2

Forecast verification

ALLAN H. MURPHY

1. Introduction

Forecast verification has had a long, colorful, and occasionally controversial history. The first substantial developments in verification methods were occasioned by the publication of a paper by J.P. Finley (Finley, 1884), in which he summarized some results of an experimental tornado forecasting program. As a measure of forecasting success, Finley calculated the sum of the percentages of correct forecasts of tornadoes and no tornadoes (multiplied by 100) and reported a percentage correct value of 96.6%. Finley's paper attracted the attention of several individuals (e.g., Doolittle, 1885; Gilbert, 1884; Peirce, 1884), and (*inter alia*) it was pointed out that always forecasting no tornadoes would have led to a percentage correct value of 98.2%! These and other individuals offered various criticisms of Finley's method of verification and proposed alternative measures of overall forecasting performance, most of which are still in use today. This signal event in the history of forecast verification — dubbed the *Finley affair* — is described in detail in a recent paper (Murphy, 1996).

The 100+ years since the Finley affair have witnessed the development of a relatively wide variety of methods and measures tailored to particular verification problems. Moreover, considerable attention has been devoted in recent years to studies related to the properties of various methods/measures. Space limitations preclude any attempt to review these developments here. Readers interested in the history of forecast verification, or the numerous methods and measures formulated since 1884, are referred to Brier and Allen (1951), Meglis (1960), Murphy and Daan (1985), Stanski, Wilson, and Burrows (1989), and Wilks (1995), as well as the references cited therein. Some of the methods developed during this period will be introduced in subsequent sections of the chapter.

Forecast verification is usually defined as the process of assessing the degree of correspondence between forecasts and observations (e.g., Murphy and Daan, 1985). In practice, the verification process has generally consisted of (i) calculating quantitative measures of one or two aspects of forecasting performance such as bias, accuracy, or skill (these and other terms are defined in Section 2) and (ii) drawing conclusions regarding absolute and/or relative performance on the basis of numerical values of these measures. Thus, traditional methods can be said to constitute a *measures-oriented (MO) approach* to the verification process.

In this chapter, forecast verification is defined as the process of assessing the quality of forecasts, where forecast quality is understood to consist of the totality of statistical characteristics embodied in the joint distribution of forecasts and observations (Murphy and Winkler, 1987). This definition leads to a *distributions-oriented (DO) approach* to the verification process, an approach that appears to offer clear theoretical and practical advantages over the traditional MO approach. For a recent case study contrasting the MO and DO approaches to forecast verification, see Brooks and Doswell (1996). The concepts, methods, and measures described here are based on this DO approach to forecast verification.

An important advantage of the DO approach is that it imposes some structure on verification problems themselves as well as on the body of verification methodology. With regard to the latter, it provides insight into the relationships among verification methods and measures, and it creates a sound basis for developing and/or choosing particular methods/measures in specific applications. From a practical point of view, this approach places particular emphasis on the assessment of various basic aspects of forecast quality, thereby providing the producers of forecasts (e.g., modelers, forecasters) with information that they need to guide their efforts to improve forecasting performance.

This chapter focuses on concepts and methods related to the verification of forecasts of weather variables. The body of verification methods described here is applicable to all types of forecasts — for example, probabilistic *and* nonprobabilistic forecasts — as long as they are evaluated on a location-by-location basis (or the forecasts from different locations are pooled to create a single verification data sample). The framework upon which these methods

are based is consistent with the multidimensional structure of verification problems and the multifaceted nature of forecast quality. These considerations are important in evaluating and comparing forecasting systems to avoid reaching inappropriate conclusions regarding absolute or relative forecasting performance. A distinction is made here between the problems of *absolute verification* (i.e., evaluating the forecasts produced by a single forecasting system) and *comparative verification* (i.e., evaluating and comparing the forecasts produced by two or more forecasting systems).

Section 2 describes the basic elements of the conceptual and methodological framework upon which the suite of DO verification methods to be presented here are based. These elements include the verification data sample, the joint distribution of forecasts and observations, the factorization of this distribution into conditional and marginal distributions, the complexity and dimensionality of verification problems, forecast quality and its various aspects, the nature of verification measures, and criteria for screening verification measures. General DO methods appropriate to the problems of absolute and comparative verification are described in some detail in Sections 3 and 4, respectively. To illustrate the use and interpretation of the basic suite of DO verification methods, Sections 3 and 4 include applications of these methods to samples of nonprobabilistic temperature and precipitation probability forecasts. Other verification problems and methods are briefly considered — and references to relevant works are provided — in Section 5. Section 6 contains a short discussion of the relationship between forecast quality and forecast value. Section 7 includes a summary and some concluding remarks.

2. Conceptual and methodological framework

2.1. Forecasts, observations, and verification data samples

The verification problems of interest here involve the evaluation and comparison of forecasting systems (forecasting methods, forecasting models, forecasters). These systems are denoted by F, G, ... Generic forecasts produced by systems F, G, ..., are denoted by f, g, ..., respectively, and these forecasts represent descriptions of future weather conditions.

To evaluate and compare forecasting systems, observations that describe the actual weather conditions (i.e., the weather conditions that occur at or during the valid time of the forecasts) must be available. Here, the observations are assumed to represent sets of weather situations characterized by variables denoted by X, Y, ... Generic values of these observations are denoted by x, y, ..., respectively.

The underlying weather variables, to which both the forecasts and observations refer, possess various basic characteristics. Some of these variables are continuous quantities (e.g., temperature, precipitation amount) and others are discrete quantities (e.g., cloud amount, precipitation type). However, all weather variables are generally treated as discrete — and defined in terms of a finite set of categories (of values) or events — for the purposes of prediction and evaluation. This practice is followed here, and the body of verification methods described in Sections 3 and 4 can be applied to forecasts of all (discrete) weather variables.

The forecasts themselves can be expressed in a probabilistic or nonprobabilistic format. Probabilistic forecasts indicate the likelihood of occurrence of the various possible categories or events that define the underlying variable, whereas nonprobabilistic forecasts specify that a particular category or event will occur (with complete certainty). The observations are assumed to be expressed in a nonprobabilistic format.

In the case of nonprobabilistic forecasts, the sets of possible forecasts and observations are generally identical. However, in the case of probabilistic forecasts, the set of distinct forecasts is usually larger (often considerably larger) than the set of observations. The implications of this difference for the verification process are discussed in Section 2.4.

Traditionally, the *verification data sample* associated with forecasting system F and forecasting situations X is viewed as a sample of n matched pairs of forecasts and observations $\{(f_i, x_i),\ i = 1, \ldots, n\}$, where n is the sample size. In this chapter the verification data sample is arranged instead by distinct pairs of forecasts and observations. For example, a sample of temperature forecasts and observations might contain some occasions on which the forecast temperature is 60°F and the observed temperature is 58°F, as well as other occasions on which the forecast temperature is 60°F and the observed temperature is 63°F. These two subsam-

ples correspond to distinct combinations of forecast and observed temperatures. The phrase "all (f, x)-pairs" is used to denote this arrangement of the verification data sample.

It is convenient here to introduce some basic notation to describe the verification data sample. Let n^f and n^x denote the number of distinct forecasts and observations, respectively. Moreover, let n_{ij} denote the joint frequency of forecast f_i and observation x_j $(i = 1, \ldots, n^f;\ j = 1, \ldots, n^x)$. Then the n^f-by-n^x matrix $\underline{N} = (n_{ij})$ contains the joint frequencies of all (f, x)-pairs in the verification data sample. The marginal frequencies of forecast f_i and observation x_j are denoted by $n_i^f = \sum_j n_{ij}$ and $n_j^x = \sum_i n_{ij}$, respectively $(\sum_i n_i^f = \sum_j n_j^x = \sum_i \sum_j n_{ij} = n)$.

2.2. Joint distribution of forecasts and observations

The joint distribution of forecasts and observations, denoted here by $p(f, x)$, constitutes the basic framework for the DO approach to forecast verification (Murphy and Winkler, 1987). This joint distribution contains information about the forecasts, about the observations, and about the relationship between the forecasts and observations. Under the assumptions that the bivariate time series of forecasts and observations is (i) serially independent and (ii) stationary in a statistical sense, the distribution $p(f, x)$ contains all the relevant information in the verification data sample. *Forecast quality*, as defined here, is the totality of the statistical characteristics of the forecasts, the observations, and their relationship embodied in this distribution.

Let p_{ij} denote the joint relative frequency of forecast f_i and observation x_j in the verification data sample $(i = 1, \ldots, n^f;\ j = 1, \ldots, n^x)$. It follows that $p_{ij} = n_{ij}/n$. Moreover, the n^f-by-n^x matrix $\underline{P} = (p_{ij})$ contains the joint relative frequencies of all (f, x)-pairs. These joint relative frequencies are treated here as joint probabilities (i.e., the estimation problem is ignored). Thus, $p_{ij} = \Pr(f = f_i, x = x_j)$, and the matrix $\underline{P}$ contains the joint probabilities of all (f, x)-pairs. These probabilities can be viewed as parameters of a primitive model of the joint distribution $p(f, x)$ (see Section 2.5).

As noted above, the joint distribution $p(f, x)$ plays a fundamental role in the verification process. In essence, forecast verification

consists of describing and summarizing the statistical characteristics of this distribution. This process is facilitated by factoring the joint distribution into conditional and marginal distributions.

2.3. Factorizations of joint distribution: conditional and marginal distributions

The information relevant to forecast quality contained in the joint distribution $p(f, x)$ becomes more accessible when this distribution is factored into conditional and marginal distributions. Two such factorizations can be defined:

$$p(f, x) = q(x|f)s(f) \tag{2.1}$$

and

$$p(f, x) = r(f|x)t(x) \tag{2.2}$$

(Murphy and Winkler, 1987). The expression in equation (2.1), the *calibration-refinement (CR) factorization* of $p(f, x)$, involves the conditional distributions of the observations given the forecasts, $q(x|f)$, and the marginal distribution of the forecasts, $s(f)$. A conditional distribution $q(x|f)$ can be defined for each of the n^f possible forecasts, with $q(x_j|f_i) = q_{ij} = \Pr(x = x_j|f = f_i)$. These conditional distributions can be depicted in the form of an n^f-by-n^x matrix $\underline{Q} = (q_{ij})$, where the elements of the ith row are the components of the conditional distribution $q(x|f_i)$. The marginal distribution $s(f)$ specifies the unconditional probabilities of the n^f forecasts. That is, $s(f_i) = s_i = \Pr(f = f_i)$ $[= \Sigma_j p_{ij} (j = 1, \ldots, n^x)]$. This distribution can be depicted in the form of a n^f-by-1 vector $\underline{s} = (s_i)$.

The expression in equation (2.2), the *likelihood-base rate (LBR) factorization* of $p(f, x)$, involves the conditional distributions of the forecasts given the observations, $r(f|x)$, and the marginal distribution of the observations, $t(x)$. A conditional distribution $r(f|x)$ can be defined for each of the n^x possible observations, with $r(f_i|x_j) = r_{ij} = \Pr(f = f_i|x = x_j)$. These conditional distributions can be depicted in the form of an n^f-by-n^x matrix $\underline{R} = (r_{ij})$, where the elements of the jth column are the components of the distribution $r(f|x_j)$. The marginal distribution $t(x)$ specifies the unconditional probabilities of the n^x observations. That is, $t(x_j) = t_j =$

$\Pr(x = x_j)$ $[= \Sigma_i p_{ij} (i = 1, \ldots, n^f)]$. This distribution can be depicted in the form of a 1-by-n^x vector $\underline{t} = (t_j)$.

Since the joint distribution $p(f, x)$ can be reconstructed from the components on the right-hand side of either equation (2.1) or equation (2.2), forecast verification based on either factorization is equivalent to forecast verification based on $p(f, x)$ itself. In fact, DO verification can be conducted within frameworks based on $p(f, x)$, $q(x|f)$ and $s(f)$, or $r(f|x)$ and $t(x)$. However, since the three DO frameworks provide insights into different aspects of forecast quality (see Section 2.6), it is more appropriate to view these frameworks as complementary rather than alternative approaches to verification problems. Of course, the fundamental equivalence of the frameworks embodied in equations (2.1) and (2.2) implies that some relationships must exist among these various aspects of quality (see Murphy and Winkler, 1987).

2.4. Verification problems and their complexity

Verification problems are of two basic types: (i) *absolute verification* and (ii) *comparative verification* (Murphy, 1991). Absolute verification problems are concerned with the evaluation of forecasts produced by individual forecasting systems. In the case of a forecasting system F and forecasting situations described by X, absolute forecast verification is based on the joint distribution $p(f, x)$ and the conditional and marginal distributions associated with the factorizations of $p(f, x)$. Absolute verification also includes the comparison of the forecasts of interest with naive forecasts derived (solely) from the marginal distribution of observations $t(x)$ (e.g., forecasts based on sample climatology or possibly persistence).

Comparative verification problems involve the evaluation and comparison of two or more forecasting systems. If two forecasting systems, denoted by F and G, formulate forecasts for the same set of forecasting situations (i.e., a common set of situations described by X), then comparative verification consists of evaluating and comparing $p_F(f, x)$ and $p_G(g, x)$ and the components of their respective factorizations. On the other hand, if F's and G's forecasts relate to different forecasting situations (i.e., different sets of situations X and Y), then comparative verification consists of evaluating and comparing $p_{F,X}(f, x)$ and $p_{G,Y}(g, y)$, as well as the

components of their respective factorizations. These two types of comparative verification are referred to as *matched comparative verification* and *unmatched comparative verification.*

The complexity of verification problems can be characterized by the number of distinct sets of quantities (i.e., forecasts and observations) associated with the underlying joint distribution(s). According to this definition, absolute verification problems involve two sets of quantities and comparative verification problems involve three or four (or more) sets of quantities. Thus, the latter are more complex than the former (Murphy, 1991). Moreover, within the context of comparative verification, unmatched comparative verification is more complex than matched comparative verification. The treatment of comparative verification in this chapter is restricted to the case of matched comparative verification.

