribution itself).

b. Comparative verification

When two or more forecasting systems (or forecasters) are compared, it is important to recognize that the two sets of forecasts may have been made under identical conditions or under different conditions (the phrase "identical conditions" implies the same forecasting situations, weather variable, geographic location, lead time, etc.). Since these circumstances can have a profound influence on the complexity of *comparative verification* (CV), it is necessary to distinguish between the two situations. The former is referred to as *matched comparative verification* (MCV) and the

latter is referred to as *unmatched comparative verification* (UCV).

In describing the basic distributions for MCV and UCV, and the factorizations of these distributions into conditional and marginal distributions, it will be useful to distinguish between *composite factors* and *basic factors*. Composite factors are distributions (conditional/unconditional), such as $p(f, x)$ in AV, that can be decomposed into other distributions. On the other hand, basic factors are distributions (conditional/marginal), such as $p(x|f)$ and $p(f)$ in AV, that cannot be decomposed into other distributions. The distinction between these two types of factors will become clear in the following paragraphs.

Matched comparative verification. In this case we are concerned with two forecasting systems that formulate forecasts under identical conditions. The variables of interest here are two sets of forecasts, f and g (for convenience, f and g are sometimes referred to as the type 1 and type 2 forecasts), and the corresponding set of observations x. The basic framework for this problem is the 3-variable (or trivariate) distribution $p(f, g, x)$. This distribution contains all of the nontime-dependent information relevant to MCV (i.e., nontime-dependent information regarding the statistical characteristics of the forecasts, the observations, and the relationships among the two types of forecasts and the observations).

The distribution $p(f, g, x)$ can be factored into conditional and marginal distributions in 6 (=3!) distinct ways:

$$p(f, g, x) = p(x|f, g)p(g|f)p(f), \quad (3)$$

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

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

$$p(f, g, x) = p(g|f, x)p(f|x)p(x), \quad (6)$$

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

and

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

where $p(x|f, g)$ represents the conditional distributions of the observations given both types of forecasts, $p(g|f)$ represents the conditional distributions of the type 2 forecasts given the type 1 forecasts, $p(f)$ represents the marginal distribution of the type 1 forecasts, etc. In (3)–(8) the factorizations are expressed in terms of basic factors. The process of deriving these expressions involves two steps: 1) decomposition of $p(f, g, x)$ into composite factors and basic factors and 2) decomposition of the composite factors obtained in step 1) into basic factors. For completeness, the intermediate expressions obtained in step 1) are included in appendix A. This appendix also briefly describes the relationship between the expressions in (3)–(8) and these intermediate expressions.

The practice of MCV is adequate if it is based on methodology that permits reconstruction of the trivariate distribution $p(f, g, x)$. Thus, verification procedures based on any of the six factorizations of $p(f, g, x)$ can be adequate. In addition, it is important to recognize that MCV necessarily involves consideration of the relationship between the two types of forecasts. For example, if MCV is based on the factorization described by (3), then the conditional distributions of the observations given both types of forecasts, $p(x|f, g)$; the conditional distributions of the type 2 forecasts given the type 1 forecasts, $p(g|f)$; and the marginal distribution of the type 1 forecasts, $p(f)$, must be considered (and/or the verification methodology used must permit reconstruction of these distributions). In general, however, MCV based on the two bivariate distributions of the forecasts and observations, $p(f, x)$ and $p(g, x)$, and the factors (conditional and marginal distributions) associated with their respective factorizations would be inadequate.

Since the framework for MCV admits six factorizations, six different (but not unrelated) approaches to this problem can be taken. As noted earlier in the case of AV, these approaches involve different distributions (in some cases) and thereby focus on different attributes or characteristics of the two types of forecasts, the corresponding observations, and their relationships. Thus, a complete MCV would involve examination of all of the conditional and marginal distributions associated with the six factorizations. Of course, some approaches may possess more (or less) intuitive appeal, or may be more meaningful or useful than others for particular purposes. For example, the approaches associated with the factorizations (3) and (4), or (6) and (8), would appear to be of particular interest, since they represent bivariate analogues of the CR and LBR factorizations, respectively, in absolute verification [see (A1) and (A6)].

In view of the attention devoted recently to the sufficiency relation in the context of comparative evaluation (e.g., see Ehrendorfer and Murphy 1988; Krzysztofowicz and Long 1991; Murphy and Ye 1990), it may be appropriate to discuss briefly its relationship to MCV as described here. According to the sufficiency relation, the forecasts f are sufficient for the forecasts g if it can be shown that the latter can be derived from the former by a stochastic transformation. In effect, the sufficiency relation involves the distributions $p(f|x)$, $p(g|x)$, and $p(g|f)$. The importance of the sufficiency relation resides in the fact that if the forecasts f are sufficient for the forecasts g, then *all* users will prefer f to g. However, this approach to comparative evaluation ignores the marginal distribution of the observations, $p(x)$. Thus, from the point of view of employing verification methodology that permits reconstruction of the basic distribution for MCV [i.e., $p(f, g, x)$], the sufficiency relation is inadequate.

Unmatched comparative verification. In this case we

are concerned with two forecasting systems that formulate forecasts under different conditions. The variables of interest are the two types of forecasts, f and g, and the corresponding types of observations, x and y, respectively (x and y are referred to as the type 1 and type 2 observations). The basic framework for this problem is the 4-variable distribution $p(f, g, x, y)$. As in the other cases (AV and MCV), the distribution $p(f, g, x, y)$ contains all of information relevant to UCV (i.e., all of the nontime-dependent information required to describe the statistical characteristics of both types of forecasts, both types of observations, and their relationships).

