General Decompositions of MSE-Based Skill Scores: Measures of Some Basic Aspects of Forecast Quality

ALLAN H. MURPHY

Prediction and Evaluation Systems, Corvallis, Oregon

(Manuscript received 26 January 1995, in final form 22 January 1996)

ABSTRACT

Skill scores defined as measures of relative mean square error—and based on standards of reference representing climatology, persistence, or a linear combination of climatology and persistence—are decomposed. Two decompositions of each skill score are formulated: 1) a decomposition derived by conditioning on the forecasts and 2) a decomposition derived by conditioning on the observations. These general decompositions contain terms consisting of measures of statistical characteristics of the forecasts and/or observations and terms consisting of measures of basic aspects of forecast quality. Properties of the terms in the respective decompositions are examined, and relationships among the various skill scores—and the terms in the respective decompositions—are described.

Hypothetical samples of binary forecasts and observations are used to illustrate the application and interpretation of these decompositions. Limitations on the inferences that can be drawn from comparative verification based on skill scores, as well as from comparisons based on the terms in decompositions of skill scores, are discussed. The relationship between the application of measures of aspects of quality and the application of the sufficiency relation (a statistical relation that embodies the concept of unambiguous superiority) is briefly explored.

The following results can be gleaned from this methodological study. 1) Decompositions of skill scores provide quantitative measures of—and insights into—multiple aspects of the forecasts, the observations, and their relationship. 2) Superiority in terms of overall skill is no guarantor of superiority in terms of other aspects of quality. 3) Sufficiency (i.e., unambiguous superiority) generally cannot be inferred solely on the basis of superiority over a relatively small set of measures of specific aspects of quality.

Neither individual measures of overall performance (e.g., skill scores) nor sets of measures associated with decompositions of such overall measures respect the dimensionality of most verification problems. Nevertheless, the decompositions described here identify parsimonious sets of measures of basic aspects of forecast quality that should prove to be useful in many verification problems encountered in the real world.

1. Introduction

Skill scores are often used to assess the accuracy of forecasts produced by numerical, statistical, and/or conceptual models relative to the accuracy of forecasts based on simple forecasting methods such as climatology or persistence. Although these measures may be useful as a means of assessing the relative accuracy of forecasts in an overall sense, accuracy is not the only aspect of forecast quality of potential interest or importance. Consideration of the underlying nature of verification problems reveals that these problems are multidimensional and that forecast quality is multifaceted (Murphy and Winkler 1987; Murphy 1991). Thus, one-dimensional measures of skill at best provide a myopic view of forecast quality. From the perspective of comparative verification, skill scores are inadequate—and potentially misleading—when used as the sole or principal means of judging relative forecasting performance.

To obtain more realistic and insightful assessments of forecasting performance, in an absolute or relative sense, the multifaceted nature of forecast quality must be taken into account. The extent to which the forecasts of interest possess various basic aspects of quality is of particular interest. Decompositions of performance measures such as skill scores can play an important role in this assessment process since the terms in these decompositions represent (a) measures of specific aspects of quality *and* (b) contributions to overall skill. In this regard, Murphy (1988) and Murphy and Epstein (1989) recently used the covariance decomposition of the mean square error to investigate the relationships between skill scores based on the mean square error and correlation coefficients as well as to assess correlation-related and other contributions to skill.

The primary purposes of this paper are (a) to present two general decompositions of skill scores based on the

Corresponding author address: Dr. Allan H. Murphy, Prediction and Evaluation Systems, 3115 NW McKinley Drive, Corvallis, OR 97330-1139.
E-mail: murphya@ucs.orst.edu.

mean square error, in which the contributions—to overall skill—of measures of specific basic aspects of quality can be distinguished and (b) to describe the properties of, relationships between, and interpretation and use of these decompositions. The decompositions of the mean square error of interest here were introduced previously in situations involving relatively severe restrictions on the nature of the underlying variables and/or the type of forecasts. These decompositions are now shown to be applicable to all variables and forecasts. Moreover, decompositions are presented for skill scores based on three standards of reference; namely, forecasts produced by climatology, persistence, and a linear combination of climatology and persistence.