Reductions in the complexity of verification problems can be achieved by decreasing the number of these basic quantities. For example, a considerable reduction in the complexity of absolute verification problems can be accomplished by assuming that attention can be focused on the (univariate) distribution of forecast errors, $u(e)$ $[= u(f - x)]$, rather than the joint distribution of forecasts and observations, $p(f, x)$. Whether or not such a strong assumption is warranted depends on the verification problem at hand (including the nature of the underlying variable). In any case, a large amount of potentially useful information regarding various aspects of forecast quality becomes inaccessible when absolute verification is based on $u(e)$ rather than on $p(f, x)$.

2.5. Models of basic distributions: dimensionality of verification problems

The dimensionality (d) of an absolute verification problem relates to the number of probabilities (or parameters) that must be determined in order to reconstruct the basic joint distribution $p(f, x)$ (Murphy, 1991). In situations in which this distribution is described by the elements of the matrix of joint probabilities, $\underline{P} = (p_{ij})$,

$$d = n^f n^x - 1, \tag{2.3}$$

since the sum over all elements in $\underline{P}$ must equal unity. Equivalent definitions of d can be formulated in terms of the elements of the

matrix $\underline{Q}$ or $\underline{R}$ and vector $\underline{s}$ or $\underline{t}$ describing the conditional and marginal distributions, respectively.

Thus, an absolute verification problem involving nonprobabilistic forecasts for a two-category (or dichotomous) variable is a three-dimensional problem ($d = 2 \times 2 - 1$). That is, it is necessary to determine three joint probabilities (or, for example, two conditional probabilities and one marginal probability) to describe forecast quality completely in this situation. On the other hand, an absolute verification problem involving probabilistic forecasts for a dichotomous variable, in which 11 equally spaced probability values are used, is a 21-dimensional problem ($d = 11 \times 2 - 1$). Twenty-one joint probabilities (or, for example, 11 conditional probabilities and 10 marginal probabilities) must be determined in order to describe forecast quality completely in this situation. The corresponding comparative verification problems possess considerably higher (approximately two or more times higher) dimensionality, since they involve two or more joint distributions.

When viewed from this perspective, most verification problems possess relatively high dimensionality. In general, the higher the dimensionality of a verification problem, the more quantities (i.e., joint, conditional, and/or marginal probabilities, or numerical measures) must be determined in order to provide a complete — or even an adequate — description of forecast quality. Clearly, a single overall measure of forecasting performance, such as the mean square error, cannot provide a complete — or even a very insightful or useful — description of forecast quality in most problems. Considerations related to the dimensionality of verification problems have generally been ignored in the traditional MO approach to forecast evaluation. To provide a reasonably complete description of forecast quality, the dimensionality of verification problems must be respected.

The description of the distribution $p(f, x)$ in terms of joint probabilities represents a model — albeit a primitive model — of this distribution (equivalent statements could be made with respect to the probabilities that constitute the conditional and marginal distributions). This primitive model is consistent with the usual approach to forecast verification, in the sense that the verification process is based on the empirical frequencies (or relative frequencies) derived from the verification data sample. Since these primitive models generally possess many parameters (in this case,

joint, conditional, and/or marginal probabilities), the verification process suffers from the "curse of dimensionality."

Although the verification methods to be described in this chapter are based — explicitly or implicitly — on this primitive model, it may be useful to consider briefly the possibility of reducing the dimensionality of verification problems by modeling the basic distributions with parametric statistical models. If acceptable fits to the basic (empirical) distributions could be obtained, then such models would provide parsimonious descriptions of the corresponding verification data samples. Forecast quality could then be characterized in terms of a relatively few model parameters, thereby substantially reducing the dimensionality of these verification problems. Moreover, characterizing forecast quality in terms of the parameters of one or more statistical models should reduce the undesirable effects of sampling variability on assessments of forecast quality (and its aspects). Currently, these effects are generally ignored, even though it is recognized that another verification data sample acquired under similar conditions will yield at least somewhat different results. Reducing the effects of sampling variability would lead to more credible estimates of the various aspects of forecast quality. Finally, the availability of models of forecast quality should facilitate studies of the relationship between forecast quality and forecast value (e.g., Katz and Murphy, 1990; Katz, Murphy, and Winkler, 1982; see also Chapter 6 in this volume).

Most studies in which statistical models have been used to describe forecast quality are of relatively recent vintage. Moreover, many of these models were identified in the context of forecast-value studies rather than in the context of forecast verification problems. For example, Katz et al. (1982) used a bivariate Gaussian model to characterize the relationship between daily minimum temperature forecasts expressed in a nonprobabilistic format and the corresponding observations in the context of a fruit-frost decision-making problem. In this case a potential verification problem of relatively high dimensionality was reduced to no more than 5 dimensions ($d \leq 5$), represented by two means, two variances, and one covariance (or correlation coefficient).

More recently, Krzysztofowicz and Long (1991b) investigated the use of the sufficiency relation (see Section 4.1) as a means of comparing the performance of objective and subjective precipita-

tion probability forecasts. To simplify the process of evaluation, they used beta distributions to fit the conditional distributions (or likelihoods) $r(f|x = 1)$ and $r(f|x = 0)$ for both types of forecasts. This approach reduced a comparative verification problem of approximately 40 dimensions to a problem of 8 dimensions ($d = 8$), represented by 4 parameters associated with the conditional distributions for each type of forecast.

As noted previously, it may not always be possible to identify mathematically convenient statistical models that fit the relevant joint, conditional, and/or marginal distributions in a satisfactory manner. For example, Clemen and Winkler (1987) used a Gaussian log-odds model to fit the likelihood functions — that is, $r(f|x = 0)$ and $r(f|x = 1)$ — for 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 summary, the use of statistical models offers a means of describing forecast quality in a parsimonious manner, thereby simplifying many verification problems. This approach — a substantial departure from the traditional approach (involving the use of empirical relative frequencies) adopted here — clearly warrants further investigation. Some results of a recent effort to use statistical models to reduce the dimensionality of verification problems and the effects of sampling variability, in the context of precipitation probability forecasting, are reported by Murphy and Wilks (1996).

2.6. Forecast quality and its aspects

As noted in Section 2.2, forecast quality is defined as the totality of the statistical characteristics of the forecasts, the observations, and their relationship embodied in the joint distribution $p(f, x)$. Moreover, specific aspects of quality can be related to $p(f, x)$ or to the conditional and marginal distributions associated with its factorizations. These aspects of quality are identified and defined in Table 2.1. This table also indicates the basic distributions associated with each aspect of quality.

Bias (also referred to as systematic or unconditional bias) relates to the degree of correspondence between the average forecast

Table 2.1. Names and definitions of aspects of quality, including basic distribution(s) related to each aspect

Name	Definition	Basic distribution(s)
Bias	Degree to which μ_f corresponds to μ_x	$s(f)$ and $t(x)$
Association	Overall strength of linear relationship between f and x	$p(f, x)$
Accuracy	Average degree of correspondence between f and x	$p(f, x)$
Skill (relative accuracy)	Accuracy of forecasts relative to accuracy of forecasts based on standard of reference (e.g., climatology)	$p(f, x)$
Type 1 conditional bias (reliability, calibration)	Degree of correspondence between $\mu_{x\|f}$ and f, averaged over all values of f	$q(x\|f)$ and $s(f)$
Resolution	Degree to which $\mu_{x\|f}$ differs from μ_x, averaged over all values of f	$q(x\|f)$ and $s(f)$
Sharpness (refinement)	Degree to which probability forecasts approach zero or one	$s(f)$
Type 2 conditional bias	Degree of correspondence between $\mu_{f\|x}$ and x, averaged over all values of x	$r(f\|x)$ and $t(x)$
Discrimination	Degree to which $\mu_{f\|x}$ differs from μ_f, averaged over all values of x	$r(f\|x)$ and $t(x)$
Uncertainty (variability)	Degree of variability in observations	$t(x)$

μ_f and the average observation μ_x. This aspect of quality is generally measured in terms of the difference between μ_f and μ_x. For example, if $\mu_f = 65.4°\text{F}$ and $\mu_x = 63.8°\text{F}$ for a sample of nonprobabilistic temperature forecasts, then these forecasts exhibit a positive bias of 1.6°F.

Association refers to the strength of the linear relationship between the forecasts and observations. This aspect of quality is usually measured by the correlation coefficient ρ_{fx}. The square of ρ_{fx} represents the proportionate reduction in the variance of the observations when they are regressed on the forecasts.

Accuracy relates to the average degree of correspondence between individual forecasts and observations in the verification data sample. It is generally defined in terms of the joint distribution $p(f, x)$, but it also can be defined in terms of conditional and marginal distributions. In the context of forecast verification, common measures of accuracy include the mean square error or mean absolute error in the case of (essentially) continuous variables and the fraction (or percentage) of correct forecasts in the case of discrete variables.

Skill is usually defined as the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by a naive forecasting system such as climatology or persistence. Skill scores, defined as measures of relative accuracy, are used to assess this aspect of quality. Positive (negative) skill scores indicate that the accuracy of the forecasts of interest is greater (less) than the accuracy of the forecasts produced by the standard of reference. In the case of operational temperature forecasts that possess a mean square error of 3.6 $(°\mathrm{F})^2$, when temperature forecasts based on persistence possess a mean square error of 5.4 $(°\mathrm{F})^2$, the skill of the forecasts is positive with a numerical skill score value of 0.333 $[= 1 - (3.6/5.4)]$.

Reliability (calibration or type 1 conditional bias) relates to the degree of correspondence between the mean observation given a particular forecast, $\mu_{x|f}$, and the forecast f. Suppose that $\mu_{x|f} = 0.424$ for a subsample of precipitation probability forecasts for which $f = 0.40$. Clearly, this subsample of forecasts is not perfectly reliable (or conditionally unbiased). Forecasts that exhibit perfect correspondence between $\mu_{x|f}$ and f over all values of f are said to be completely reliable (or, equivalently, well-calibrated or conditionally unbiased overall). Forecasts that are completely reliable are necessarily unbiased (but the converse relationship does not hold).

Resolution relates to the difference between the mean observation given a particular forecast, $\mu_{x|f}$, and the overall unconditional mean observation μ_x. A verification data sample for which

$\mu_{x|f} = \mu_x$ for all f is completely lacking in resolution. Thus, larger differences between $\mu_{x|f}$ and μ_x are preferred to smaller differences. Resolution, as an aspect of forecast quality, is based on the concept that "different forecasts should be followed by different observations."

Sharpness (refinement) is an aspect of forecast quality that applies only to probabilistic forecasts. Such forecasts are perfectly sharp (refined) if only probabilities of zero and one are used in the forecasts. (Nonprobabilistic forecasts are always perfectly sharp.) On the other hand, constant forecasts of the climatological probability are completely lacking in sharpness. Since the degree of sharpness increases as more frequent use is made of relatively high and low probabilities, the variability of the forecasts is an indicator of sharpness. Forecasts involving probabilities of zero and one maximize this variability. Not surprisingly, the variance of the distribution of forecasts, σ_f^2, is frequently used as a one-dimensional measure of sharpness. When the forecasts of interest are completely reliable, sharpness and resolution become identical aspects of quality.

Type 2 conditional bias relates to the degree of correspondence between the mean forecast given a particular observation, $\mu_{f|x}$, and the observation x. Suppose that, in the case of precipitation probability forecasting systems A and B, $\mu_{f|x}(A) = 0.72$ and $\mu_{f|x}(B) = 0.68$ when $x = 1$ and $\mu_{f|x}(A) = 0.24$ and $\mu_{f|x}(B) = 0.32$ when $x = 0$. Then system A's forecasts are less conditionally biased in the type 2 sense than system B's forecasts. Forecasts that exhibit complete correspondence between $\mu_{f|x} = x$ for all x are said to be completely conditionally unbiased in this sense (probabilistic forecasts that satisfy this condition are necessarily perfect forecasts).

Discrimination relates to the difference between the mean forecast given a particular observation, $\mu_{f|x}$, and the overall mean forecast μ_f. In the hypothetical situation involving forecasting systems A and B considered above, A's forecasts exhibit greater discrimination than B's forecasts. When $\mu_{f|x} = \mu_f$ for all x the verification data sample is completely lacking in discrimination. Thus, larger differences between $\mu_{f|x}$ and μ_f are preferable to smaller differences.

Uncertainty relates to the variability of the observations (as primitive descriptors of the forecasting situations). The variance

of the distribution of observations, σ_x^2, is sometimes used as a measure of uncertainty. It should be noted that this aspect of quality does not depend in any way on the forecasts. Thus, uncertainty is a characteristic of the observations, or forecasting situations, rather than a characteristic of the forecasts.

2.7. Measures of aspects of quality: verification measures

A *verification measure* is defined here as any (mathematical) function of the forecasts, the observations, or their relationship. Thus, the variance of the forecasts and the variance of the observations are verification measures, even though they are not concerned directly with the correspondence between forecasts and observations. Clearly, these two variances — and their relative magnitude — are of considerable interest in the context of many verification problems.

A *performance measure* is defined here as a verification measure that focuses on the correspondence between forecasts and observations, on either an individual or collective basis. The mean square error (a measure of accuracy), as well as the measure of bias defined as the difference between the mean forecast and the mean observation, represent examples of performance measures. Obviously, performance measures constitute a subset of the set of all verification measures.

A *scoring rule* is a performance measure that is defined for individual pairs of forecasts and observations. Thus, the mean square error is a scoring rule, whereas the measure of bias defined in the previous paragraph is not a scoring rule. Scoring rules represent a subset of the set of performance measures.

The set of all verification measures is essentially infinite. For the purposes of this chapter, two verification measures are considered to be equivalent if they are linearly related. This definition of equivalence is used in conjunction with the discussion of screening criteria for verification measures (see Section 2.8). Other definitions of equivalence may be used in other contexts.

As defined, verification measures in general — and performance measures in particular — may possess positive or negative orientations. In the case of performance measures, positive (negative) orientation implies that larger (smaller) scores are indicative of

better performance (with regard to the aspect of quality of interest). Since a linear transformation of a performance measure leads to an equivalent measure, it is always possible to change the orientation of a performance measure from negative to positive (or vice versa).

It is sometimes desirable to establish a standard range of values (or scores) for a verification or performance measure. If the range of values of a measure are finite, then it can be transformed linearly in such a way as to produce any desired range of values. Ranges frequently considered desirable include [0, 1] with one (zero) representing the best (worst) possible score or $[-1, +1]$ with plus (minus) one representing the best (worst) possible score. Some measures (e.g., skill scores; see Section 3.2) do not possess finite ranges of values; in this case, it is not possible to transform the measures in such a way as to obtain a desired finite range of values.