The distribution $p(f, g, x, y)$ can be factored into conditional and marginal distributions in 24 (=4!) distinct ways:

$$p(f, g, x, y) = p(y \mid f, g, x)p(x \mid f, g)p(g \mid f)p(f), \quad (9)$$

$$p(f, g, x, y) = p(y \mid f, g, x)p(x \mid f, g)p(f \mid g)p(g), \quad (10)$$

$$p(f, g, x, y) = p(y \mid f, g, x)p(g \mid f, x)p(x \mid f)p(f), \quad (11)$$

$$p(f, g, x, y) = p(y \mid f, g, x)p(g \mid f, x)p(f \mid x)p(x), \quad (12)$$

$$p(f, g, x, y) = p(y \mid f, g, x)p(f \mid g, x)p(x \mid g)p(g), \quad (13)$$

$$p(f, g, x, y) = p(y \mid f, g, x)p(f \mid g, x)p(g \mid x)p(x), \quad (14)$$

$$p(f, g, x, y) = p(x \mid f, g, y)p(y \mid f, g)p(g \mid f)p(f), \quad (15)$$

$$p(f, g, x, y) = p(x \mid f, g, y)p(y \mid f, g)p(f \mid g)p(g), \quad (16)$$

$$p(f, g, x, y) = p(x \mid f, g, y)p(g \mid f, y)p(y \mid f)p(f), \quad (17)$$

$$p(f, g, x, y) = p(x \mid f, g, y)p(g \mid f, y)p(f \mid y)p(y), \quad (18)$$

$$p(f, g, x, y) = p(x \mid f, g, y)p(f \mid g, y)p(y \mid g)p(g), \quad (19)$$

$$p(f, g, x, y) = p(x \mid f, g, y)p(f \mid g, y)p(g \mid y)p(y), \quad (20)$$

$$p(f, g, x, y) = p(g \mid f, x, y)p(y \mid f, x)p(x \mid f)p(f), \quad (21)$$

$$p(f, g, x, y) = p(g \mid f, x, y)p(y \mid f, x)p(f \mid x)p(x), \quad (22)$$

$$p(f, g, x, y) = p(g \mid f, x, y)p(x \mid f, y)p(y \mid f)p(f), \quad (23)$$

$$p(f, g, x, y) = p(g \mid f, x, y)p(x \mid f, y)p(f \mid y)p(f), \quad (24)$$

$$p(f, g, x, y) = p(g \mid f, x, y)p(f \mid x, y)p(y \mid x)p(x), \quad (25)$$

$$p(f, g, x, y) = p(g \mid f, x, y)p(f \mid x, y)p(x \mid y)p(y), \quad (26)$$

$$p(f, g, x, y) = p(f \mid g, x, y)p(y \mid g, x)p(x \mid g)p(g), \quad (27)$$

$$p(f, g, x, y) = p(f \mid g, x, y)p(y \mid g, x)p(g \mid x)p(x), \quad (28)$$

$$p(f, g, x, y) = p(f \mid g, x, y)p(x \mid g, y)p(y \mid g)p(g), \quad (29)$$

$$p(f, g, x, y) = p(f \mid g, x, y)p(x \mid g, y)p(g \mid y)p(y), \quad (30)$$

$$p(f, g, x, y) = p(f \mid g, x, y)p(g \mid x, y)p(y \mid x)p(x), \quad (31)$$

and

$$p(f, g, x, y) = p(f \mid g, x, y)p(g \mid x, y)p(x \mid y)p(y), \quad (32)$$

where $p(y \mid f, g, x)$ represents the conditional distributions of the type 2 observations given both types of forecasts and the type 1 observations, $p(x \mid f, g)$ represents the conditional distributions of the type 2 observations given both types of forecasts, $p(g \mid f)$ represents the conditional distributions of the type 2 forecasts given the type 1 forecasts, $p(f)$ represents the marginal distribution of the type 1 forecasts, etc.

As in the case of MCV, the factorizations of $p(f, g, x, y)$ in (9)–(32) are expressed in terms of basic factors. The process of deriving these expressions involves three steps in which the basic distribution is decomposed into composite (or composite and basic) factors in step 1, into composite and basic factors in step 2, and then into basic factors in step 3. For completeness, the intermediate expressions obtained in steps 1 and 2 are reproduced in appendix B. This appendix also briefly describes the relationship between the expressions in (9)–(32) and these intermediate expressions.

The practice of UCV is adequate if it is based on methodology that permits reconstruction of the 4-variable distribution $p(f, g, x, y)$. Thus, verification practices based on any of the 24 factorizations of $p(f, g, x, y)$ can be adequate. In this regard, it should be noted that UCV necessarily involves relationships between the two types of forecasts, between the two types of observations, and among the types of forecasts and observations. Once again, it is generally *not* adequate to consider only the bivariate distributions $p(f, x)$ and $p(g, y)$ when undertaking comparative verification in this context.

Thus, 24 different (but related) approaches to UCV are available. As in the case of MCV, these various approaches focus on different characteristics of the basic variables (i.e., f, g, x, and y) and their relationships, and some approaches may prove to be more useful than others. The relative merits of the different approaches warrant further study, but such investigations are beyond the scope of this paper.

c. Complexity

Comparison of the frameworks for AV, MCV, and UCV provides insight into the complexity (C) of these verification problems. Three characteristics of these frameworks can be readily identified: 1) the number of factorizations of the basic distribution (C_F); 2) the number of basic factors in each factorization (C_{BF}); and 3) the total number of basic factors (C_{TBF}). These characteristics of the frameworks for AV, MCV, and UCV are summarized in Table 1a. Note that C_{BF} also is equal to the number of variables involved in the basic distribution.