Section 2 contains various expressions for the mean square error of the generic (i.e., general) forecasts of interest, as well as expressions for the mean square errors of forecasts based on climatology, persistence, and a linear combination of climatology and persistence. Two general decompositions of the mean square error are described in section 3, one derived by conditioning on the forecasts and the other derived by conditioning on the observations. Basic expressions for the skill scores of interest, and expressions for the various decompositions of these skill scores, are introduced in section 4. The properties of the terms in these decomposed skill scores are also examined and compared in this section. Relationships among the skill scores—and the terms in their respective decompositions—are briefly explored in section 5. Section 6 describes an application of these decompositions to a verification problem involving binary forecasts and observations. Some issues related to the interpretation and use of these decompositions are discussed in section 7, with particular reference to the inferences concerning relative forecasting performance that can be drawn from traditional skill scores, from relatively small sets of measures of basic aspects of forecast quality, and from the sufficiency relation. Section 8 consists of a summary and some concluding remarks.

2. Mean square error

a. MSE for generic forecasts

Let F and X denote the forecasts and observations, respectively, of the underlying variable of interest, and let f and x denote the respective numerical values of these quantities. The mean square error (MSE) of a sample of forecasts and observations can be expressed as follows:

$$\mathrm{MSE} = \sum_f \sum_x p(f, x)(f - x)^2, \qquad (1)$$

where $p(f, x) = \Pr(F = f, X = x)$ represents the empirical joint distribution of forecasts and observations derived from the sample of verification data. Since MSE in (1) is concerned with the average correspondence between forecasts and observations on an individual basis, it represents a measure of accuracy (e.g., Murphy and Daan 1985). Note that MSE ≥ 0, with equality only if $p(f, x) = 0$ for all $f \neq x$.

The joint distribution $p(f, x)$ can be factored into conditional and marginal distributions in two different ways: (i) $p(f, x) = p(x \mid f)p(f)$ and (ii) $p(f, x) = p(f \mid x)p(x)$ (Murphy and Winkler 1987, p. 1332–1333). In these expressions, the distributions $p(x \mid f) = \Pr(X = x \mid F = f)$ and $p(f \mid x) = \Pr(F = f \mid X = x)$ represent the empirical conditional distributions of the observations given the forecasts and the empirical conditional distributions of the forecasts given the observations, respectively. The distributions $p(f) = \Pr(F = f)$ and $p(x) = \Pr(X = x)$ represent the empirical marginal (or unconditional) distributions of the forecasts and observations, respectively.

Murphy and Winkler (1987, 1332–1333) identified expressions (i) and (ii) as the calibration-refinement and likelihood-base rate factorizations, respectively, of the joint distribution $p(f, x)$. In this paper, we refer to the factorization (i) as the conditioning on f (or "cof") factorization and to the factorization (ii) as the conditioning on x (or "cox") factorization. Consideration of the cof and cox factorizations here is motivated by the fact that these factorizations provide conceptual frameworks within which the general decompositions of skill scores of interest can be formulated.

For the purposes of this paper, it is useful to rewrite MSE in (1) in terms of conditional and marginal distributions. To distinguish between these expressions and the basic expression in (1), the MSEs associated with the cof and cox factorizations are denoted by MSE_f and MSE_x, respectively. In the case of the cof factorization, we can rewrite (1) as

$$\mathrm{MSE}_f = \sum_f p(f) \sum_x p(x \mid f)(f - x)^2. \qquad (2)$$

Moreover, denoting the MSE of all forecasts for which $F = f$ by $\mathrm{MSE}(f)$, it follows that

$$\mathrm{MSE}(f) = \sum_x p(x \mid f)(f - x)^2 \qquad (3)$$

and, from (2), that

$$\mathrm{MSE}_f = \sum_f p(f)\mathrm{MSE}(f). \qquad (4)$$

In the case of the cox factorization, (1) can be rewritten as

$$\mathrm{MSE}_x = \sum_x p(x) \sum_f p(f \mid x)(f - x)^2. \qquad (5)$$

Moreover, denoting the MSE of all forecasts for which $X = x$ by $\mathrm{MSE}(x)$, it follows that

$$\mathrm{MSE}(x) = \sum_f p(f \mid x)(f - x)^2 \qquad (6)$$

and, from (5), that

$$\mathrm{MSE}_x = \sum_x p(x)\mathrm{MSE}(x). \tag{7}$$

The expressions denoted by MSE_f in (2) and MSE_x in (5) are used in section 3 as points of departure in the formulation of the cof and cox decompositions, respectively, of the MSE.