2.8. Criteria for screening verification measures

Four criteria that can be used to screen alternative verification measures are briefly described in this section. Each criterion relates to a desirable property or characteristic that a particular measure may or may not possess. Application of screening criteria is not intended to identify the single best measure (this goal is generally inappropriate and unattainable), but rather to eliminate from further consideration those measures that do not possess one or more of these desirable characteristics.

Sufficiency. The concept of sufficiency — and the sufficiency relation — will be considered in detail in conjunction with the discussion of comparative evaluation of forecasting systems in Section 4 (see Section 4.1). In brief, the sufficiency relation identifies the conditions under which one forecasting system can be judged to be unambiguously superior to another forecasting system (i.e., superior in terms of both quality and value). Under certain conditions, it may be possible to use this relation as a criterion for screening verification measures. Verification measures that are consistent with the sufficiency relation are generally preferred to measures that are not consistent with this relation.

For example, Krzysztofowicz (1992) has shown that in the case of nonprobabilistic forecasts for a continuous variable with a Gaus-

sian distribution, it is possible to formulate a one-dimensional measure of quality — the so-called Bayesian correlation score (BCS) — that is consistent with the sufficiency relation. That is, the values of the BCS order forecasting systems by their relative quality and value (i.e., a larger score indicates higher quality and greater value). In situations in which these conditions and assumptions are satisfied, the BCS obviously offers advantages over alternative measures as a basis for comparative verification. It remains to be seen whether or not it is possible to formulate other one-dimensional measures that, under specific conditions and/or assumptions, are consistent with the sufficiency relation. Until such measures have been defined, this concept will remain of limited use as a screening criterion for verification measures.

Propriety. This screening criterion is based on the principle that a verification measure should not encourage forecasters to make forecasts that differ from their true judgments (differences between forecasts and judgments are indicative of "hedging"). It applies to a particular class of verification measures called "scoring rules" (see Section 2.7). Since forecasters' judgments are inherently probabilistic, this principle is applicable only in contexts in which forecasts are expressed in a probabilistic format. Moreover, the fact that the propriety criterion relates to the correspondence between forecasters' judgments and their forecasts implies that it is especially relevant in contexts involving *subjective* probabilistic forecasts.

Consider a situation involving an underlying variable whose range of values has been divided into m mutually exclusive and collectively exhaustive (m.e.c.e.) categories or events. Let $S = S(f, x)$ denote a generic scoring rule with positive orientation (i.e., larger scores are better), and let p_k denote the forecaster's judgment concerning the likelihood of occurrence of the kth event $(k = 1, \ldots, m)$. Further, let $S_k(f)$ denote the score assigned (by S) to the forecast f when the kth event $(x = x_k)$ occurs and let $E[S(f, p)]$ denote the forecaster's subjective expected score, where

$$E[S(f, p)] = \sum_k p_k S_k(f). \tag{2.4}$$

$E[S(f, p)]$ in equation (2.4) is a subjective expected score in the sense that it represents the weighted average of the actual scores [i.e., the $S_k(f)$ for $k = 1, \ldots, m$], where the weights are the forecaster's subjective probabilities that she will receive these scores.

The scoring rule $S(f,x)$ is *strictly proper* if $E[S(p,p)] > E[S(f,p)]$ for all $f \neq p$, *proper* if $E[S(p,p)] \geq E[S(f,p)]$ for all f, and *improper* if $E[S(p,p)] < E[S(f,p)]$ for some $f \neq p$ (Murphy and Daan, 1985; Winkler and Murphy, 1968). Thus, a scoring rule is strictly proper if the maximum expected score can be achieved only when the forecaster makes her forecast f correspond exactly to her judgment p. A scoring rule is proper if forecasts other than $f = p$ lead to the same maximum expected score. Finally, a scoring rule is improper if the maximum expected score is achieved by making a forecast $f \neq p$.

Examples of scoring rules that are strictly proper include the Brier or quadratic score, the logarithmic score, and the spherical score (Winkler and Murphy, 1968). Skill scores based on the Brier score and defined in the usual way (see Section 3.2) are approximately strictly proper for large data samples (Murphy, 1973). The expression for (expected) forecast value in Section 6 is an example of a proper scoring rule. Finally, the so-called linear scoring rule, defined in terms of the absolute difference between forecast probabilities and event occurrence or nonoccurrence (in the case of a dichotomous variable), is an example of an improper scoring rule. This scoring rule leads to an extreme form of hedging, in that forecasters are encouraged to transform their probabilistic judgments into nonprobabilistic forecasts.

Use of the propriety criterion provides a means of screening alternative scoring rules as overall measures of the quality of probabilistic forecasts. Clearly, it is desirable to restrict the choice of such measures to the class of strictly proper scoring rules whenever possible and to avoid the use of improper scoring rules.

Consistency. Although the concept of propriety cannot be applied to nonprobabilistic forecasts, a related but weaker concept called consistency can be invoked in contexts involving such forecasts. This concept is especially applicable in situations involving nonprobabilistic forecasts of continuous variables. It is based on the premise that forecasters follow — or are instructed to follow — a specific directive (or rule) when translating their probabilistic judgments into nonprobabilistic forecasts.

The consistency criterion states that the primary measure used to verify the forecasts should be consistent with the directive followed by the forecaster (Murphy and Daan, 1985). For example, if the directive states "forecast the mean value of your judgmental

probability distribution," then the mean square error is an appropriate — that is, consistent — verification measure. This measure is consistent with the directive because the score assigned by the mean square error is minimized by choosing the mean of the distribution as the forecast. On the other hand, if the directive states "forecast the median value of your subjective probability distribution," then the mean absolute error would be a consistent measure. This measure is minimized by forecasting the median of the forecaster's distribution. Thus, the consistency concept can be used to screen alternative measures of forecast quality for nonprobabilistic forecasts, identifying those measures that are — and are not — consistent with the directive given to the forecaster.

Equitability. The concept of equitability applies to verification measures involving nonprobabilistic forecasts of discrete variables. This concept is based on the principle that constant forecasts of any event — as well as forecasts produced by a procedure in which the forecast event is chosen at random — should receive the same expected score (Gandin and Murphy, 1992). According to this principle, the expected scores attained by constant forecasts of events that occur frequently and constant forecasts of events that occur infrequently should be identical.

An example of a measure that satisfies this criterion in dichotomous situations is Kuiper's performance index (see Murphy and Daan, 1985; Wilks, 1995). Verification measures that do not satisfy this criterion include the fraction of correct forecasts, as well the threat score or critical success index. The concept of equitability does not provide a unique solution in the case of polychotomous situations, so that additional assumptions are required to apply this criterion to screen verification measures in such problems.

Application of screening criteria. As previously noted, application of these screening criteria is generally not intended to identify a single best verification measure. The objective is instead to identify those measures that do — and those that do not — satisfy the criteria. Presumably, measures that are judged acceptable on the basis of such screening tests would be preferred to measures that are judged unacceptable. It may be possible to identify other criteria that can be added to the current set, thereby reducing still further the class of acceptable measures. In any case, it is important to keep in mind that no single verification measure can assess all potentially relevant aspects of forecast quality.

3. Absolute verification: methods and applications

Consideration of the problem of absolute verification from a DO perspective leads to the identification of three classes of verification methods: (i) the basic joint, conditional, and marginal distributions themselves; (ii) summary measures of these distributions; and (iii) measures of various aspects of forecast quality. These methods are defined in this section, and their use and interpretation are illustrated by applying the methods to samples of nonprobabilistic maximum temperature (Tmax) forecasts for Minneapolis, Minnesota (Murphy, Brown, and Chen, 1989), and probability of precipitation (PoP) forecasts for St. Louis, Missouri (Murphy and Winkler, 1992).

Both verification data samples include forecasts of two types: so-called objective forecasts produced by numerical-statistical models, and subjective forecasts formulated by U.S. National Weather Service forecasters. Here we focus on the evaluation of various basic aspects of the quality of the objective and subjective forecasts separately. Use of these verification methods to compare the two types of forecasts is described in Section 4.2.

The verification methods and measures defined in this section constitute tailored versions of a body of evaluation methodology that is applicable in principle to all verification data samples. In particular, this set of methods and measures is consistent with the multidimensional structure of verification problems and the multifaceted nature of forecast quality. Brief discussions of the way in which these methods/measures can be tailored to verification problems involving different types of forecasts — as well as overviews of some other verification methods — are included in Section 5.

3.1. Basic distributions and summary measures

In view of the fact that the joint distribution $p(f, x)$ contains all the information relevant to forecast quality (see Section 2.2), absolute verification should begin with an examination of this distribution. Evaluation of $p(f, x)$ provides overall insight into the relationship between forecasts and observations. The joint distribution can be depicted in several different ways. For example, it can be displayed graphically in the form of a scatter diagram or a bivariate

histogram or numerically in the form of a contingency table. Bivariate histograms for samples of 24-hour nonprobabilistic Tmax forecasts are presented in Figure 2.1. In this case, strong relationships between f and x appear to exist for both types of forecasts. Further insight can be obtained from careful examination of these diagrams. For example, Figure 2.1a reveals a tendency toward overforecasting in the case of the objective forecasts (i.e., forecast temperatures exceed observed temperatures more often than vice versa).

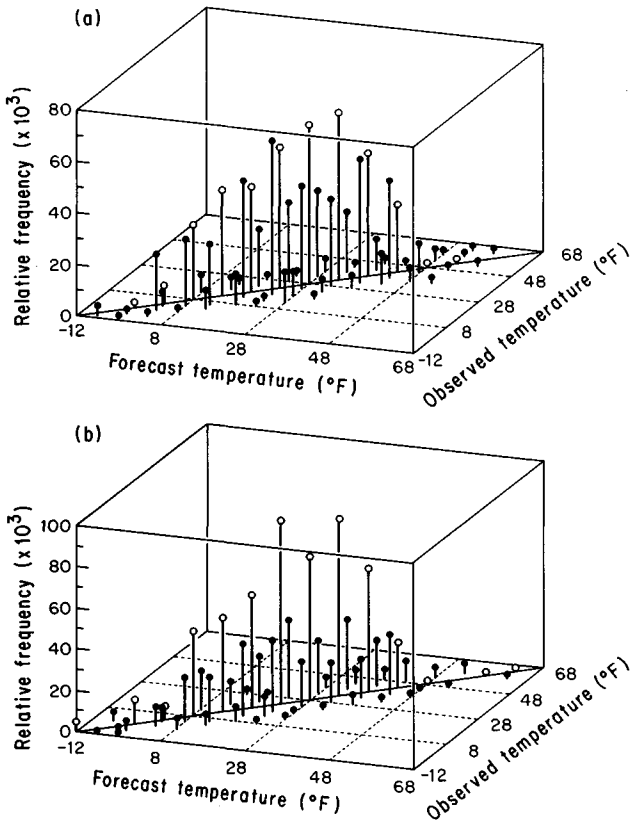

Figure 2.1. Bivariate histograms displaying $p(f, x)$ for 24-hour maximum temperature forecasts in the winter season for Minneapolis, Minnesota (open circles represent case of $f = x$): (a) objective forecasts; (b) subjective forecasts. (From Murphy, Brown, and Chen, 1989)

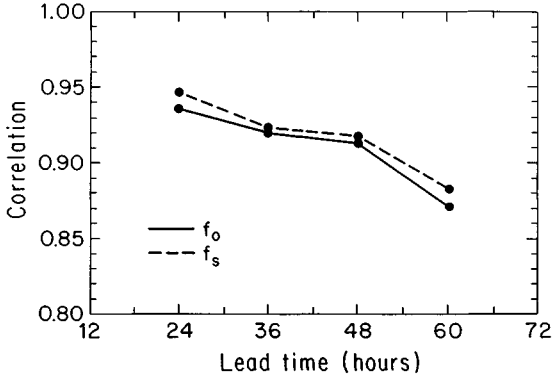

Figure 2.2. Line diagram displaying ρ_{fx} as a function of lead time for objective (f_o) and subjective (f_s) maximum temperature forecasts in the winter season for Minneapolis, Minnesota. (From Murphy, Brown, and Chen, 1989)

The correlation coefficient ρ_{fx} is a (one-dimensional) summary measure of the joint distribution $p(f, x)$. A line diagram displaying ρ_{fx} as a function of lead time for both types of Tmax forecasts is presented in Figure 2.2. Correlations approach 0.95 for the 24-hour forecasts and decrease monotonically as lead time increases. When ρ_{fx} is employed as a quantitative measure of forecasting performance (as opposed to a summary measure), it is important to keep in mind that this measure ignores any unconditional or conditional biases in the forecasts (see Section 3.2).

Information regarding various basic aspects of forecast quality can be obtained by examining the conditional and marginal distributions. In view of the factorizations of $p(f, x)$ set forth in equations (2.1) and (2.2), it seems appropriate to consider the components of the respective factorizations together. The insights provided by these distributions are illustrated here by evaluating samples of PoP forecasts, as well as the samples of nonprobabilistic Tmax forecasts already introduced.