The complexity of the AV problem can be described by noting that its basic distribution, $p(f, x)$, admits two factorizations ($C_F = 2$), that each factorization involves two basic factors ($C_{BF} = 2$), and that the two factorizations involve a total of four basic factors ($C_{TBF} = 4$)—namely, two sets of conditional distributions, $p(x \mid f)$ and $p(f \mid x)$, and two marginal distributions, $p(f)$ and $p(x)$. Numerical values for these characteristics of the AV problem appear in Table 1a, and they provide a quantitative description of the complexity of this problem. The choice of a particular characteristic (or subset of characteristics) to describe the AV problem depends on the perspective that is taken. On the one hand, since each factorization provides the basis for a coherent and adequate approach to the verification problem, the complexity of this problem could be characterized by C_{BF} (=2). On the other hand, if the perspective is taken that it is necessary to explicitly examine all of the basic factors to perform a complete verification, then the problem could be characterized by C_{TBF} (=4).

In the case of CV, the numerical values in Table 1a indicate that MCV involves 6 factorizations, 3 basic factors in each factorization, and a total of 12 basic factors. On the other hand, UCV involves 24 factorizations, 4 basic factors in each factorization, and a total of 32 basic factors. Comparison of these numbers with the corresponding numbers for AV provides insight into the relative complexity of AV and CV as well as that of MCV and UCV.

For example, in terms of the number of basic factors (and variables) associated with each factorization, the increase in complexity from AV to MCV—and from MCV to UCV—appears relatively modest. However, in terms of the number of factorizations or the total number of basic factors, this increase in complexity appears quite substantial. In any event, these indices provide the basis for a quantitative assessment of the relative complexity of these (and other) verification problems.

It also is useful to consider the structure of the basic factors associated with the respective frameworks. This objective can be achieved by recognizing that each basic factor is a univariate distribution conditional on k variables ($k = 0, 1, 2, 3$). Information concerning this characteristic of the frameworks for AV, MCV, and UCV is summarized in Table 1b. The basic factors for AV are two marginal distributions and two distributions conditional on one variable. On the other hand, the basic factors for UCV are 4 marginal distributions, 12 distributions conditional on 1 variable, 12 distributions conditional on 2 variables, and 4 distributions conditional on 3 variables.

As yet no mention has been made of the nature of the basic variables (continuous, discrete) or the type of forecasts (probabilistic, nonprobabilistic). Complexity, as defined here, does not depend on these factors. That is, the complexity of verification problems is independent of the nature of the variable and the forecast type. As we shall discover in section 3, this statement does *not* hold for the dimensionality of verification problems.

TABLE 1. (a) Characteristics of complexity of frameworks for absolute verification (AV), matched comparative verification (MCV), and unmatched comparative verification (UCV). (b) Nature of basic factors associated with frameworks for AV, MCV, and UCV, described in terms of number of factors with k conditioning variables. See text for additional details.

(a)	Number of factorizations (C_F)	Number of basic factors in each factorization (C_{BF})	Total number of basic factors (C_{TBF})
AV	2	2	4
MCV	6	3	12
UCV	24	4	32

	Number of basic factors with k conditioning variables				
(b)	$k = 0$	$k = 1$	$k = 2$	$k = 3$	Total
AV	2	2	0	0	4
MCV	3	6	3	0	12
UCV	4	12	12	4	32

3. Dimensionality of verification problems

For the purposes of this discussion of the dimensionality of verification problems, it will be assumed that the basic distributions of interest here [i.e., $p(f, x)$ in the case of AV, $p(f, g, x)$ in the case of MCV, and $p(f, g, x, y)$ in the case of UCV] are described by joint (2-variable, 3-variable, and 4-variable) empirical relative frequencies derived from a relevant sample of forecasts and observations. These relative

frequencies are estimates of the corresponding probabilities. The dimensionality of the various verification problems obviously depends on the dimensionality of these basic distributions.

Specifically, the dimensionality D of verification problems can be defined as the number of relative frequencies (or probabilities) that must be specified in order to reconstruct the basic 2-variable, 3-variable, and 4-variable distributions of forecasts and observations. Thus, as defined, D is equivalent to the number of degrees of freedom associated with these distributions. Using the problem of AV as an example, the dimensionality (D_{AV}) is the number of probabilities required to specify $p(f, x)$. Let I and K denote the number of distinct forecasts and observations, respectively. Then $D_{AV} = IK - 1$, since the joint probabilities must sum to one. If it is assumed that the climatological probabilities of the observations, $p(x)$, are known, then the dimensionality is reduced from D_{AV} to D^*_{AV}, where $D^*_{AV} = IK - 1 - (K - 1) = (I - 1)K$.

A few specific examples are considered to illustrate the dimensionality of AV (and other) problems. In the case of nonprobabilistic (yes/no) forecasts in a dichotomous situation, $I = K = 2$ and $D_{AV} = 3 (D^*_{AV} = 2)$. Thus, three probabilities must be specified to completely describe the joint distribution $p(f, x)$ in this context (when the climatological probabilities are known, only two probabilities must be specified). In the case of probabilistic forecasts, with $I = 11$ probability values [0(0.1)1] and $K = 2$ observed values (0, 1), $D_{AV} = 21$ and $D^*_{AV} = 20$. The introduction of probabilistic forecasts leads to a substantial increase in dimensionality in this situation ($K = 2$).

The values of D_{AV} (and D^*_{AV}), for these and other typical verification problems, are included in Table 2. Note that the dimensionality of AV problems involving nonprobabilistic forecasts increases rapidly as the number of distinct forecasts and observations increases (e.g., $D_{AV} = 99$ when $I = K = 10$). Moreover, AV problems involving multicategory ($K > 2$) probabilistic forecasts exhibit particularly high dimensionality ($D_{AV} = 197$ when $K = 3$; see footnote in Table 2).