b. *MSEs for reference forecasts*

Three standards of reference are used in the skill scores considered in this paper: 1) climatological forecasts, 2) persistence forecasts, and 3) forecasts based on an optimal linear combination of climatological and persistence forecasts (combined climatological–persistence forecasts). Expressions for the MSEs of these reference forecasts are presented in Table 1. The general expression appears in column (i) and the corresponding expression under the condition of complete sample representativeness (see below) is given in column (ii). The conditions under which these expressions were derived are described first and then the expressions are briefly compared.

1) CLIMATOLOGICAL FORECASTS

The mean square error of climatological forecasts, MSE_c, is derived on the basis of a constant forecast equal to the long-term mean of the underlying variable μ (i.e., $f = \mu$ for all f). Note that MSE_c in column (i) (Table 1) is the sum of two nonnegative quantities. These quantities are the sample variance of the observations and the square of the difference between the long-term mean μ and sample mean $\langle x \rangle$.

The degree of correspondence between $\langle x \rangle$ and μ is referred to here as the representativeness of the sample in the mean or, simply, the *sample representativeness* (SR). When $\langle x \rangle = \mu$, SR is said to be complete. Thus, this term in the general expression for MSE_c is a measure of the degree of SR.

Although complete SR seldom occurs in the real world, this special case is still of some interest. Moreover, complete SR might be closely approximated for large verification data samples under the condition of statistical stationarity. The mean square error in this case, MSE_c^*, is given in column (ii) (Table 1).

Comparison of MSE_c and MSE_c^* can be accomplished by computing their ratio. Thus,

$$\frac{\mathrm{MSE}_c}{\mathrm{MSE}_c^*} = 1 + \left(\frac{\mu - \langle x \rangle}{s_x}\right)^2 \tag{8}$$

or, letting $d = (\mu - \langle x \rangle)/s_x$,

$$\frac{\mathrm{MSE}_c}{\mathrm{MSE}_c^*} = 1 + d^2. \tag{9}$$

It is evident that $\mathrm{MSE}_c \geq \mathrm{MSE}_c^*$, with equality only when $d = 0$ (i.e., $\langle x \rangle = \mu$). The quantity d^2, which is a scaled measure of SR, appears in various expressions throughout the paper.

2) PERSISTENCE FORECASTS

The mean square error of persistence forecasts, MSE_p, is derived on the basis of a forecast equal to the observed value of the underlying variable at the initial time, x_0 ($f = x_0$ for all f). The expression for MSE_p in column (i) (Table 1) holds approximately under the condition of negligible end effects. Under this condition, the sample means and sample variances of x_0 and x are equal (i.e., $\langle x_0 \rangle = \langle x \rangle$ and $s_{x_0}^2 = s_x^2$). The quantity r in the expression for MSE_p denotes the sample autocorrelation coefficient between x_0 and x. Since the

TABLE 1. Expressions for the mean square error of the reference forecasts.