Conditional quantiles of the distributions $q(x|f)$ for the Tmax forecasts are depicted in Figure 2.3. If the curve representing the conditional medians is taken to be a close approximation to the conditional means (a reasonable assumption in this case), comparison of this curve with the 45° line provides insight into the reliability of these forecasts. In this regard, the objective forecasts (Figure 2.3a) exhibit some overforecasting, at least for relatively

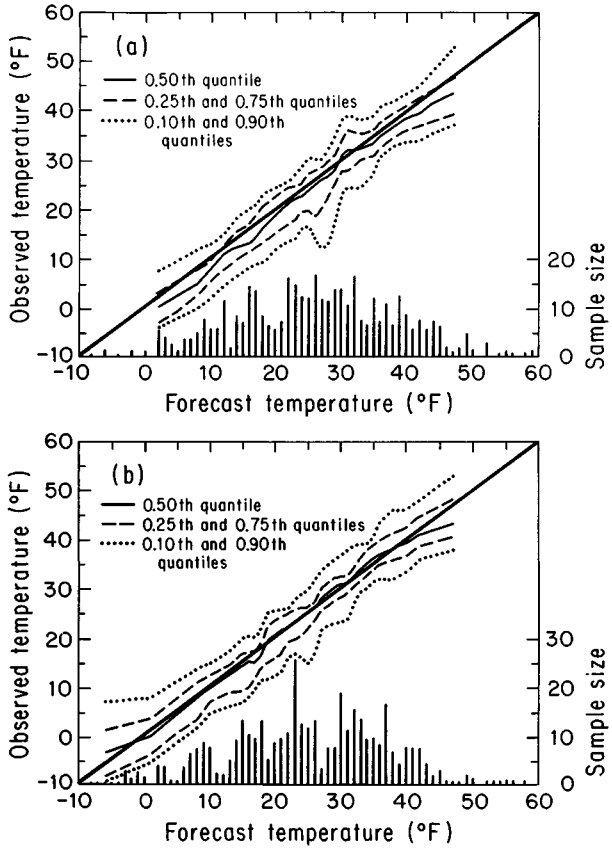

Figure 2.3. Quantiles of conditional distributions $q(x|f)$ (above), and marginal distribution $s(f)$ (below), for 24-hour maximum temperature forecasts in the winter season for Minneapolis, Minnesota: (a) objective forecasts; (b) subjective forecasts. (From Murphy, Brown, and Chen, 1989)

high and low Tmax forecasts. The subjective forecasts (Figure 2.3b) appear to be quite reliable over the entire range of forecast temperatures. Similar diagrams can be produced for the conditional distributions $r(f|x)$; they are omitted here because of space considerations.

The spread of the conditional quantiles (i.e., the difference between the 0.25th and 0.75th quantiles, or between the 0.10th and 0.90th quantiles) provides insight into the accuracy of the Tmax forecasts as a function of the forecast temperature. This spread

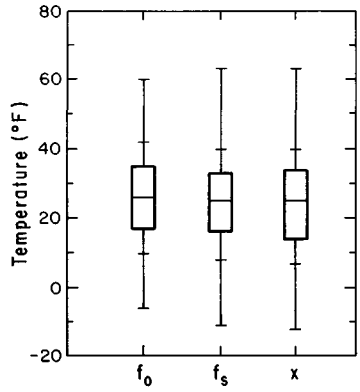

Figure 2.4. Box plots of marginal distributions for 24-hour maximum temperature forecasts — and the corresponding observations — in the winter season for Minneapolis, Minnesota. (From Murphy, Brown, and Chen, 1989)

exhibits no clear pattern of behavior in the case of the objective forecasts (Figure 2.3a). In the case of the subjective forecasts (Figure 2.3b), however, the spread appears to be smaller for intermediate forecasts than it is for relatively high or low forecasts.

The diagrams in Figure 2.3 also contain, in the form of bar charts, the marginal distributions of the forecasts [i.e., $s(f)$]. These distributions, together with the marginal distribution of observations, $t(x)$, are summarized in the form of box plots in Figure 2.4. Among other things, these plots confirm the existence of a bias in the objective Tmax forecasts (initially identified in the bivariate histogram). In addition, the box plots indicate that the variability of the observations exceeds the variability of both types of forecasts.

Conditional distributions $q(x|f)$ and marginal distributions $s(f)$ for matched samples of PoP forecasts are depicted in Figure 2.5. Since these forecasts relate to a binary variable (i.e., $x = 1$ or $x = 0$), each conditional distribution $q(x|f)$ contains only one independent probability [i.e., $q(x = 1|f) + q(x = 0|f) = 1$ for each f]. As a result, it is reasonable to begin the verification process in this case with the conditional and marginal distributions (rather than the joint distribution). Moreover, it should be noted that $q(x = 1|f) = \mu_{x|f}$ in the case of these PoP forecasts.

Examination of the lower part of Figure 2.5 reveals that both types of forecasts are quite reliable; that is, the conditional rela-

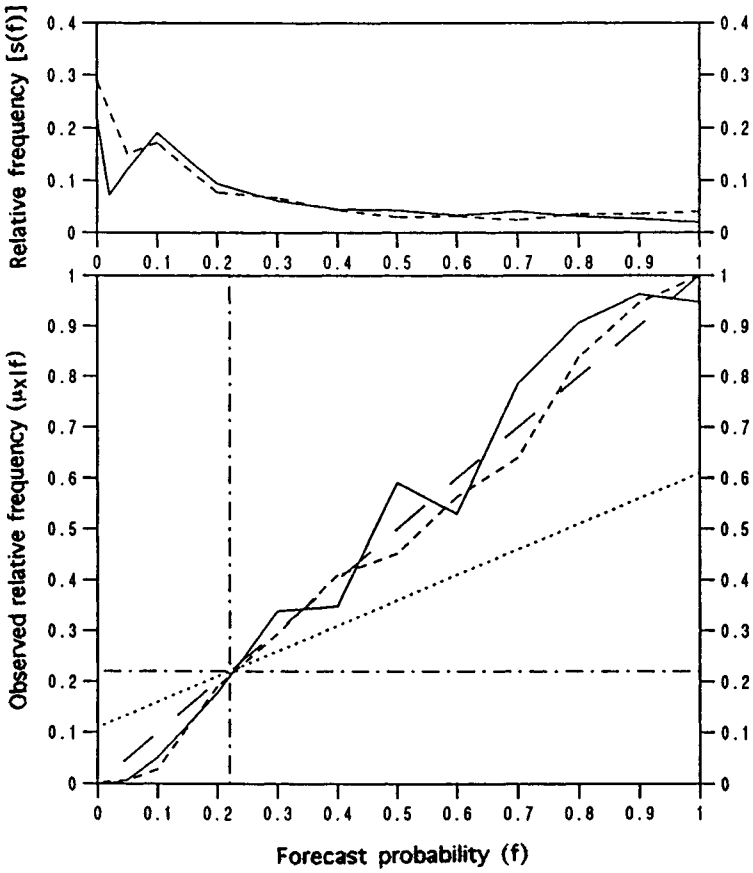

Figure 2.5. Reliability diagram (bottom) and sharpness diagram (top) for 12- to 24-hour objective (solid line) and subjective (dashed line) PoP forecasts in the cool season for St. Louis, Missouri. (From Murphy and Winkler, 1992)

tive frequencies $q(x = 1|f)$ defining the empirical reliability curve correspond quite closely to the forecast probabilities f over the entire range of forecasts. The upper part of Figure 2.5 depicts the marginal distributions $s(f)$, which describe the sharpness (or refinement) of the forecasts. Sharp (refined) forecasts are characterized by u-shaped distributions, with probabilities near zero and one used relatively frequently and intermediate probabilities used relatively infrequently. Although these PoP forecasts use probabilities less than or equal to 0.2 quite frequently, it is evident that neither the objective nor subjective forecasts are very sharp.

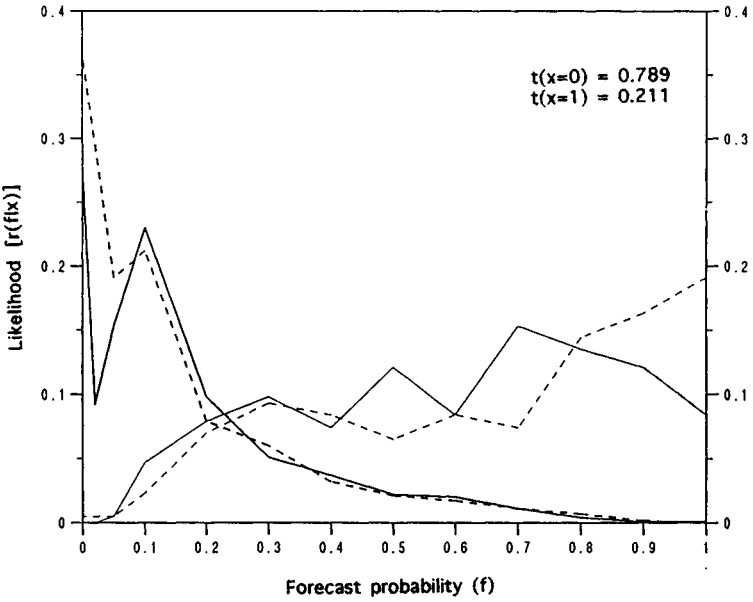

Figure 2.6. Discrimination diagram for 12- to 24-hour objective (solid line) and subjective (dashed line) PoP forecasts in the cool season for St. Louis, Missouri. (From Murphy and Winkler, 1992)

Insight into other aspects of quality can also be obtained from the diagrams in Figure 2.5. The horizontal line $\mu_{x|f} = \mu_x$ represents no resolution (see Section 2.5), and the line intermediate between this horizontal line and the 45° line (which represents perfect reliability) represents no skill (Hsu and Murphy, 1986). This no-skill line is defined in terms of a skill score based on the mean square error (see equation 2.7), in which the standard of reference is a constant forecast of the sample climatological probability (see Section 3.2). Subsamples of forecasts possess positive skill if the corresponding points $(\mu_{x|f}, f)$ lie to the right (left) of the vertical line $f = \mu_x$ and above (below) the no-skill line. Comparison of the empirical reliability curves with the horizontal line $\mu_{x|f} = \mu_x$, keeping in mind the distribution $s(f)$, indicates that both types of forecasts possess substantial resolution. Moreover, most if not all of the points defining these curves lie in regions of positive (subsample) skill, thereby implying that the overall skill of the forecasts should be strongly positive (see Section 3.2).

Table 2.2. Summary measures of marginal distributions $s(f)$ and $t(x)$ and joint distribution $p(f, x)$ for objective (f_o) and subjective (f_s) PoP forecasts in the cool season for St. Louis, Missouri

Lead time (h)	Type of forecast	Means		Variances		Correlation coefficient	Sample size
		μ_f	μ_x	σ_f^2	σ_x^2	ρ_{fx}	n
12–24	f_o	0.220	0.211	0.074	0.166	0.722	1,021
12–24	f_s	0.230	0.211	0.089	0.166	0.755	1,021
24–36	f_o	0.234	0.222	0.067	0.173	0.643	1,019
24–36	f_s	0.227	0.222	0.075	0.173	0.615	1,019
36–48	f_o	0.208	0.212	0.047	0.167	0.565	1,007
36–48	f_s	0.209	0.212	0.054	0.167	0.543	1,007

Source: Adapted from Murphy and Winkler, 1992.

Conditional distributions $r(f|x)$ and marginal distributions $t(x)$ for the PoP forecasts are depicted in Figure 2.6. This figure contains the likelihoods $r(f|x = 0)$ and $r(f|x = 1)$. Although these distributions overlap, substantial discrimination exists for both types of forecasts. That is, high probabilities are used much more frequently than low probabilities when precipitation occurs ($x = 1$), and low probabilities are used much more frequently than high probabilities when precipitation does not occur ($x = 0$). Since PoP forecasts relate to a binary event, the marginal distribution of the observations consists simply of $t(x = 1)$ and $t(x = 0)$, where $t(x = 1) + t(x = 0) = 1$.

Summary measures of the joint and marginal distributions for the PoP forecasts and observations are included in Table 2.2. These measures indicate a slight tendency toward overforecasting for both types of forecasts (i.e., $\mu_f > \mu_x$) for the 12- to 24-hour and 24- to 36-hour lead times. Moreover, the variances of the forecasts are considerably smaller than the variances of the observations for all three lead times (in this regard, recall that the forecasts are probabilities, whereas the observations are binary variables). The correlation coefficient, a summary measure of $p(f, x)$, indicates that, as expected, the association between forecasts and observations tends to decrease as lead time increases.

Table 2.3. Summary measures of conditional distributions of forecasts given observations, $r(f|x)$, for objective (f_o) and subjective (f_s) PoP forecasts in the cool season for St. Louis, Missouri

Lead time (h)	Type of forecast	Means		Variances		Sample sizes	
		$\mu_{f\|x=0}$	$\mu_{f\|x=1}$	$\sigma^2_{f\|x=0}$	$\sigma^2_{f\|x=1}$	$n_{x=0}$	$n_{x=1}$
12–24	f_o	0.118	0.598	0.026	0.070	806	215
12–24	f_s	0.114	0.665	0.027	0.080	806	215
24–36	f_o	0.145	0.547	0.031	0.068	793	226
24–36	f_s	0.137	0.542	0.036	0.083	793	226
36–48	f_o	0.144	0.445	0.026	0.055	794	213
36–48	f_s	0.144	0.453	0.032	0.061	794	213

Source: Adapted from Murphy and Winkler, 1992.

Table 2.3 contains summary measures of the conditional distributions of forecasts given observations, $r(f|x)$. The conditional means, $\mu_{f|x=0}$ and $\mu_{f|x=1}$, for both types of PoP forecasts generally increase and decrease, respectively, as lead time increases. Thus, the difference between these means, which represents an overall measure of discrimination, decreases as lead increases. The conditional variances indicate that the variability of the forecasts when precipitation occurs ($x = 1$) is considerably greater than the variability of the forecasts when precipitation does not occur ($x = 0$). Since the means associated with the conditional distributions of observations given forecasts, $q(x|f)$, together with the forecasts themselves, define the points on the empirical reliability curves (see Figure 2.5), these conditional means are not presented here in tabular form.

3.2. Some basic measures of aspects of quality

In this section we define basic quantitative measures of the various aspects of forecast quality. These performance measures include the mean error, the mean square error, and a skill score based on the mean square error, as well as terms in decompositions of the mean square error and skill score. To illustrate the interpretation of these measures, they are then applied to the verification

data samples — containing Tmax and PoP forecasts — considered previously.

The mean error (ME) of the forecasts in the verification data sample is defined here as the difference between the mean forecast μ_f and the mean observation μ_x. That is,

$$\mathrm{ME}(f, x) = \mu_f - \mu_x. \tag{2.5}$$

The ME is a measure of (unconditional or systematic) bias, with positive (negative) values of ME being indicative of overforecasting (underforecasting). When ME = 0, the forecasts are (unconditionally) unbiased. In some contexts, overall bias is measured in terms of the ratio of — rather than the difference between — these means (in this case, $\mu_f/\mu_x = 1$ represents unbiased forecasts).