For CV problems, the dimensionality is even higher because these problems involve additional variables. In the case of MCV, for example, $D_{MCV} = IJK - 1$ [and $D^*_{MCV} = K(IJ - 1)$], where I denotes the number of distinct type 1 forecasts, J denotes the number of distinct type 2 forecasts, and K denotes the number of distinct observations (it is assumed here that I and J are not necessarily equal). The values of D_{MCV} and D^*_{MCV} for the examples considered in conjunction with AV problems also are included in Table 2. Note that $D_{MCV} = 7$ when $I = J = K = 2$ (nonprobabilistic forecasts in a dichotomous situation) and $D_{MCV} = 241$ when $I = J = 11$ and $K = 2$ (probabilistic forecasts in a dichotomous situation). Moreover, the dimensionality attains very high values for nonprobabilistic forecasts with many values (e.g., $D_{MCV} = 999$ when $I = J = K = 10$) and for multicategory probabilistic forecasts ($D_{MCV} = 13\,067$ when $I = J = 66$ and $K = 3$).

In the case of UCV, $D_{UCV} = IJKL - 1 (D^*_{UCV} = IJKL - K - L + 1)$, where K and L denote the number of type 1 and type 2 observations, respectively (with K and L not necessarily equal). Since UCV involves an additional variable (i.e., the type 2 observations), the dimensionality is even higher for these problems. For example, $D_{UCV} = 15$ when $I = J = K = L = 2$ and $D_{UCV} = 483$ when $I = J = 11$ and $K = L = 2$ (see Table 2). As before, dimensionality increases further still for problems involving additional categories. Note that $D_{UCV} = 9999$ for $I = J = K = L = 10$, and $D_{UCV} = 39\,204$ for $I = J = 66$ and $K = L = 3$.

From the discussion in this section, it is clear that the dimensionality of verification problems depends on the treatment of the basic variables and the type of

TABLE 2. Dimensionality of some typical verification problems. Here D_{AV} (D_{MCV}, D_{UCV}) denotes total dimensionality, and D^*_{AV} (D^*_{MCV}, D^*_{UCV}) denotes dimensionality under the assumption that the climatological probabilities $p(x)$ and $p(y)$ are known (AV = absolute verification, MCV = matched comparative verification, and UCV = unmatched comparative verification).

Type of forecasts	Number of distinct forecasts	Number of distinct observations	D_{AV} (D^*_{AV})	D_{MCV} (D^*_{MCV})	D_{UCV} (D^*_{UCV})
Nonprobabilistic	2	2	3 (2)	7 (6)	15 (13)
Probabilistic	11	2	21 (20)	241 (240)	483 (481)
Nonprobabilistic	3	3	8 (6)	26 (24)	80 (76)
Probabilistic	66*	3	197 (195)	13067 (13065)	39204 (39200)
Nonprobabilistic	5	5	24 (20)	124 (120)	624 (616)
Nonprobabilistic	10	10	99 (90)	999 (990)	9999 (9981)

* In the case of 3-category probabilistic forecasts, where $f = (f_1, f_2, f_3)$ with $f_i = 0(0.1)1$ ($i = 1, 2, 3$) and $f_1 + f_2 + f_3 = 1$, 66 distinct forecasts can be identified.

forecasts. Dimensionality is high when the basic variable possesses many values or is divided into a large number of categories and/or when the forecasts are expressed in a modestly resolved probabilistic format. Thus, the practical consequences of the difference between probabilistic and nonprobabilistic forecasts in this context consists of an increase in the dimensionality of verification problems rather than an increase in their complexity. The relatively high dimensionality of some of these problems (see Table 2) raises questions related (inter alia) to the possibility of reducing the dimensionality of verification problems and the size of verification datasets, and these issues are discussed in section 4.

4. Discussion: Some implications and consequences

The discussions of the complexity and dimensionality concepts in sections 2 and 3, respectively, reveal that CV is considerably more complex than AV and that the dimensionality of many AV and CV problems is quite high. Since it should now be evident that verification problems are of greater complexity and higher dimensionality than generally recognized heretofore, several questions arise regarding the implications and/or consequences of these concepts for verification procedures and practices. For example, what are the dangers of ignoring considerations of complexity and dimensionality in the practice of forecast verification? How can the complexity and/or dimensionality of verification problems be reduced? What are the implications of these new perspectives concerning the nature of AV and CV problems for verification datasets? In this section we briefly discuss these issues.

a. Dangers of ignoring complexity and/or dimensionality

In the practice of forecast verification, the complexity and dimensionality of the problem at hand are seldom considered. For example, comparative verification of precipitation probability forecasts under unmatched conditions (e.g., at two different locations) is frequently performed using a climatological skill score, with the differences between the two sets of observations being represented solely by the respective climatological probabilities. This practice also assumes that skill (i.e., relative accuracy) is the only relevant aspect of forecast quality. In fact, most absolute verification is performed in terms of one or two overall performance measures (e.g., measures of accuracy and/or skill), and these measures generally do not permit reconstruction of the basic distribution of forecasts and observations.

What are the dangers inherent in these practices? First, it is evident that such practices necessarily overlook various characteristics of the forecasts, the observations, and/or their relationship(s). With regard to the latter (i.e., the relationship between forecasts and observations), important characteristics of forecast quality may be ignored. Moreover, when a single overall performance measure is used to evaluate the forecasts of interest, the likelihood of overlooking important characteristics undoubtedly increases as complexity and/or dimensionality increase. This discussion raises a fundamental question: under what conditions do overall performance measures (such as traditional measures of accuracy and skill) capture the essential features of forecast quality? For example, what important features of the quality of precipitation probability forecasts are overlooked when they are evaluated using a skill score? Moreover, to what extent does the answer to this question vary from one dataset to another dataset? At the moment, little if any information exists that bears directly or indirectly on the answers to such questions. Nevertheless, these basic questions appear to warrant very careful consideration in the future.