Standard of reference (abbreviation)	Mean square error (i) MSE	Mean square error (ii) MSE*
Climatology (c)	$\mathrm{MSE}_c = (d^2 + 1)s_x^2$	$\mathrm{MSE}_c^* = s_x^2$
Persistence (p)	$\mathrm{MSE}_p = 2(1 - r)s_x^2$	$\mathrm{MSE}_p^* = 2(1 - r)s_x^2$
Combined climatology–persistence (cp)	$\mathrm{MSE}_{cp} = [(d^2 + 1)(1 - k)^2 + 2k(1 - r)]s_x^2$	$\mathrm{MSE}_{cp}^* = (1 - r^2)s_x^2$

Key:

$s_x^2 = \Sigma_x\, p(x)(x - \langle x \rangle)^2$

$d^2 = [(\mu - \langle x \rangle)/s_x]^2$

$k = (d^2 + r)/(d^2 + 1)$

$r = s_{x_0 x}/s_x^2$

$s_{x_0 x} = \Sigma_{x_0}\, \Sigma_x p\,(x_0, x)(x_0 - \langle x \rangle)(x - \langle x \rangle)$

Note: The expressions for MSE* in column (ii) hold under the condition of complete SR (i.e., $d = 0$).

long-term climatological mean, μ, does not enter into the expression for MSE_p, the expressions in columns (i) and (ii) (Table 1) are identical.

Note that $MSE_p = 2s_x^2$ when $r = 0$, $MSE_p = s_x^2$ when $r = 1/2$, and $MSE_p = 0$ when $r = 1$. Since the autocorrelation coefficient r is positive for most meteorological variables, the behavior of MSE_p for positive values of r is of primary interest here.

3) Combined climatological–persistence forecasts

The mean square error of combined climatological–persistence forecasts, MSE_{cp}, is derived on the basis of a forecast equal to an optimal linear combination of climatological forecasts ($f = \mu$) and persistence forecasts ($f = x_0$). The general expression for MSE_{cp} in column (i) (Table 1) holds approximately under the condition of negligible end effects (see the appendix). The mean square error of these combined forecasts in the case of complete SR (i.e., $d = 0$), MSE_{cp}^*, is given in column (ii) (Table 1).

Note that $MSE_{cp} = [(1 + 2d^2)s_x^2](1 + d^2)^{-1}$ when $r = 0$ and $MSE_{cp} = 0$ when $r = 1$. Moreover, in the special case in which $d = 0$, $MSE_{cp} = MSE_{cp}^* = s_x^2$ when $r = 0$.

4) Relationships among reference MSEs

The relationships among MSE_c, MSE_p, and MSE_{cp} can be determined by comparing the expressions in Table 1. For example, comparison of MSE_c and MSE_p reveals that $MSE_c < (=, >) MSE_p$ if $r < (=, >)(1/2)(1 - d^2)$. Moreover, in the special case of complete SR (i.e., $d = 0$), $MSE_c < (=, >) MSE_p$ if $r < (=, >) 1/2$ (see Table 1; see also Murphy 1992).

Comparison of MSE_c and MSE_p with MSE_{cp} indicates that $MSE_{cp} \leq \min(MSE_c, MSE_p)$, with equality between MSE_{cp} and MSE_c only when $r = 0$ and with equality between MSE_{cp} and MSE_p only when $r = 1$. Thus, $MSE_{cp} \leq MSE_c \leq MSE_p$ if $r \leq (1/2)(1 - d^2)$ and $MSE_{cp} \leq MSE_p \leq MSE_c$ if $r \geq (1/2)(1 - d^2)$.

3. Decompositions of MSE

As noted in section 1, two general decompositions of the MSE are of interest here. One decomposition is associated with the cof factorization and the other decomposition is associated with the cox factorization. Similar two-step processes are employed to formulate each decomposition.

a. Decomposition associated with cof factorization

In formulating the decomposition of the MSE based on conditioning on the forecasts, the first step consists of expanding the quadratic expression in parentheses on the right-hand side (rhs) of (2) and then applying the summation over x to the various terms in this expansion. As a result,

$$MSE_f = \sum_f p(f)(f^2 - 2f\langle x_f\rangle + \langle x_f^2\rangle), \quad (10)$$

where $\langle x_f\rangle = \Sigma_x p(x|f)x$ is the conditional mean of X given $F = f$ and $\langle x_f^2\rangle = \Sigma_x p(x|f)x^2$ is the conditional mean of X^2 given $F = f$. Then adding and subtracting the quantity $(\langle x_f\rangle)^2$ within the parentheses on the rhs of (10) yields

$$MSE_f = \sum_f p(f)(f - \langle x_f\rangle)^2 + \sum_f p(f)[\langle x_f^2\rangle - (\langle x_f\rangle)^2]. \quad (11)$$