The mean square error (MSE) of the forecasts in the verification data sample can be defined as follows:

$$\mathrm{MSE}(f, x) = \sum_f \sum_x p(f, x)(f - x)^2. \tag{2.6}$$

The MSE is a measure of accuracy (see Section 2.6), and its values are nonnegative [MSE $\geq$ 0, with equality only when $p(f, x) = 0$ for all $f \neq x$]. Smaller values of the MSE are indicative of greater accuracy. As expressed in equation (2.6), the MSE makes no distinction between nonprobabilistic and probabilistic forecasts. In the meteorological literature, it is common practice to refer to the mean square error of probabilistic forecasts as the Brier score (BS) (Brier, 1950). We will make use of both the MSE and BS notations here.

Skill scores (SSs) are usually defined as the fractional (or percentage) improvement in the accuracy of the forecasts of interest over the accuracy of forecasts based on a naive forecasting method such as climatology or persistence (e.g., Brier and Allen, 1951; Murphy and Daan, 1985). Thus, if the MSE is taken as the measure of accuracy, then the $\mathrm{SS}_{\mathrm{MSE}}$ can be defined as follows:

$$\mathrm{SS}_{\mathrm{MSE}}(f, r, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(r, x)], \tag{2.7}$$

where $\mathrm{MSE}(r, x)$ is the MSE for forecasts r based on the naive standard of reference R [note that $\mathrm{MSE}(f, x) = 0$ for perfect forecasts]. Values of the $\mathrm{SS}_{\mathrm{MSE}}$ are positive if $\mathrm{MSE}(f, x) < \mathrm{MSE}(r, x)$, zero if $\mathrm{MSE}(f, x) = \mathrm{MSE}(r, x)$, and negative if $\mathrm{MSE}(f, x) > \mathrm{MSE}(r, x)$.

Since $\mathrm{MSE}(f, x) = 0$ when $p(f, x) = 0$ for all $f \neq x$, the value of the $\mathrm{SS}_{\mathrm{MSE}}$ for perfect forecasts is unity.

Climatology and persistence are the standards of reference most often used when skill scores are defined. In the case of climatology, the mean observation (or probability of the event) based on the verification data sample itself is a particularly convenient choice as a standard of reference. In those situations in which the verification data base is relatively large, this sample mean or probability (denoted by μ_x) should be approximately equal to the long-term historical climatological mean or probability. (It should be noted, however, that the use of sample climatology as a standard of reference fails to give the forecasting system or forecaster credit for recognizing differences between sample and long-term climatological probabilities.) The MSE of forecasts based solely on this mean or probability is $\mathrm{MSE}(\mu_x, x) = \sigma_x^2$, in which case equation (2.7) becomes

$$\mathrm{SS}_{\mathrm{MSE}}(f, \mu_x, x) = 1 - [\mathrm{MSE}(f, x)/\sigma_x^2]. \tag{2.8}$$

Under this assumption, skill is positive (negative) when the MSE of the forecasts is less (greater) than the variance of the observations.

The values of ME (bias) in equation (2.5), MSE (accuracy) in equation (2.6), and $\mathrm{SS}_{\mathrm{MSE}}$ (skill) in equation (2.8) for the PoP forecasts for St. Louis, Missouri, are presented in Table 2.4. Overall, the forecasts exhibit relatively little bias. The accuracy and skill of the forecasts are greater in the cool season than in the warm season. Both forecast accuracy and forecast skill decrease as lead time increases. Comments related to the relative performance of the objective and subjective forecasts are reserved for Section 4.2.

The MSE can be decomposed in several different ways, with the terms in these decompositions representing quantitative measures of various aspects of quality. A basic decomposition of the MSE is

$$\mathrm{MSE}(f, x) = (\mu_f - \mu_x)^2 + \sigma_f^2 + \sigma_x^2 - 2\sigma_f \sigma_x \rho_{fx} \tag{2.9}$$

(Murphy, 1988). This decomposition is easily motivated by reference to the familiar statistical expression for the variance of a difference; that is, $\mathrm{Var}(f - x) = \sigma_{f-x}^2 = E[(f - x)^2] - [E(f - x)]^2$. In equation (2.9), a measure of accuracy (MSE) has been decomposed into a measure of unconditional bias $[(\mu_f - \mu_x)^2]$, a measure of sharpness (or variability) (σ_f^2), a measure of uncertainty (σ_x^2),

Table 2.4. Overall measures of bias (ME), accuracy (BS), and skill (SS_{BS}) for objective (f_o) and subjective (f_s) PoP forecasts for St. Louis, Missouri

Lead time (h)	Type of forecast	Sample size n	Mean error (ME)	Brier score (BS)	Skill score (SS_{BS})
(a) Cool season					
12–24	f_o	1,021	0.009	0.080	0.517
12–24	f_s	1,021	0.020	0.072	0.567
24–36	f_o	1,019	0.013	0.101	0.412
24–36	f_s	1,019	0.005	0.108	0.376
36–48	f_o	1,007	–0.004	0.114	0.318
36–48	f_s	1,007	–0.002	0.118	0.295
(b) Warm season					
12–24	f_o	993	0.020	0.110	0.321
12–24	f_s	993	0.021	0.103	0.365
24–36	f_o	990	–0.014	0.133	0.228
24–36	f_s	990	–0.006	0.130	0.240
36–48	f_o	980	–0.017	0.131	0.196
36–48	f_s	980	–0.013	0.129	0.210

Source: Adapted from Murphy and Winkler, 1992.

and a measure of association ($2\sigma_f\sigma_x\rho_{fx} = 2\sigma_{fx}$, where σ_{fx} denotes the covariance of the forecasts and observations). The last three terms on the right-hand side of equation (2.9), taken together, constitute the variance of the forecast errors (i.e., σ^2_{f-x}). This decomposition is applicable to all verification data samples, regardless of the nature (or treatment) of the underlying variable and the format of the forecasts.

The terms in this basic decomposition for the Tmax forecasts are presented in Table 2.5. Some bias is exhibited by the objective forecasts (1.8–2.4°F), whereas the subjective forecasts are almost completely unbiased. The variability of both types of forecasts is less than the variability of the observations. Moreover, the variability of the objective and subjective forecasts — and the

Table 2.5. Decomposition of MSE(f, x) for objective (f_o) and subjective (f_s) maximum temperature forecasts in the winter season for Minneapolis, Minnesota

Lead time (h)	Type of forecast	Sample size n	MSE	$(\mu_f - \mu_x)^2$	σ_f^2	σ_x^2	$2\sigma_f\sigma_x\rho_{fx}$
24	f_o	417	24.9	3.2	148.6	174.9	302.0
24	f_s	417	18.0	0.0	154.0	174.9	310.8
36	f_o	405	34.4	5.8	149.3	184.1	304.8
36	f_s	405	26.9	0.1	154.4	184.1	311.6
48	f_o	416	33.9	4.4	143.8	177.8	292.0
48	f_s	416	28.4	0.0	137.9	177.8	287.4
60	f_o	397	49.6	5.3	129.8	182.7	268.2
60	f_s	397	40.5	0.0	129.6	182.7	271.8

Source: Adapted from Murphy, Brown, and Chen, 1989.

covariability between the respective forecasts and the observations — generally decreases as lead time increases.

Other decompositions of the MSE can be formulated by conditioning on either the forecast f or the observation x. These decompositions are related to the CR and LBR factorizations of the joint distribution $p(f, x)$ (see Section 2.3). Conditioning on the forecast f leads to the following CR decomposition of the MSE:

$$\text{MSE}_{\text{CR}}(f, x) = \sigma_x^2 + E_f(\mu_{x|f} - f)^2 - E_f(\mu_{x|f} - \mu_x)^2, \qquad (2.10)$$

where E_f denotes an expectation with respect to the distribution of forecasts [i.e., a weighted average — using weights $s(f)$ — over all forecasts]. The expression in equation (2.10) represents a decomposition of a measure of accuracy (MSE) into a measure of uncertainty (σ_x^2), a measure of reliability (or type 1 conditional bias) $[E_f(\mu_{x|f} - f)^2]$, and a measure of resolution $[E_f(\mu_{x|f} - \mu_x)^2]$.

The measure of reliability in equation (2.10) is simply the weighted squared deviation of the points defining the empirical reliability curve from the 45° line representing perfect reliability (see Figure 2.5), where the weights are the probabilities that constitute the marginal distribution of forecasts, $s(f)$. This nonnegative term

Table 2.6. Decomposition of $\mathrm{MSE}_{\mathrm{CR}}(f, x)$ related to the calibration-refinement factorization of $p(f, x)$ for objective (f_o) and subjective (f_s) PoP forecasts in the cool season for St. Louis, Missouri

Lead time (h)	Type of forecast	Sample size n	$\mathrm{MSE}_{\mathrm{CR}}$	σ_x^2	$E_f(\mu_{x\|f} - f)^2$	$E_f(\mu_{x\|f} - \mu_x)^2$
12–24	f_o	1,021	0.080	0.166	0.002	0.088
12–24	f_s	1,021	0.072	0.166	0.002	0.096
24–36	f_o	1,019	0.101	0.173	0.002	0.073
24–36	f_s	1,019	0.108	0.173	0.002	0.067
36–48	f_o	1,007	0.114	0.167	0.002	0.055
36–48	f_s	1,007	0.118	0.167	0.003	0.052

Source: Adapted from Murphy and Winkler, 1992.

vanishes only for completely reliable forecasts (i.e., $\mu_{x|f} = f$ for all f). The measure of resolution in equation (2.10) is simply the weighted squared deviation of the points defining the same empirical (reliability) curve from the horizontal line representing no resolution (i.e., $\mu_{x|f} = \mu_x$ for all f), where the weights are once again the components of the marginal distribution $s(f)$. Since it is preceded by a negative sign, larger values of this nonnegative term are indicative of greater resolution.

The results of applying this decomposition of the MSE (or BS) to the PoP forecasts are summarized in Table 2.6. It is evident that the lack of complete reliability, noted in Section 3.1, contributes very little to the MSE for either the objective or subjective forecasts. The resolution term is substantially larger than the reliability term for both types of forecasts, and it decreases as lead time increases.

Conditioning on the observation x leads to the following LBR decomposition of the MSE:

$$\mathrm{MSE}_{\mathrm{LBR}}(f, x) = \sigma_f^2 + E_x(\mu_{f|x} - x)^2 - E_x(\mu_{f|x} - \mu_f)^2, \qquad (2.11)$$

where E_x denotes an expectation with respect to the distribution of observations. The expression in equation (2.11) represents a decomposition of a measure of accuracy (MSE) into a measure of

Table 2.7. Decomposition of $\mathrm{MSE}_{\mathrm{LBR}}(f,x)$ related to the likelihood-base rate factorization of $p(f,x)$ for objective (f_o) and subjective (f_s) PoP forecasts in the cool season for St. Louis, Missouri

Lead time (h)	Type of forecast	Sample size n	$\mathrm{MSE}_{\mathrm{LBR}}$	σ_f^2	$E_x(\mu_{f\|x}-x)^2$	$E_x(\mu_{f\|x}-\mu_f)^2$
12–24	f_o	1,021	0.080	0.074	0.045	0.038
12–24	f_s	1,021	0.072	0.089	0.034	0.050
24–36	f_o	1,019	0.101	0.067	0.062	0.028
24–36	f_s	1,019	0.108	0.075	0.061	0.028
36–48	f_o	1,007	0.114	0.047	0.081	0.015
36–48	f_s	1,007	0.118	0.054	0.080	0.016

Source: Adapted from Murphy and Winkler, 1992.

sharpness (σ_f^2), a measure of type 2 conditional bias $[E_x(\mu_{f|x}-x)^2]$, and a measure of discrimination $[E_x(\mu_{f|x}-\mu_f)^2]$. These latter two terms — and their signs — indicate that it is desirable for the conditional mean forecasts, $\mu_{f|x=1}$ and $\mu_{f|x=0}$, to approach the respective observations $x=1$ and $x=0$ as closely as possible (to decrease type 2 conditional bias) and, at the same time, for these conditional means to differ as much as possible from the overall unconditional mean forecast μ_f (to increase discrimination).

Application of the LBR decomposition of the MSE (or BS) to the PoP forecasts yields the results summarized in Table 2.7. The term that measures type 2 conditional bias (the mean squared difference between $\mu_{f|x}$ and x averaged over all x) and contributes positively to the magnitude of the MSE increases as lead time increases. On the other hand, the term that measures discrimination (the mean squared difference between $\mu_{f|x}$ and μ_f averaged over all x) and contributes negatively to the magnitude of the MSE decreases as lead time increases. Comparisons of the terms in this decomposition for the two types of forecasts are considered in Section 4.2.

A decomposition of the MSE-based skill score in equation (2.7) can be obtained by substituting the basic decomposition of the MSE in equation (2.9) into the expression for $\mathrm{SS}_{\mathrm{MSE}}$ in equation (2.8). After rearranging terms and completing a square, it can be

Table 2.8. Decomposition of $SS_{MSE}(f, \mu_x, x)$ for objective (f_o) and subjective (f_s) maximum temperature forecasts in the winter season for Minneapolis, Minnesota

Lead time (h)	Type of forecast	Sample size n	SS_{MSE}	ρ_{fx}^2	$[\rho_{fx} - (\sigma_f/\sigma_x)]^2$	$[(\mu_f - \mu_x)/\sigma_x]^2$
24	f_o	417	0.858	0.876	0.000	0.018
24	f_s	417	0.897	0.897	0.000	0.000
36	f_o	405	0.813	0.846	0.001	0.031
36	f_s	405	0.854	0.854	0.000	0.000
48	f_o	416	0.809	0.834	0.000	0.025
48	f_s	416	0.840	0.843	0.001	0.000
60	f_o	397	0.728	0.759	0.001	0.029
60	f_s	397	0.778	0.780	0.002	0.000

Source: Adapted from Murphy, Brown, and Chen, 1989.

seen that

$$SS_{MSE}(f, \mu_x, x) = \rho_{fx}^2 - [\rho_{fx} - (\sigma_f/\sigma_x)]^2 - [(\mu_f - \mu_x)/\sigma_x]^2. \quad (2.12)$$

The terms on the right-hand side of equation (2.12) are all nonnegative, and they can be interpreted by reference to a linear regression model in which the forecasts are regressed on the observations (e.g., see Murphy and Winkler, 1992). In this context, the first term is the square of the correlation coefficient ρ_{fx}, a measure of (linear) association between f and x. As noted in Section 2.6, ρ_{fx}^2 represents the fraction of the variability in the observations accounted for (or "explained") by the forecasts.