Although the "scientific" dangers inherent in ignoring complexity and/or dimensionality must await the results of studies designed to answer the questions posed in the previous paragraph, the "economic" dangers are already apparent. The failure to respect the full dimensionality of verification problems can lead to rather surprising results concerning the relative value of forecasting systems. In this regard, it should be noted that in comparative verification, measures of forecast accuracy (or skill) are frequently used as surrogates for measures of economic value. In particular, more accurate forecasts usually are assumed to be more valuable to users. However, Murphy and Ehrendorfer (1987) have shown that even in a simple situation in which only two probabilities are required to characterize forecast quality completely, the use of a one-dimensional measure of performance (e.g., a measure of accuracy such as the Brier score) can lead to reversals in the usual accuracy/value relationship. That is, forecasts with a larger (i.e., worse) Brier score actually can be of greater value to some users. (Since forecast quality is related to forecast value, such results also indirectly demonstrate the scientific dangers inherent in arbitrarily reducing the dimensionality of verification problems.) This result underscores the need to measure forecasting performance in its full dimensionality, or at least to reduce dimensionality in such a way that the essential components of forecast quality are retained and the likelihood of quality/value reversals is minimized.

b. Reducing complexity

Since the complexity of AV problems cannot be reduced, this discussion will focus on ways of reducing the complexity of MCV and UCV problems. Perhaps the most obvious way to reduce the complexity of UCV problems is to transform these problems into MCV problems by comparing the forecasting systems of interest under identical conditions. Such an approach

may require that greater attention be given to the design of forecasting experiments and/or that more effective use be made of existing datasets. In any case, the substantial decrease in complexity achieved by reducing an UCV problem to an MCV problem implies that very careful consideration should be given to this approach. Of course, it is not always possible to compare forecasting systems under identical conditions. Some comparisons (e.g., the comparison of forecasts of the same variable at two different locations) are fundamentally problems of unmatched comparative verification.

In some situations it may be possible to invoke the assumptions of independence or conditional independence to reduce complexity. In the case of UCV, for example, it might be reasonable in some circumstances to assume that the two types of observations (x and y) are independent. Such circumstances might include comparisons involving observations from different time periods at the same location or comparisons involving observations from two widely separated locations. When this assumption can be justified, $p(y|x) = p(y)$ and $p(x|y) = p(x)$. As a result, the number of basic factors with one conditioning variable (i.e., $k = 1$) decreases from 12 to 10, and the total number of basic factors associated with the UCV problem is reduced from 32 to 30 (see Table 1b).

Conditional independence implies that two variables are independent conditional on a third variable. For example, in the case of MCV, the variables g and x are conditionally independent given the variable f when it can be shown that $p(x|f, g) = p(x|f)$. Since $p(x|f, g) = p(x|f)$ implies that $p(g|f, x) = p(g|f)$, the assumption of conditional independence in this context reduces the number of basic factors with two conditioning variables (i.e., $k = 2$) from three to one, and the total number of basic factors associated with the MCV problem is reduced from 12 to 10 (see Table 1b). The concept of conditional independence has been employed in the context of comparative verification as a means of investigating the incremental information content in objective and subjective weather forecasts (e.g., Clemen and Murphy 1986; Murphy et al. 1988).

It may be of interest here to describe briefly the relationship between conditional independence and sufficiency (recall the discussion of the sufficiency relation in section 2b). Suppose that, in the context of MCV, g and x are conditionally independent given f. Then, it is quite easy to show that f is sufficient for g, and proof of this result is sketched in appendix C. However, the converse is not true; that is, sufficiency does *not* imply conditional independence. Thus, conditional independence is a stronger result than sufficiency.

c. *Reducing dimensionality*

The dimensionality of verification problems can be decreased by reducing the number of probabilities that must be specified to reconstruct the basic distribution. As noted in section 3 (see also Table 2), a small reduction in dimensionality can be achieved when it can be assumed that the distribution of observations $p(x)$ is known. This assumption is equivalent to assuming that the sample climatology is identical to the long-term historical climatology (an assumption that is more likely to be satisfied as the sample size increases, under conditions of stationarity).

A potentially efficient and effective way to reduce the dimensionality of verification problems is to model the conditional and/or unconditional distributions. In this approach, parametric statistical models are fit to the relevant distributions, and the evaluation of forecast quality is then based on the parameters of the model(s). Such an approach can lead to quite substantial reductions in dimensionality, and it has the added feature that the effect of sampling variability on the results of the verification process may be reduced.

Studies involving the use of statistical models to characterize the relationship between forecasts and observations are for the most part of relatively recent vintage. In this regard, parametric models have been employed in some decision-analytic investigations of the value of weather and climate forecasts. For example, Katz et al. (1982) used a bivariate normal distribution to characterize the relationship between daily minimum temperature forecasts (expressed in a nonprobabilistic format) and the corresponding observations. In effect, such a model reduces the dimensionality of the verification problem to (no more than) five dimensions, represented by two means, two variances, and a covariance.

Of more direct relevance to this discussion, Krzysztofowicz and Long (1991) used a parametric modeling approach to reduce the dimensionality of an MCV problem involving objective and subjective precipitation probability forecasts. Specifically, they used beta densities to fit the conditional distributions (or likelihoods) $p(f|x)$ and $p(g|x)$ in order to facilitate their study of the conditions under which one set of forecasts could be judged to be sufficient for another set of forecasts. In effect, this approach reduced the dimensionality of the problem to four dimensions, represented by the two parameters associated with the respective beta distributions.

It should be noted that it may not always be possible to find parametric models that fit the relevant distributions in a satisfactory manner. For example, Clemen and Winkler (1987) used a normal log-odds model to fit the likelihood functions [i.e., $p(f|x)$] of samples of precipitation probability forecasts and the corresponding observations in a calibration and combining study. They found that these models tended to yield distributions that were appreciably more skewed than the empirical distributions.