The second step makes use of a basic relationship that exists between the expectations and variances of any two variables. With the cof decomposition in mind, this relationship can be written as

$$V(X) = V[E(X|F)] + E[V(X|F)], \quad (12)$$

where E and V denote the expectation (or mean) and variance, respectively (e.g., Rice 1988, p. 132). Since the second term on the rhs of (11) represents the mean of the sample variance of the variable X_f, it follows from (12) that

$$MSE_f = s_x^2 + \sum_f p(f)(f - \langle x_f\rangle)^2 - \sum_f p(f)(\langle x_f\rangle - \langle x\rangle)^2. \quad (13)$$

The decomposition in (13) is the first of the two general decompositions of interest in this paper.

A brief discussion of the interpretation of the terms on the rhs of (13) is in order here. The first term, s_x^2, is a summary measure of the marginal distribution of observations, $p(x)$. It represents the variability in the forecasting situations (as characterized by the values of X), and it is independent of the forecasts. This term, which is denoted here symbolically by VARX, generally makes a positive contribution to MSE_f.

The second term is a measure of the average squared degree of correspondence between the forecast f and the mean observation given that forecast, $\langle x_f\rangle$, averaged over all f. This aspect of forecasting performance has usually been referred to as *reliability* (e.g., Murphy and Winkler 1987, p. 1332). In this paper, however, it is identified as *type 1 conditional bias;* namely, the bias that exists in the forecasts $F = f$, averaged over all f. As a result, we denote this term by CB_f. (The subscript f is included to identify the term as a measure of the conditional bias associated with the cof decomposition.) Note that CB_f generally contributes positively to (i.e., increases) MSE_f, and it vanishes only if $\langle x_f\rangle = f$ for all f for which $p(f) \neq 0$ (i.e., for type 1 conditionally unbiased—or perfectly reliable—forecasts).

The third term is a measure of the average squared degree of correspondence between the conditional mean observation $\langle x_f\rangle$ and the overall unconditional

mean observation $\langle x \rangle$, again averaged over all f. Since this aspect of forecasting performance is usually referred to as *resolution* (Murphy and Winkler 1987, p. 1337), this term will be denoted by RES. Note that RES generally contributes negatively to (i.e., decreases) MSE_f, and it vanishes only if $\langle x_f \rangle = \langle x \rangle$ for all f for which $p(f) \neq 0$ (i.e., for completely unresolved forecasts).

The general decomposition in (13) can be rewritten in symbolic form as follows:

$$\text{MSE}_f = \text{VARX} + \text{CB}_f - \text{RES}. \quad (14)$$

This decomposition contains one term that depends only on the observations (i.e., VARX) and two terms that depend on the relationship between the forecasts and observations (i.e., CB_f and RES).

The decomposition in (13) was first described by Murphy (1973). At that time, it was derived as a partition of the Brier score (Brier 1950), a special form of the MSE for probability forecasts. As formulated here, it is clear that this decomposition is applicable to all types of forecasts and observations.

b. *Decomposition associated with cox factorization*

The steps involved in formulating a decomposition of the MSE based on conditioning on the observations are identical to the steps described in section 3a, except that the respective roles of the variables F and X are reversed. The first step consists of expanding the quadratic expression in parentheses on the rhs of (5) and then applying the summation over f to the various terms in this expansion. As a result,

$$\text{MSE}_x = \sum_x p(x)(\langle f_x^2 \rangle - 2\langle f_x \rangle x + x^2), \quad (15)$$

where $\langle f_x \rangle = \Sigma_f p(f \mid x) f$ is the conditional mean of F given $X = x$ and $\langle f_x^2 \rangle = \Sigma_f p(f \mid x) f^2$ is the conditional mean of F^2 given $X = x$. Then adding and subtracting the quantity $(\langle f_x \rangle)^2$ within the parentheses on the rhs of (15) yields

$$\text{MSE}_x = \sum_x p(x)(\langle f_x \rangle - x)^2 + \sum_x p(x)[\langle f_x^2 \rangle - (\langle f_x \rangle)^2]. \quad (16)$$