Forecasts are completely reliable in the context of the regression model only when the regression line possesses zero intercept *and* unit slope (i.e., only when it coincides with the 45° line). Under this condition, $\sigma_f = \rho_{fx}\sigma_x$, and it follows that the second term on the right-hand side of equation (2.12) is a measure of reliability (calibration, conditional bias in the type 1 sense). This term vanishes for completely reliable forecasts, and otherwise acts to reduce the skill score. Since the third term is the squared difference between μ_f and μ_x, scaled by the square of σ_x, it is a measure of

Table 2.9 Decomposition of $SS_{MSE}(f, \mu_x, x)$ for objective (f_o) and subjective (f_s) PoP forecasts in the cool season for St. Louis, Missouri

Lead time (h)	Type of forecast	Sample size n	SS_{MSE}	ρ_{fx}^2	$[\rho_{fx} - (\sigma_f/\sigma_x)]^2$	$[(\mu_f - \mu_x)/\sigma_x]^2$
12–24	f_o	1,021	0.517	0.521	0.003	0.000
12–24	f_s	1,021	0.567	0.570	0.001	0.002
24–36	f_o	1,019	0.412	0.414	0.000	0.001
24–36	f_s	1,019	0.376	0.378	0.002	0.000
36–48	f_o	1,007	0.318	0.319	0.001	0.000
36–48	f_s	1,007	0.295	0.295	0.001	0.000

Source: Adapted from Murphy and Winkler, 1992.

overall (or unconditional) bias. It vanishes for completely unbiased forecasts, and otherwise acts to reduce the skill score.

Examination of the decomposition in equation (2.12) reveals that SS_{MSE} and ρ_{fx}^2 are equal when the forecasts are conditionally unbiased. As noted in Section 2.6, a verification data sample that is conditionally unbiased for all forecasts is also unconditionally unbiased, but not necessarily vice versa. In this sense, ρ_{fx}^2 can be viewed as a measure of potential skill (Murphy and Epstein, 1989).

The results of applying this decomposition of the SS to the Tmax and PoP forecasts are summarized in Tables 2.8 (p. 53) and 2.9, respectively. Since these samples of forecasts are relatively reliable, the terms measuring type 1 conditional bias (or reliability) and unconditional (or systematic) bias are quite small in most cases. The only exception to this general result is the sample of objective Tmax forecasts, which exhibit a small but not insignificant contribution to the SS from the term that measures unconditional bias. The fact that these two terms are quite small in most cases implies that the values of the SS and ρ_{fx}^2 are similar and that actual skill and potential skill differ very little in most cases for these forecasts.

4. Comparative verification: methods and applications

In this section we address the problem of comparative verification of two or more forecasting systems. Since the process of comparing such systems is transitive, it suffices to compare forecasting systems on a pairwise basis. Section 4.1 discusses the use of methods based on the sufficiency relation as a means of screening alternative forecasting systems. The application of the basic set of verification methods and measures introduced in Section 3.2 to the problem of comparative verification is illustrated in Section 4.2.

4.1. Screening forecasting systems: the sufficiency relation

The conditions under which one forecasting system can be unambiguously judged to be better, in terms of both quality and value, than another forecasting system are embodied in the *sufficiency relation.* In the context of matched comparative verification, forecasting system F is sufficient for forecasting system G if and only if a stochastic transformation $h(g|f)$ exists such that

$$\sum_f h(g|f) r_F(f|x) = r_G(g|x) \text{ for all } x \tag{2.13}$$

(e.g., DeGroot and Fienberg, 1982; Ehrendorfer and Murphy, 1988). The function $h(g|f)$ qualifies as a stochastic transformation if $0 \leq h(g|f) \leq 1$ for all f and g and $\sum_g h(g|f) = 1$ for each f. Note that, under the assumption that the marginal distribution $t(x)$ is known, the likelihoods $r_F(f|x)$ and $r_G(g|x)$ in equation (2.13) provide a complete description of the quality of F's and G's forecasts, respectively (see equation 2.2).

Since the sufficiency relation defined by equation (2.13) implies that G's likelihoods can be obtained by an auxiliary randomization of F's likelihoods, the former obviously contain greater uncertainty than the latter. The importance of the sufficiency relation resides in the fact that if it can be shown that F's forecasts are sufficient for G's forecasts, then it follows that F's forecasts are of greater value than G's forecasts to all users regardless of the structure of their individual payoff (or loss) functions.

The sufficiency relation was developed originally by Blackwell (1953) in the context of statistical experiments. It was then applied to information systems by Marschak (1971) and subsequently

introduced into the forecasting literature by DeGroot and Fienberg (1982). The sufficiency relation is a quasi-order (Krzysztofowicz and Long, 1991a), in the sense that the comparison of systems F and G can lead to any of three possible results: (i) F is sufficient for G; (ii) G is sufficient for F; or (iii) F and G are insufficient for each other. Result (iii) implies that no function $h(g|f)$ or $h(f|g)$ can be found that satisfies the conditions for a stochastic transformation. In this case, some users may prefer F's forecasts and other users may prefer G's forecasts. The conditions under which forecasting systems are sufficient and insufficient for each other clearly is an important issue in applications of the sufficiency relation to problems of comparative verification.

As a method of screening alternative forecasting systems, the sufficiency relation can be used to determine whether or not a particular forecasting system is dominated by another forecasting system (Clemen, Murphy, and Winkler, 1995). However, as a quasi-order, the method will not always be able to identify the "better" system. Nevertheless, in the context of comparative verification, this relation leads to coherent methods of comparing alternative forecasting systems, eliminating those systems whose forecasts are dominated by the forecasts of other systems, and thereby reducing the number of alternative systems that must be considered. Various applications of the sufficiency relation as a screening method are described briefly in this section.

One screening method based on the sufficiency relation involves the direct search for auxiliary randomizations that satisfy the conditions for a stochastic transformation. This approach was taken by Ehrendorfer and Murphy (1988, 1992a) in assessing the relative quality and value of primitive weather and climate forecasts. For example, Ehrendorfer and Murphy (1988) investigated the conditions under which one forecasting system is sufficient for another forecasting system in the case of nonprobabilistic forecasts for a dichotomous variable (i.e., $f = 0$ or 1 and $x = 0$ or 1). These conditions are described graphically in the two-dimensional sufficiency diagram depicted in Figure 2.7, in which the conditional probabilities $q_{11} = \Pr(x = 1|f = 1)$ and $q_{10} = \Pr(x = 1|f = 0)$ represent the coordinate axes (equivalent descriptions can be formulated in terms of likelihoods or joint probabilities; see Ehrendorfer and Murphy, 1988).

In Figure 2.7 system F is taken as the reference system, with $q_{11}(F) = 0.571$ and $q_{10}(F) = 0.276$ [the sample climatological probability $t_1 = \Pr(x = 1) = 0.4$, the same for all systems under consideration here]. These conditional probabilities completely determine the geometry of the sufficiency diagram, which is defined by horizontal and vertical lines passing through the point $(q_{11}, q_{10}) = (0.571, 0.276)$ in this framework. Three regions are identified: (i) regions S, which consist of all systems G (e.g., $G1$) for which system F is sufficient; (ii) regions S', which consist of all systems G (e.g., $G3$) that are sufficient for system F; and (iii) regions I, which consist of all systems G (e.g., $G2$) that are insufficient for system F.

The constraints that the conditional probabilities of the respective forecasting systems must satisfy to fall in these regions are summarized in Table 2.10. In essence, for an alternative system G to be sufficient for the reference system F, the interval defined by the former's conditional probabilities $q_{11} = \Pr(x = 1|f = 1)$ and $q_{10} = \Pr(x = 1|f = 0)$ must "straddle" the interval defined by the latter's conditional probabilities. Conversely, if F's interval straddles G's interval, then F's forecasts are sufficient for G's forecasts. If the intervals overlap, then systems F and G are insufficient for each other. This latter condition implies that (i) the quality of F's forecasts is, at the same time, superior and inferior to the quality of G's forecasts in one or more respects and (ii) the user population includes at least some individuals who prefer each system's forecasts. In this case neither F nor G can be said to dominate the other system.

The dashed lines in Figure 2.7 are isopleths of the expected Brier score (EBS) [see Ehrendorfer and Murphy (1988) for a definition of the EBS]. It should be noted that these isopleths generally traverse regions of sufficiency (S or S') and regions of insufficiency (I). Thus, from the relative values of EBS alone it is not possible to determine whether F is sufficient for G (or vice versa) or F and G are insufficient for each other. On the other hand, if $\mathrm{EBS}(F) < (>)\ \mathrm{EBS}(G)$, system G (F) cannot be sufficient for system F (G).

Another screening method based on the sufficiency relation involves the formulation of the *forecast sufficiency characteristic* (FSC) (Krzysztofowicz and Long, 1991a). This method, which is based on a theorem presented by Blackwell and Girshick (1954), is applicable only in situations involving dichotomous variables (i.e.,

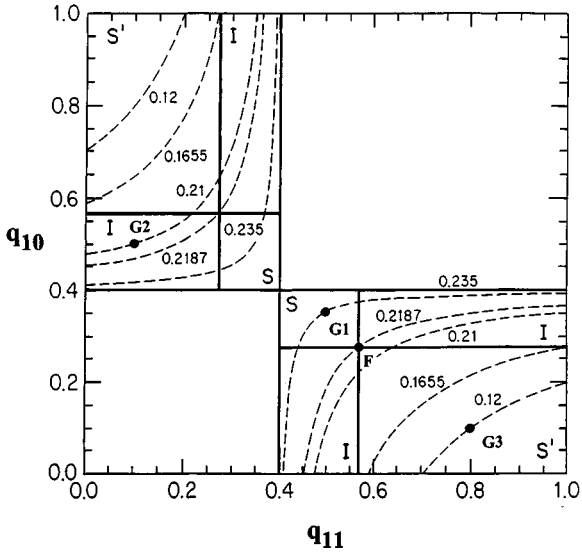

Figure 2.7. Sufficiency diagram for nonprobabilistic forecasts in a dichotomous situation, described in terms of the conditional probabilities $q_{11} = \Pr(x = 1|f = 1)$ and $q_{10} = \Pr(x = 1|f = 0)$. Isopleths represent lines of equal expected half-Brier score. See text for additional details. (From Ehrendorfer and Murphy, 1988)

$x = 0$ or 1). It involves the construction and comparison of FSCs, which are derived from the likelihoods of the respective forecasting systems [i.e., from $r(f|x = 1)$ and $r(f|x = 0)$].

As an example of an application of this screening method, the FSCs for objective and subjective precipitation probability forecasts for Portland, Oregon, are depicted in Figure 2.8 [the FSCs are denoted by $C(t)$ in this figure]. If the FSC of one forecasting system is *superior* to (i.e., lies everywhere to the left and above) the FSC of the other forecasting system, then the former's forecasts are sufficient for the latter's forecasts. In the case of the cool season forecasts (Figure 2.8a), the FSCs clearly do not satisfy the conditions for sufficiency. However, in the case of the warm season forecasts (Figure 2.8b), the FSC for the subjective forecasts is superior to the FSC for the objective forecasts, implying that the former are sufficient for the latter.

Although experience with the use of FSCs to compare forecasting systems is quite limited, this experience suggests that the FSCs

Table 2.10. Constraints on conditional probabilities of reference forecasting system F, $q_{11}(F) = 0.571$ and $q_{10}(F) = 0.276$, and alternative forecasting systems G, $q_{11}(G)$ and $q_{10}(G)$, associated with the regions S, S', and I in Figure 2.7 [with $t_1 = \Pr(x = 1) = 0.4$]

Region	Constraints	Example
S	$q_{10}(F) \leq q_{10}(G) < t_1 < q_{11}(G) \leq q_{11}(F)$	$G1: \ q_{11} = 0.50, \ q_{10} = 0.35$
S'	$q_{10}(G) \leq q_{10}(F) < t_1 < q_{11}(F) \leq q_{11}(G)$	$G3: \ q_{11} = 0.80, \ q_{10} = 0.10$
I	$q_{10}(F) \leq q_{10}(G) < t_1 < q_{11}(F) \leq q_{11}(G)$ or $q_{10}(G) \leq q_{10}(F) < t_1 < q_{11}(G) \leq q_{11}(F)$	$G2: \ q_{11} = 0.10, \ q_{10} = 0.50$

Note: These constraints define the regions contained in the lower right-hand portion of the unit square in Figure 2.7. The regions in the upper left-hand portion of the unit square represent the situation in which the values of q_{11} and q_{10} are interchanged.

of highly competitive forecasting systems seldom will satisfy the conditions for sufficiency. For example, in 24 different situations (i.e., combinations of location, season, and lead time) considered by Murphy and Ye (1990), sufficiency could be demonstrated in only two situations. Similar results are reported by Krzysztofowicz and Long (1991a). In a companion paper, Krzysztofowicz and Long (1991b) used beta distributions to model the likelihoods associated with the objective and subjective probability forecasts and then compared the respective FSCs derived from the models. This approach served to suppress the sampling variability present in the empirical data sets and, as a result, yielded less ambiguous results. That is, the conditions for sufficiency were met in a substantially larger fraction of the situations. For a discussion of other models of distributions of forecasts and/or observations, see Section 2.5.