In the absence of reasonable parametric models, it still may be possible to reduce the dimensionality of

verification problems when the empirical probabilities that constitute the basic distribution exhibit relatively small or negligible differences for various combinations of forecasts and observations. In this context, we can distinguish between warranted and unwarranted reductions in dimensionality. Roughly speaking, reductions in dimensionality would be warranted when these probabilities are indistinguishable (in a statistical sense); conversely, they would be unwarranted when the probabilities are clearly distinguishable. Statistical tests could be used to determine whether or not such differences are distinguishable and, as a result, whether or not corresponding reductions in dimensionality are justified.

d. Implications for verification datasets

The fact that forecast verification problems are of greater complexity and higher dimensionality than generally recognized heretofore has important implications for verification datasets. In particular, since both AV and CV involve the evaluation of conditional distributions (and measures based on conditional distributions), adequate verification based on the frameworks described in this paper requires larger datasets than those required for traditional verification procedures involving the computation of overall performance measures. Moreover, since the degree of conditionality associated with the framework for MCV (UCV) is greater than that for AV (MCV), larger datasets generally will be required for MCV (UCV) than for AV (MCV). The practical significance of this sample-size issue will become clear only after some additional experience is gained in applying the respective frameworks.

If it is possible to model conditional distributions in a satisfactory manner, then the impact of the sample-size problem may be reduced. In any case, it should be possible to perform considerably more informative and insightful CV studies than those conducted heretofore without evaluating the distributions that involve the highest degree of conditionality (e.g., $k = 3$ in Table 1b). Moreover, in some verification studies, consideration could be given to combining datasets from different regions (or locations) or time periods to obtain large enough samples to permit application of (at least) portions of the CV frameworks described here.

In summary, the size of the dataset may limit the evaluation of forecasting performance in its *full* dimensionality in some verification studies. However, sample sizes generally should be adequate to permit the examination of some conditional distributions and/or conditional performance measures. Reliable information of this type will yield considerably more insight into the basic characteristics of forecast quality than that provided by overall performance measures.

5. Conclusion

This paper represents a contribution toward the development of a deeper understanding of the true nature of forecast verification problems. With this overall goal in mind, these problems have been considered from the perspective of general frameworks for absolute and comparative verification. The framework for absolute verification, described in MW87, is based on the bivariate distribution of forecasts and observations and on factorizations of this distribution into conditional and marginal distributions. This framework was extended here to frameworks for the problems of matched and unmatched comparative verification. The latter are based on 3-variable and 4-variable distributions, respectively, as well as on factorizations of these distributions into conditional and marginal distributions. Since the basic distributions—and each of the respective factorizations—contain all of the nontime-dependent information relevant to forecast verification, the practice of verification is *adequate* only if it is possible to reconstruct these distributions from the methodology actually employed. Moreover, since different distributions relate to different characteristics of performance, forecast verification can be said to be *complete* only if it involves consideration of the basic factors associated with all of the relevant factorizations.

Two fundamental characteristics of verification problems—complexity and dimensionality—were described here, and quantitative measures of these characteristics were defined. Complexity relates to the structure and components of the frameworks that undergird such problems. Several measures of complexity were identified, including indices defined in terms of the number of factorizations, the number of basic factors associated with each factorization, and the total number of basic factors associated with a particular framework. Although it has been understood heretofore (implicitly if not explicitly) that comparative verification is more complex than absolute verification and that unmatched comparative verification is more complex than matched comparative verification, these indices can serve as *quantitative* measures of the relative complexity of these problems. Moreover, according to the definitions of complexity introduced here, this characteristic of verification problems is not influenced by the format of the forecasts. Within a given framework, the complexity of verification problems is the same whether the forecasts are expressed in a probabilistic or nonprobabilistic format.

Dimensionality relates to the number of probabilities that must be specified, within a particular framework, in order to reconstruct the basic distribution relevant to that framework. Thus, the dimensionality of a verification problem is one less than the product of the number of distinct combinations of forecasts and observations (note that comparative verification problems

involve two types of forecasts and one or two types of observations). Examination of the dimensionality of several typical verification problems revealed that comparative verification problems—and all verification problems involving probabilistic forecasts or nonprobabilistic forecasts with many distinct forecast values or categories—possess relatively high dimensionality.

In the process of developing quantitative definitions of complexity and dimensionality, it has become evident that comparative verification problems are relatively complex and that many absolute and comparative verification problems are of relatively high dimensionality. Several questions arise regarding the practical implications of these concepts. These questions relate to the dangers of ignoring complexity and dimensionality in conducting verification studies, the possible ways of reducing complexity and/or dimensionality, and the consequences of these conceptual developments for verification datasets. The dangers involved in the current practice of (largely) ignoring complexity and dimensionality can be summarized by indicating that such an approach may lead to erroneous conclusions regarding the absolute and relative quality and value of alternative forecasting systems.

Several ways of reducing complexity or dimensionality were briefly discussed. For example, it may be feasible to design some forecasting studies in such a way that unmatched comparative verification problems are transformed into matched comparative verification problems, with the important benefit that the underlying framework for the latter is considerably less complex than that for the former. Moreover, it may be possible to invoke assumptions such as independence or conditional independence to simplify the structure of verification problems. This latter possibility clearly warrants further investigation.

The use of parametric statistical models of the conditional or unconditional distributions that constitute the basic factors in the underlying frameworks offers a promising means of reducing the dimensionality of verification problems. Comparative verification could then be based on the parameters of these distributions, yielding a substantial reduction in dimensionality. In addition to reducing dimensionality, the use of such models would serve to minimize the effects of sampling variability on the results of forecast verification studies. Moreover, it may be possible to reduce the (apparent) dimensionality of some verification problems by recognizing that distinctions between some values of the relevant variables are unwarranted (as revealed through comparisons of the respective empirical probabilities).