The second step makes use of the basic relationship between the expectations and variances of F and X described by (12), when the roles of these variables are reversed. With the cox decomposition in mind, this relationship becomes

$$V(F) = V[E(F \mid X)] + E[V(F \mid X)]. \quad (17)$$

Since the second term on the rhs of (16) represents the mean of the sample variance of the variable F_x, it follows that

$$\text{MSE}_x = s_f^2 + \sum_x p(x)(\langle f_x \rangle - x)^2 - \sum_x p(x)(\langle f_x \rangle - \langle f \rangle)^2, \quad (18)$$

where $\langle f \rangle = \Sigma_f p(f) f$ is the sample mean of the forecasts and $s_f^2 = \Sigma_f p(f)(f - \langle f \rangle)^2$ is the sample variance of the forecasts. The decomposition in (18) is the second of the two general decompositions of interest in this paper.

With regard to the interpretation of the terms on the rhs of (18), the first term, s_f^2, is a summary measure of the marginal distribution of forecasts, $p(f)$. This term generally makes a positive contribution to MSE_x, and it is denoted here symbolically by VARF.

The second term is a measure of the average squared degree of correspondence between the observation x and the mean forecast given that observation, $\langle f_x \rangle$, averaged over all x. This aspect of forecasting performance is referred to here as *type 2 conditional bias;* namely, the bias that exists in the subsample of forecasts for which $X = x$, averaged over the all x. [Murphy and Winkler (1992) referred to this aspect of quality as type 1 discrimination. The terminology introduced here seems more appropriate.] As a result, it is denoted by CB_x. Note that CB_x generally contributes positively to MSE_x, and it vanishes only if $\langle f_x \rangle = x$ for all x for which $p(x) \neq 0$.

The third term is a measure of the average squared degree of correspondence between the conditional mean forecast $\langle f_x \rangle$ and the overall unconditional mean forecast $\langle f \rangle$, again averaged over all x. This aspect of forecasting performance is referred to here as *discrimination* and it is denoted by DIS. [In essence, this terminology is consistent with the terminology introduced by Murphy and Winkler (1992).] Note that DIS generally contributes negatively to MSE_x, and it vanishes only if $\langle f_x \rangle = \langle f \rangle$ for all x for which $p(x) \neq 0$.

The general decomposition in (18) can be rewritten in symbolic form as follows:

$$\text{MSE}_x = \text{VARF} + \text{CB}_x - \text{DIS}. \quad (19)$$

This decomposition contains one term that depends only on the forecasts (i.e., VARF) and two terms that depend on the relationship between the forecasts and observations (i.e., CB_x and DIS).

The decomposition of MSE_x in (18) was first described in the refereed literature by Murphy and Winkler (1987, p. 1337). On that occasion, it was introduced as a decomposition of the MSE for forecasts of binary variables. As formulated here, it is clear that this decomposition is applicable to all types of variables and forecasts.

4. Skill scores: Decompositions

a. *Basic skill scores*

The basic skill score based on the MSE is denoted here by SS, where

$$SS = 1 - \frac{MSE}{MSE_r}, \qquad (20)$$

in which MSE is the mean square error of the forecasts of interest and MSE_r is the mean square error of the reference forecasts (e.g., see Murphy 1988). According to this definition, SS is the fractional increase (or decrease) in the accuracy of the forecasts of interest over the accuracy of the reference forecasts. In effect, skill as measured by SS represents a reorientation and rescaling of accuracy as measured by MSE. Note that $SS > 0$ for $MSE < MSE_r$, $SS = 0$ for $MSE = MSE_r$, and $SS < 0$ for $MSE > MSE_r$.

For reference forecasts based on a particular simple forecasting method, the corresponding skill score is defined by replacing MSE_r in (20) by the mean square error of the forecasts produced by this method. The skill scores based on climatological forecasts, persistence forecasts, and combined climatological–persistence forecasts are denoted here by SS_c, SS_p, and SS_{cp}, respectively. Expressions for SS_c, SS_p, and SS_{cp}, in terms of the respective mean square errors, are presented in Table 2.