The utility of the sufficiency relation as a means of screening alternative forecasting systems is limited because it imposes strict conditions on the relationship between the likelihood functions [i.e., the conditional distributions $r(f|x)$] associated with the respective forecasting systems. The strictness of these conditions arises from the fact that no assumptions are made concerning the

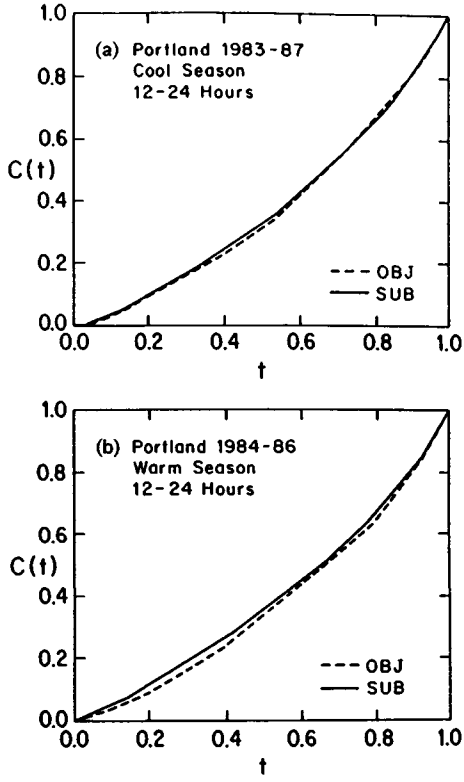

Figure 2.8. FSCs for 12- to 24-hour objective (dashed curve) and subjective (solid curve) PoP forecasts for Portland, Oregon: (a) cool season; (b) warm season. (From Murphy and Ye, 1990)

structure of the payoff or loss functions of the users of the forecasts. Moreover, straightforward application of this relation is difficult in comparative verification problems of moderate or high dimensionality. It may be possible to develop less restrictive — and more easily applicable — versions of the sufficiency relation in situations in which some knowledge is available concerning users' payoff functions. These "tailored" sufficiency relations presumably would be more successful in identifying dominant or dominated forecasting systems in particular situations.

4.2. Distributions-oriented methods and their application

Application of the screening methods described in Section 4.1 may reveal that a forecasting system of potential interest is dominated by another system. In general, however, this screening process will not lead to the identification of a single "best" forecasting system. In such cases, it is necessary to compare the various aspects of forecast quality using methods similar to those described in Section 3.2. To illustrate this approach to comparative verification we will make use of the objective and subjective weather forecasts — and associated results — presented in Section 3. The Tmax forecasts are considered first, and then the PoP forecasts. To conserve space, the objective and subjective forecasts are referred to simply as f_o and f_s, respectively.

Tmax forecasts. From the box plots for the Tmax forecasts and the corresponding observations in Figure 2.4, it is evident that f_o possesses a modest positive (i.e., overforecasting) bias, whereas f_s is relatively unbiased. The magnitude of the respective unconditional biases can be inferred from Table 2.5 — the bias of f_o ranges from 1.8°F to 2.4°F, whereas the bias of f_s is very close to zero. Despite the bias for f_o, the measures of bias in the decomposition of SS_{MSE} detract very little from the overall skill of both types of forecasts (see Table 2.8).

The bivariate histograms in Figure 2.1 provide qualitative insight into the relative association of f_o and f_s for the Tmax forecasts. From a quantitative point of view, the line diagram in Figure 2.2 indicates that ρ_{fx} is slightly higher for f_s than for f_o at all lead times. This difference is consistent with the numerical values of $2\sigma_f\sigma_x\rho_{fx}$ and ρ_{fx}^2 in Tables 2.5 and 2.8, respectively.

The accuracy of the Tmax forecasts, as measured by the MSE, is reported in Table 2.5. According to this measure, f_s is noticeably more accurate than f_o at all lead times. The skill scores for these forecasts in Table 2.8 reveal that f_s is more skillful than f_o across all lead times. In percentage terms, this difference in skill ranges from 3.1% to 5.0%.

Under the assumption that the conditional median temperatures correspond closely to the conditional mean temperatures, the conditional quantile diagrams in Figure 2.3 characterize the reliability of the 24-hour Tmax forecasts. Comparison of these diagrams indicates that the subjective forecasts (Figure 2.3b) are

more reliable than the objective forecasts (Figure 2.3a). Despite this apparent difference in reliability, the magnitude of the term in the decomposition of SS_{MSE} that measures reliability is the same for both types of forecasts and detracts very little from the skill score at all lead times (see Table 2.8).

The relative variability of the objective and subjective Tmax forecasts is indicated in Table 2.5. These results reveal that f_s is somewhat more variable than f_o for the 24-hour and 36-hour forecasts, whereas the opposite is true for the 48-hour forecasts.

Information regarding discrimination in the case of the Tmax forecasts has been omitted to conserve space. However, results presented by Murphy et al. (1989) indicate that f_s exhibits greater discrimination than f_o at the 24-hour lead time, but that very little difference in discrimination between the two types of forecasts exists at longer lead times.

PoP forecasts. Comparison of the values of the ME for the objective and subjective PoP forecasts in Table 2.4 (see also Table 2.2) reveals that neither f_o nor f_s possesses any substantial bias. The impact of the unconditional biases that do exist on the skill score SS_{MSE} is nil (see Table 2.9). With regard to the association between the PoP forecasts and observations, the values of ρ_{fx} in Table 2.2 indicate that association is somewhat greater for f_s than for f_o in the case of the 12- to 24-hour forecasts and vice versa in the cases of the 24- to 36-hour and 36- to 48-hour forecasts. Values of ρ_{fx}^2 for both types of forecasts are included in Table 2.9.

The values of the MSE (or BS) for the PoP forecasts in Table 2.4 (see also Tables 2.6 and 2.7) indicate that f_s is slightly more accurate than f_o for the 24-hour forecasts in both seasons. However, at longer lead times, either little if any difference in accuracy exists or f_o slightly exceeds f_s in accuracy (e.g., the 24- to 36-hour forecasts in the cool season). With regard to skill, the skill scores in Table 2.4 (see also Table 2.9) indicate that f_s is noticeably more skillful than f_o for the 12- to 24-hour PoP forecasts in both seasons. Differences in skill are much smaller at longer lead times, with f_o (f_s) possessing slightly higher skill in the cool (warm) season.

Comparison of the empirical reliability curves for the objective and subjective PoP forecasts in Figure 2.5 suggests that the subjective forecasts are more reliable than the objective forecasts. However, the quantitative measure of reliability in the CR decomposition of the MSE (see Table 2.6) reveals little or no difference

in overall reliability between f_o and f_s at all lead times. In the case of resolution, the quantitative measure in Table 2.6 indicates that f_s (f_o) is slightly more resolved than f_o (f_s) for the 12- to 24-hour (24- to 36-hour) forecasts. Comparison of the values of σ_f^2 in Table 2.2 (see also Table 2.7) reveals that f_s is sharper than f_o at all lead times, with the difference in sharpness being largest for the 12- to 24-hour forecasts.

The discrimination diagrams for the PoP forecasts in Figure 2.6 suggest that f_s possesses greater discrimination than f_o for the 12- to 24-hour forecasts. This inference is supported by a comparison of the difference in the means of the conditional distributions, $\mu_{f|x=1} - \mu_{f|x=0}$, for the two types of forecasts (see Table 2.3). The respective differences are 0.551 for f_s and 0.480 for f_o. It is interesting to note that the difference in magnitude between these two differences in means is almost entirely due to the fact that $\mu_{f|x=1}$ is much larger for f_s than for f_o (in particular, the difference between the respective values of $\mu_{f|x=0}$ is quite small). At longer lead times the respective differences in these conditional means are similar in magnitude, indicating that the levels of discrimination for the two types of forecasts are about the same. These results are supported by the values of the quantitative measure of discrimination in Table 2.7.

Quality — overall assessment of f_o vis-à-vis f_s. With regard to an overall assessment of the relative quality of the objective and subjective forecasts, it appears that 12- to 24-hour subjective forecasts are of higher quality — in terms of most if not all of its various aspects — than the 12- to 24-hour objective forecasts (this statement applies to both the Tmax and PoP forecasts). However, at longer lead times differences in the various aspects of quality for the two types of forecasts are generally much smaller, and in some cases appear to favor the objective forecasts over the subjective forecasts. In this regard, it should be noted that the results of comparative verification are usually not definitive, in the sense that one forecasting system can be judged unambiguously to be better than another forecasting system. The fact that one system is better than another system on "almost all" aspects of quality does not guarantee that such a strong conclusion is warranted. Only comparisons based on the sufficiency relation can yield unambiguous results regarding the relative quality (and value) of alternative forecasting systems.

5. Other methods and measures

The DO approach and its three classes of methods can be applied to most if not all verification problems. In particular applications, however, it may be desirable (or necessary) to tailor the body of DO methodology to take into account such factors as the nature of the underlying variable and the type of forecasts. For example, the DO methods introduced in Sections 3 and 4 were tailored to nonprobabilistic forecasts of a continuous weather variable in the case of Tmax forecasts and to probabilistic forecasts of a dichotomous weather variable in the case of PoP forecasts. In this section we briefly describe ways in which DO methods can be tailored to verification problems involving nonprobabilistic forecasts of discrete weather variables and probabilistic forecasts of discrete but polychotomous weather variables.

In the case of nonprobabilistic forecasts of discrete variables (e.g., cloud amount, weather types), the basic joint, conditional, and marginal distributions of forecasts and/or observations can be conveniently summarized in the form of contingency tables. Summary measures of the basic distributions — as well as measures of various aspects of forecast quality — can be calculated directly from the joint, conditional, and marginal probabilities that constitute the elements of these tables. In choosing among alternative measures of some aspects of quality, it may be necessary to consider the nature of the underlying weather variable. If the variable is ordinal (i.e., consists of categories with a natural order, such as categories of cloud amount), verification measures that take into account the "distance" between the forecast and observed categories may be appropriate. On the other hand, when the variable is nominal (i.e., consists of events without a natural order, such as weather types), the concept of distance is not meaningful and verification measures need not take this factor into account. A recent discussion of verification methods for nonprobabilistic forecasts of discrete variables can be found in Wilks (1995, pp. 238–250).

Verification problems involving probabilistic forecasts of polychotomous variables are problems of relatively high dimensionality. In the case of nominal variables, such problems can be simplified — with little or no loss of information — by treating the forecasts and observations associated with each event separately. If this approach is followed, then the original polychotomous problem

would be replaced by n^x dichotomous problems (i.e., a dichotomous problem associated with each event defining the underlying polychotomous variable), and the body of methods used to evaluate PoP forecasts in Sections 3 and 4 could be applied to the verification data sample associated with each dichotomous problem. A similar approach could be followed in the case of ordinal variables (in this case, the polychotomous problem would be replaced by $n^x - 1$ dichotomous problems, one associated with each threshold defining the categories of the underlying polychotomous variable), but it would generally lead to the loss of information concerning some aspects of forecast quality. Moreover, in the case of ordinal polychotomous variables, the appropriate squared-error measure of overall accuracy is the ranked probability score (RPS) (Epstein, 1969; Murphy, 1971). In addition to taking distance into account, the RPS — and skill scores based on the RPS — can be decomposed into measures of other aspects of quality in a manner analogous to the BS and the SS_{BS} in Section 3.2. The full range of DO methods have yet to be applied to probabilistic forecasts in (unmodified) polychotomous verification problems. For further discussion of verification problems involving probabilistic forecasts, see Murphy and Daan (1985) and Wilks (1995).

Verification methods and measures almost always involve various explicit and/or implicit assumptions, and the body of DO methodology described in this chapter is no exception. For example, the basic measures of accuracy and skill introduced in Section 3.2 are defined in terms of the square of the overall degree of correspondence between forecasts and observations. In effect, it has been assumed here that the importance of differences between forecasts and observations is proportional to the square of the magnitude of these differences. The mean square error criterion seems reasonable in many situations, and it facilitates the processs of decomposing the basic measures into measures of other aspects of forecast quality. However, verification methods and measures of overall accuracy or skill based on other criteria have been proposed. For example, Mielke (1991) advocates the use of a linear measure of correspondence between forecasts and observations; a measure in effect analogous to the mean absolute error. Another approach has been proposed by Potts et al. (1996); it involves measuring the correspondence between forecasts and observations

in probability space rather than in the space of values of the underlying variable. In this approach, forecast errors are defined in terms of (linear) differences between the respective cumulative probabilities of the forecasts and observations according to the climatological distribution of the variable, thereby assigning larger penalties to errors in regions of probability space in which forecasts/observations are more likely to occur (see Wilks, 1995, pp. 257–258).

In recent years, verificaton methods based on concepts derived from signal detection theory (SDT) (e.g., Swets, 1988) have been used to assess and compare forecasting performance. The SDT approach to verification problems involving weather forecasts was initially exploited by Mason (1982), who described SDT-based methods in some detail and then reported the results of applying these methods to samples of probabilistic forecasts. The basic tool in the SDT approach is the receiver operating characteristic (ROC) — a curve representing the relationship between $r(f|x = 1)$ (the "probability of detection") versus $r(f|x = 0)$ (the "false alarm ratio") for all possible values of f. As a result, SDT methods are similar in some respects to verification methods associated with the likelihood-base rate factorization of the joint distribution $p(f, x)$ (see Section 2.3). The ROC can be calculated using empirical estimates of the conditional probabilities $r(f|x = 1)$ and $r(f|x = 0)$, or it can be determined using Gaussian models of these conditional probabilities (in this latter case, the SDT approach provides an example of a model-based approach to forecast verification; see Section 2.5). Various measures of overall performance can be derived from the estimated or modeled ROCs. Although SDT methods are applicable to all types of forecasts (including forecasts expressed in qualitative terms), these methods are limited to situations involving dichotomous variables. For a recent discussion and application of SDT methods to forecast verification, see Harvey et al. (1992).

As described in this chapter (see Section 2.1), forecast verification is the process and practice of evaluating weather forecasts at specific points (e.g., geographical locations, grid points). In effect, this process/practice ignores the relationship that may exist between forecasts (and observations) at different points. Thus, the problem of *model verification*, which involves the evaluation of forecasts produced by numerical models in the form of spatial arrays (or two-dimensional fields), presents a somewhat different

challenge. This problem generally involves nonprobabilistic forecasts, and it can be approached in at least two quite different ways. Traditionally, model verification has been performed by comparing the forecasts and the corresponding observations (or analyzed values) on a point-by-point basis. If this approach is followed, then the DO methods described in Sections 3 and 4 can be used to assess or compare various aspects of forecast quality. In practice, model verification has usually consisted of calculating one or two overall measures of performance such as the mean square error or the (anomaly) correlation coefficient (e.g., see Wilks, 1995, pp. 272–281).