The implementation of adequate verification procedures and practices implies the need for relatively large datasets. In practice, however, the sample sizes of the relevant datasets will be limited, and evaluators will be required to make judicious compromises between traditional practices based on one or two overall measures of performance and adequate procedures as defined here. Nevertheless, since traditional procedures are clearly *inadequate,* the available datasets should be exploited to the greatest extent possible to obtain detailed insight into the basic characteristics of forecasting performance.

Before considering possible directions for future work in this area, it seems appropriate to underline once again the deficiencies inherent in current practices. In this regard, studies based solely on one or two overall measures necessarily fail to describe potentially important characteristics of forecasting performance. In fact, except in the simplest situations, approaches involving overall performance measures may be inadequate in this sense. How serious is the loss of information concerning forecasting performance that occurs when the complexity and/or dimensionality of verification problems is arbitrarily reduced by restricting the scope of the methodology actually employed? Since explicit recognition of the concepts of complexity and dimensionality—and their implications for verification procedures and practices—is quite recent, this question cannot be answered at the present time. However, it appears to be a *very* fundamental question and a question that needs to be addressed by those concerned with developing a coherent, adequate (or complete), and useful body of verification methods.

Future work in this context should include studies of alternative ways of reducing complexity and dimensionality, as well as efforts to extend these concepts to other verification problems. Since many such problems are relatively complex and of relatively high dimensionality, it is essential to find rational ways of simplifying these problems. This work will require a sound knowledge of verification problems (i.e., frameworks, methods, etc.), as well as a willingness to explore and test the appropriateness of various assumptions and models using real data. With regard to the concepts of complexity and dimensionality themselves, it would be desirable to extend these concepts to other verification problems, including problems involving forecasts expressed in the form of multidimensional fields.

Acknowledgments. This work was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8714108. Work on the first draft of this paper was completed in September 1990 at which time the author was a visiting scientist at the Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado. The author would like to express his appreciation to A. E. MacDonald, Director of NOAA's Forecast Systems Laboratory, for making the visit to CIRES possible. Currently, the author is a University Corporation for Atmospheric Research visiting scientist at the National Meteorological Center.

E. S. Epstein, L. S. Gandin, R. L. Winkler, and an anonymous reviewed provided valuable comments on earlier versions of the manuscript.

APPENDIX A

Matched Comparative Verification: Decompositions Involving Composite and Basic Factors

The expressions obtained from the first step in the process of decomposing the basic distribution $p(f, g, x)$ into conditional and marginal distributions are presented in this appendix. These expressions involve basic factors and composite factors, and they can be written as follows:

$$p(f, g, x) = p(x \mid f, g)p(f, g), \quad \text{(A1)}$$

$$p(f, g, x) = p(g \mid f, x)p(f, x), \quad \text{(A2)}$$

$$p(f, g, x) = p(f \mid g, x)p(g, x), \quad \text{(A3)}$$

$$p(f, g, x) = p(g, x \mid f)p(f), \quad \text{(A4)}$$

$$p(f, g, x) = p(f, x \mid g)p(g), \quad \text{(A5)}$$

and

$$p(f, g, x) = p(f, g \mid x)p(x). \quad \text{(A6)}$$

Note that each expression contains one composite factor and one basic factor.

The second step in the process, which leads to (3)–(8), involves decomposing the composite factors in (A1)–(A6) into basic factors. For example, $p(f, g)$ in (A1) can be decomposed into $p(g \mid f)p(f)$ or $p(f \mid g)p(g)$, yielding (3) and (4), respectively. Analogously, $p(g, x \mid f)$ in (A4) can be decomposed into $p(x \mid f, g)p(g \mid f)$ or $p(g \mid f, x)p(x \mid f)$, yielding (3) and (5), respectively. Thus, this step leads to a set of 12 ($=6 \times 2$) expressions, with each of the six factorizations [i.e., (3)–(8)] appearing twice in this set.

APPENDIX B

Unmatched Comparative Verification: Decompositions Involving Composite and Basic Factors

The expressions obtained from the first and second steps in the process of decomposing the basic distribution $p(f, g, x, y)$ into conditional and marginal distributions are presented in this appendix. These expressions involve basic factors and/or composite factors, and they can be written as follows:

$$\begin{aligned} p(f, g, x, y) &= p(y \mid f, g, x)p(f, g, x), \\ &= p(y \mid f, g, x)p(x \mid f, g)p(f, g), \\ &= p(y \mid f, g, x)p(g \mid f, x)p(f, x), \\ &= p(y \mid f, g, x)p(f \mid g, x)p(g, x), \quad \text{(B1)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(x \mid f, g, y)p(f, g, y), \\ &= p(x \mid f, g, y)p(y \mid f, g)p(f, g), \\ &= p(x \mid f, g, y)p(g \mid f, y)p(f, y), \\ &= p(x \mid f, g, y)p(f \mid g, y)p(g, y), \quad \text{(B2)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(g \mid f, x, y)p(f, x, y), \\ &= p(g \mid f, x, y)p(y \mid f, x)p(f, x), \\ &= p(g \mid f, x, y)p(x \mid f, y)p(f, y), \\ &= p(g \mid f, x, y)p(f \mid x, y)p(x, y), \quad \text{(B3)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(f \mid g, x, y)p(g, x, y), \\ &= p(f \mid g, x, y)p(y \mid g, x)p(g, x), \\ &= p(f \mid g, x, y)p(x \mid g, y)p(g, y), \\ &= p(f \mid g, x, y)p(g \mid x, y)p(x, y), \quad \text{(B4)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(x, y \mid f, g)p(f, g), \\ &= p(y \mid f, g, x)p(x \mid f, g)p(f, g), \\ &= p(x \mid f, g, y)p(y \mid f, g)p(f, g), \quad \text{(B5)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(g, y \mid f, x)p(f, x), \\ &= p(y \mid f, g, x)p(g \mid f, x)p(f, x), \\ &= p(g \mid f, x, y)p(y \mid f, x)p(f, x), \quad \text{(B6)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(g, x \mid f, y)p(f, y), \\ &= p(x \mid f, g, y)p(g \mid f, y)p(f, y), \\ &= p(g \mid f, x, y)p(x \mid f, y)p(f, y), \quad \text{(B7)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(f, y \mid g, x)p(g, x), \\ &= p(y \mid f, g, x)p(f \mid g, x)p(g, x), \\ &= p(f \mid g, x, y)p(y \mid g, x)p(g, x), \quad \text{(B8)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(f, x \mid g, y)p(g, y), \\ &= p(x \mid f, g, y)p(f \mid g, y)p(g, y), \\ &= p(f \mid g, x, y)p(x \mid g, y)p(g, y), \quad \text{(B9)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(f, g \mid x, y)p(x, y), \\ &= p(g \mid f, x, y)p(f \mid x, y)p(x, y), \\ &= p(f \mid g, x, y)p(g \mid x, y)p(x, y), \quad \text{(B10)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(g, x, y \mid f)p(f), \\ &= p(x, y \mid f, g)p(g \mid f)p(f), \\ &= p(g, y \mid f, x)p(x \mid f)p(f), \\ &= p(g, x \mid f, y)p(y \mid f)p(f), \quad \text{(B11)} \end{aligned}$$

$$\begin{aligned} p(f, g, x, y) &= p(f, x, y \mid g)p(g), \\ &= p(x, y \mid f, g)p(f \mid g)p(g), \\ &= p(f, y \mid g, x)p(x \mid g)p(g), \\ &= p(f, x \mid g, y)p(y \mid g)p(g), \quad \text{(B12)} \end{aligned}$$

$$p(f, g, x, y) = p(f, g, y|x)p(x),$$
$$= p(g, y|f, x)p(f|x)p(x),$$
$$= p(f, y|g, x)p(g|x)p(x),$$
$$= p(f, g|x, y)p(y|x)p(x), \quad \text{(B13)}$$

and

$$p(f, g, x, y) = p(f, g, x|y)p(y),$$
$$= p(g, x|f, y)p(f|y)p(y),$$
$$= p(f, x|g, y)p(g|y)p(y),$$
$$= p(f, g|x, y)p(x|y)p(y). \quad \text{(B14)}$$

Examination of (B1)–(B14) reveals that each expression on lines 2–4 of (B1)–(B4) and (B11)–(B14)—and each expression on lines 2–3 of (B5)–(B10)—contains one composite factor and two basic factors.

The third step in the process, which leads to (9)–(32), involves decomposing the composite factors on lines 2–4 of (B1)–(B4) and (B11)–(B14)—and lines 2–3 of (B5)–(B10)—into basic factors. For example, $p(f, g)$ on line 2 of (B1) can be decomposed into $p(g|f)p(f)$ or $p(f|g)p(g)$ yielding (9) and (10), respectively. Analogously, $p(x, y|f, g)$ on line 2 of (B5) can be decomposed into $p(y|f, g, x)p(x|f, g)$, which also yields (9) and (10), respectively, when $p(f, g)$ is decomposed into its two possible expressions. Thus, this step leads to a set of 72 ($=8 \times 3 \times 2 + 6 \times 2 \times 2$) expressions, with each of the 24 factorizations [i.e., (9)–(32)] appearing three times in this set.

APPENDIX C

Conditional Independence and Sufficiency

Consider an MCV problem in which forecasting systems F and G produce forecasts f and g, respectively, and the observing system X produces the corresponding observations x. In this context, conditional independence between G and X can be defined as follows: G and X are conditionally independent given F if and only if $p(x|f, g) = p(x|f)$. Under this condition, the basic trivariate distribution $p(f, g, x)$ can be written as follows:

$$p(f, g, x) = p(x|f, g)p(f, g)$$
$$= p(x|f)p(g|f)p(f)$$
$$= p(g|f)p(f, x)$$
$$= p(g|f)p(f|x)p(x). \quad \text{(C1)}$$

Summing both sides of (C1) over all values of f yields

$$\sum_f p(f, g, x) = p(g, x) = p(g|x)p(x)$$
$$= \sum_f p(g|f)p(f|x)p(x), \quad \text{(C2)}$$

or

$$p(g|x) = \sum_f p(g|f)p(f|x). \quad \text{(C3)}$$

Under the assumption that the stochastic transformation relating the two sets of forecasts is represented by the function $p(g|f)$, (C3) is identical to the definition of sufficiency (e.g., see Ehrendorfer and Murphy 1988, pp. 1758–1759). In other words, (C3) indicates that forecasting system F is sufficient for forecasting system G. Thus, conditional independence implies sufficiency. As noted in the text, the converse is not true; that is, sufficiency does not imply conditional independence.

REFERENCES

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

Clemen, R. T., and A. H. Murphy, 1986: Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships. *Wea. Forecasting*, **1**, 56–65.

——, and R. L. Winkler, 1987: Calibrating and combining precipitation probability forecasts. *Probability and Bayesian Statistics*, R. Viertl, Ed., Plenum, 97–110.

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

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

Krzysztofowicz, R., and D. Long, 1991: Beta likelihood models of probabilistic forecasts. *Int. J. Forecasting*, **7**, in press.

Murphy, A. H., 1989: Probability, statistics, and weather forecasting. *Idojaras*, **93**, 84–99.

——, 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.

——, Y.-S. Chen and R. T. Clemen, 1988: Statistical analysis of interrelationships between objective and subjective temperature forecasts. *Mon. Wea. Rev.*, **116**, 2121–2131.

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