In the point-by-point approach, the correspondence between coherent features — for example, the phase or amplitude of waves — in the respective spatial arrays is not explicitly considered. Model verification in terms of coherent features — or, more generally, spatial patterns encountered in the two-dimensional fields — is not well-developed and is seldom attempted. Some insight into the extent to which differences between forecasts and observations (or analyzed values), calculated on a point-by-point basis, are due to differences in the phase or amplitude of features can be obtained by appealing to the decomposition of $\mathrm{SS}_{\mathrm{MSE}}$ in equation (2.12). In this expression, the term ρ_{fx}^2 can be interpreted as the degree of association in phase and the term $[\rho_{fx} - (s_f/s_x)]^2$ can be interpreted as the degree of correspondence in amplitude (in applications of this decomposition to model verification problems, the forecasts and observations are usually expressed as anomalies). Further discussion of this decomposition and its application to model verification problems can be found in Livezey (1995) and Murphy and Epstein (1989). Recent attempts to develop methods of model verification based on spatial patterns or features are reported by Briggs and Levine (1997) and Hoffman et al. (1995).

6. Forecast quality and forecast value

This chapter has focused on methods of assessing forecast quality and its various aspects. However, it is not possible — or necessarily even desirable — to divorce considerations of forecast quality entirely from considerations of forecast value. Even the relatively simple choice of a measure of accuracy — for example, the choice

between the mean absolute error and the mean square error — involves considerations related to forecast use and value (for an expository discussion of the principal types of goodness in forecasting and their interrelationships, see Murphy, 1993). Moreover, the primary focus of this book — in particular, Chapters 3 through 6 — is the value of weather and climate forecasts. For these reasons, some basic features of the relationship between forecast quality and forecast value are discussed briefly here. This discussion may help to connect the treatment of forecast quality in this chapter with the treatments of forecast value in subsequent chapters.

It is useful here to distinguish between two types of situations in which forecast use and value are of interest. In the first type of situation all potential users of the forecasts are of interest and nothing is known or assumed about their decision-making problems, in particular about the structure of their payoff functions. In the second type of situation, one or more specific users are of interest and knowledge is available or assumptions are made about the structure of the payoff functions of these users. To distinguish between these situations, we refer to the former as the *general-user situation* and the latter as the *specific-user situation.* Moreover, the primary concern in this discussion is the effect of changes in forecast quality (or its aspects) on forecast value, not the actual numerical values of these quantities.

In considering the nature of the quality/value relationship in the general-user situation, it is instructive to compare the general expression for the value of forecasts (VF) with an expression for the mean square error, a measure of accuracy (a specific aspect of quality). Under the assumption that the user's utility function is linear (i.e., utilities and payoffs are linearly related; see Chapters 3 and 4 of this volume), the expression for VF in a single-stage (i.e., static) decision-making problem can be written as follows:

$$\mathrm{VF} = \min_{\alpha} \sum_{x} t(x)\lambda(\alpha, x) - \sum_{f} s(f) \min_{\alpha} \sum_{x} q(x|f)\lambda(\alpha, x), \quad (2.14)$$

where $\lambda(\alpha, x)$ is the loss incurred by the decision maker when action α is taken and (weather or climate) event x occurs. The first term on the right-hand side of equation (2.14) is the expected loss incurred by the user if her decisions are based on the sample climatological probabilities of the events, $t(x)$. This term does not

involve the forecasts and can be considered to be a fixed constant [given $t(x)$] for the purposes of this discussion. The second term on the right-hand side of equation (2.14) is the user's expected loss when her decisions are based on the forecasts. Given a particular forecast f, the decision maker is assumed to choose the action that minimizes expected loss for that forecast, with the conditional distributions $q(x|f)$ containing the probabilities required to compute the relevant expected losses. Overall expected loss, as measured by this term, then involves averaging these minimal expected losses over all possible forecasts. When the expected loss associated with climatological probabilities is taken as the zero point on the value scale, the difference between the two terms in equation (2.14) represents the expected value of the forecasts.

The MSE in equation (2.6) can be written as follows:

$$\mathrm{MSE}(f,x) = \sum_f s(f) \sum_x q(x|f)(f-x)^2, \qquad (2.15)$$

since $\mathrm{p}(f,x) = q(x|f)s(f)$ (see equation 2.1). Comparison of equation (2.15) and the second term on the right-hand side of equation (2.14) reveals definite similarities between these expressions but also two important differences. First, the second term in VF involves a general loss function $\lambda(\alpha,x)$ (in which the forecast f influences the choice of an action α), whereas the MSE involves a specific "loss" defined in terms of the square of the difference between f and x. Second, the term in VF includes a process of minimization associated with the choice of an optimal action (for each f), whereas the MSE includes no such optimization process. These differences imply (*inter alia*) that the relationship between VF and the MSE is inherently nonlinear (see Chapter 6 of this volume).

The fact that VF and MSE are, in general, different functions of the joint distribution of forecasts and observations has other important implications as well. For example, changes in one or more aspects of quality, as reflected by changes in $q(x|f)$ and/or $s(f)$, generally will affect VF and the MSE (or any other one-dimensional measure of an aspect of quality) differently. As a result, the relationship between VF and MSE will be multivalued rather than single-valued. That is, a range of values of VF will exist for a specific value of the MSE, and vice versa. The existence of a multivalued relationship between VF and MSE implies that

decreases (increases) in the MSE can be associated with decreases (increases) in VF. These *accuracy/value reversals* have been investigated by Ehrendorfer and Murphy (1992b) and Murphy and Ehrendorfer (1987), among others. If decisions related to the absolute or relative performance of forecasting systems are based on measures of one or two overall aspects of quality, then they may be subject to such reversals (from the perspective of some users). It is for this reason that we place particular emphasis on the multidimensional structure of verification problems and recommend the use of a suite of verification methods to assess the various aspects of forecast quality.

In specific-user situations, in which the user's payoff function is known or is modeled in terms of a few parameters, the relationship between forecast quality and forecast value can be determined either analytically or numerically. Studies of quality/value relationships in both prototypical and real-world situations have been undertaken in recent years. These studies have confirmed the nonlinear nature of this relationship and have identified other properties (e.g., quality thresholds, convexity) that these relationships appear to possess in many situations of potential interest. Relationships between changes in quality and changes in value in real-world situations are discussed in Chapter 4 of this volume. An in-depth discussion of such relationships in the context of prototypical situations is a primary focus of Chapter 6. These chapters include most if not all of the known references to recent quality/value studies.

7. Conclusion

This chapter has focused on forecast verification, which is characterized here as the process of assessing forecast quality. Forecast quality, in turn, is defined as the totality of the statistical characteristics of the forecasts, the observations, and their relationship embodied in the joint distribution of forecasts and observations. The joint distribution, together with conditional and marginal distributions associated with factorizations of this distribution, constitute the basic elements of a distributions-oriented approach to verification problems. This approach provides insight into the complexity and dimensionality of these problems. Moreover, the various aspects of quality are shown to be directly related

to these underlying joint, conditional, and/or marginal distributions. The chapter also describes a method of screening alternative forecasting systems based on the sufficiency relation, as well as several criteria that can be used to screen alternative verification measures.

A basic set of verification methods applicable to all types of weather variables and forecasts has been described and illustrated. These methods include the basic joint, conditional, and marginal distributions themselves, summary measures of these distributions, and various verification measures and terms in decompositions of these measures. The use of these methods in both absolute verification problems and comparative verification problems has been illustrated through applications to verification data samples involving maximum temperature and precipitation probability forecasts. Although maximum temperature forecasts and precipitation probability forecasts differ with respect to both the nature and treatment of the underlying variable and the format of the forecasts, the verification methods applied to the two data sets introduced in Sections 3 and 4 are essentially equivalent. The way in which this body of methods could be tailored to verification problems involving other types of variables and/or forecasts was outlined in Section 5, and this section also contained a brief introduction to some other verification problems and methods. Section 6 discussed the relationship between forecast quality and forecast value, with the objective of providing some insight into the relationship between this chapter and the remaining chapters in the book.

Forecast verification, as described in this chapter, focuses on the use of the verification process to obtain insight into the basic strengths and weaknesses in forecasting performance. This process is an essential component of any effort to assess forecast quality in its full dimensionality or to compare forecasting systems in a rational manner. Since forecast quality is an important determinant of forecast value, detailed assessments of the various aspects of quality are a desirable adjunct to studies of the absolute and/or relative value of weather and climate forecasts.

Acknowledgments

Martin Ehrendorfer provided valuable comments on an earlier version of this chapter. This work was supported in part by the National Science Foundation under Grant SES-9106440.

References

Blackwell, D. (1953). Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, **24**, 265–272.

Blackwell, D. & Girshick, A. (1954). *Theory of Games and Statistical Decisions*. New York: Wiley.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Brier, G.W. & Allen, R.A. (1951). Verification of weather forecasts. In *Compendium of Meteorology*, ed. T.F. Malone, 841–848. Boston: American Meteorological Society.

Briggs, W.M. & Levine, R.A. (1997). Wavelets and field forecast verification. *Monthly Weather Review*, **125**, in press.

Brooks, H.E. & Doswell, C.A. (1996). A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather and Forecasting*, **11**, 288–303.

Clemen, R.T., Murphy, A.H. & Winkler, R.L. (1995). Screening candidate forecasts: contrasts between choosing and combining. *International Journal of Forecasting*, **11**, 133–146.

Clemen, R.T. & Winkler, R.L. (1987). Calibrating and combining precipitation probability forecasts. In *Probability and Bayesian Statistics*, ed. R. Viertl, 97–110. London: Plenum Press.

DeGroot, M.H. & Fienberg, S.E. (1982). Assessing probability assessors: calibration and refinement. In *Statistical Decision Theory and Related Topics III*, Volume 1, ed. S.S. Gupta & J.O. Berger, 291–314. New York: Academic Press.

Doolittle, M.H. (1885). The verification of predictions. *American Meteorological Journal*, **2**, 327–329.

Ehrendorfer, M. & Murphy, A.H. (1988). Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy. *Monthly Weather Review*, **116**, 1757–1770.

Ehrendorfer, M. & Murphy, A.H. (1992a). Evaluation of prototypical climate forecasts: the sufficiency relation. *Journal of Climate*, **5**, 876–887.

Ehrendorfer, M. & Murphy, A.H. (1992b). On the relationship between the quality and value of weather and climate forecasting systems. *Időjárás*, **96**, 187–206.

Epstein, E.S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, **8**, 985–987.

Finley, J.P. (1884). Tornado prediction. *American Meteorological Journal*, **1**, 85–88.

Gandin, L.S. & Murphy, A.H. (1992). Equitable skill scores for categorical forecasts. *Monthly Weather Review*, **120**, 361–370.

Gilbert, G.K. (1884). Finley's tornado predictions. *American Meteorological Journal*, **1**, 166–172.

Harvey, L.O., Hammond, K.R., Lusk, C.M. & Mross, E.F. (1992). The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review*, **120**, 863–883.

Hoffman, R.N., Liu, Z., Louis, J.-F. & Grassotti, C. (1995). Distortion representation of forecast errors. *Monthly Weather Review*, **123**, 2758–2770.

Hsu, W.-R. & Murphy, A.H. (1986). The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285–293.

Katz, R.W. & Murphy, A.H. (1990). Quality/value relationships for imperfect weather forecasts in a prototype multistage decision-making model. *Journal of Forecasting*, **9**, 75–86.

Katz, R.W., Murphy, A.H. & Winkler, R.L. (1982). Assessing the value of frost forecasts to orchardists: a dynamic decision-making approach. *Journal of Applied Meteorology*, **21**, 518–531.

Krzysztofowicz, R. (1992). Bayesian correlation score: a utilitarian measure of forecast skill. *Monthly Weather Review*, **120**, 208–219.

Krzysztofowicz, R. & Long, D. (1991a). Forecast sufficiency characteristic: construction and application. *International Journal of Forecasting*, **7**, 39–45.

Krzysztofowicz, R. & Long, D. (1991b). Beta likelihood models of probabilistic forecasts. *International Journal of Forecasting*, **7**, 47–55.

Livezey, R.E. (1995). The evaluation of forecasts. In *Analysis of Climate Variability: Applications of Statistical Techniques*, ed. H. von Storch & A. Navarra, 177–196. New York: Springer-Verlag.

Marschak, J. (1971). Economics of information systems. *Journal of the American Statistical Association*, **66**, 192–219.

Mason, I.B. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291–303.

Meglis, A.J. (1960). Annotated bibliography on forecast verification. *Meteorological and Geoastrophysical Abstracts and Bibliography*, **11**, 1129–1174.

Mielke, P.W. (1991). The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth–Science Reviews*, **31**, 55–71.

Murphy, A.H. (1971). A note on the ranked probability score. *Journal of Applied Meteorology*, **10**, 155–156.

Murphy, A.H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, **12**, 215–223.

Murphy, A.H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, **116**, 2417–2424.

Murphy, A.H. (1991). Forecast verification: its complexity and dimensionality. *Monthly Weather Review*, **119**, 1590-1601.

Murphy, A.H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–293.

Murphy, A.H. (1996). The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting*, **11**, 3–20.

Murphy, A.H., Brown, B.G. & Chen, Y.-S. (1989). Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485–501.

Murphy, A.H. & Daan, H. (1985). Forecast evaluation. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, ed. A.H. Murphy & R.W. Katz, 379–437. Boulder, CO: Westview Press.

Murphy, A.H. & Ehrendorfer, M. (1987). On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Weather and Forecasting*, **2**, 243–251.

Murphy, A.H. & Epstein, E.S. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, **117**, 572–581.

Murphy, A.H. & Wilks, D.S. (1996). Statistical models in forecast verification: a case study of precipitation probability forecasts. *Preprints, Thirteenth Conference on Probability and Statistics in Atmospheric Sciences*, 218–223. Boston: American Meteorological Society.

Murphy, A.H. & Winkler, R.L. (1987). A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Murphy, A.H. & Winkler, R.L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435–455.

Murphy, A.H. & Ye, Q. (1990). Comparison of objective and subjective precipitation probability forecasts: the sufficiency relation. *Monthly Weather Review*, **118**, 1783–1792.

Peirce, C.S. (1884). The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Potts, J.M., Folland, C.K, Jolliffe, I.T. & Sexton, D. (1996). Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *Journal of Climate*, **9**, 34–53.

Stanski, H.R., Wilson, L.J. & Burrows, W.R. (1989). Survey of common verification methods in meteorology. Research Report No. 89-5, 114 pp. Toronto: Atmospheric Environment Service.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences.* New York: Academic Press.

Winkler, R.L. & Murphy, A.H. (1968). "Good" probability assessors. *Journal of Applied Meteorology*, **7**, 751–758.