TABLE 2. Basic expressions for the skill scores in terms of the mean square errors of the three types of reference forecasts.

Standard of reference (abbreviation)	Skill score (SS)
Climatology (c)	$SS_c = 1 - (MSE/MSE_c)$
Persistence (p)	$SS_p = 1 - (MSE/MSE_p)$
Climatology and persistence combined (cp)	$SS_{cp} = 1 - (MSE/MSE_{cp})$

Note: The basic expression for MSE appears as Eq. (1) in the text, and the expressions for MSE_c, MSE_p, and MSE_{cp} are given in Table 1.

1) COF DECOMPOSITION

The decomposition of the basic skill score associated with the cof factorization, SS_f, is obtained (in symbolic form) by substituting the expression for the cof decomposition of MSE_f in (14) into the expression for SS in (20):

$$SS_f = 1 - \frac{VARX + CB_f - RES}{MSE_r}. \qquad (21)$$

Examination of the rhs of (21) reveals that overall skill, as measured by SS_f, consists of three distinct components; namely, a component that is related to the variability of the observations, a component that is related to the type 1 conditional bias of the forecasts, and a component that is related to the resolution of the forecasts. These contributions to the overall skill score are all scaled by the mean square error of the reference forecasts, MSE_r.

The first term on the rhs of (21) (i.e., $VARX/MSE_r$) is independent of the forecasts of interest. It depends on the variance of the observations, VARX ($=s_x^2$), as well as on various parameters associated with the reference forecasts (i.e., μ, $\langle x \rangle$, and/or r in the case of the reference forecasts of concern here). Considered in isolation, this term generally makes a negative contribution to SS_f; that is, skill decreases as the variability of the observations increases. However, it should be kept in mind that the variability of the observations can also influence the magnitudes of other terms in (21).

The second and third terms on the rhs of (21) constitute the contributions to SS_f due to two basic aspects of the quality of the forecasts of interest when the verification data sample is conditioned on the forecasts. The term CB_f/MSE_r measures type 1 conditional bias, and this term generally makes a negative contribution to skill. It will be referred to here as the penalty assigned to forecasts for which $\langle x_f \rangle \neq f$ for some f, or simply the *type 1 conditional bias penalty.* The term RES/MSE_r measures resolution, and this term generally makes a positive contribution to skill. It will be referred to here as the reward given to forecasts for which $\langle x_f \rangle \neq \langle x \rangle$ for some f, or simply the *resolution reward.*

2) COX DECOMPOSITION

The decomposition of the basic skill score associated with the cox factorization, SS_x, is obtained (in symbolic form) by substituting the expression for the cox decomposition of MSE_x in (19) into the expression for SS in (20):

$$SS_x = 1 - \frac{VARF + CB_x - DIS}{MSE_r}. \qquad (22)$$

Examination of the rhs of (22) reveals that overall skill, as measured by SS_x, also consists of three distinct components; namely, a component that is related to the variability of the forecasts, a component that is related to the type 2 conditional bias of the forecasts, and a component that is related to the discrimination of the forecasts. These contributions to the overall skill score are all scaled by the mean square error of the reference forecasts, MSE_r.

The first term on the rhs of (22) (i.e., $VARF/MSE_r$) depends on the variance of the forecasts, VARF ($=s_f^2$), as well as on various parameters associated with the reference forecasts. Considered in isolation, this term generally makes a negative contribution to SS_x; that is, skill decreases as the variability of the forecasts increases. However, it should be kept in mind that the variability of the forecasts can also influence the magnitudes of other terms in (22).

The second and third terms on the rhs of (22) constitute the contributions to SS_x due to two basic aspects of the quality of the forecasts of interest when the verification data sample is conditioned on the observations. The term CB_x/MSE_r measures type 2 conditional bias, and this term generally makes a negative contri-

bution to skill. It will be referred to here as the penalty assigned to forecasts for which $\langle f_x \rangle \neq x$ for some x, or simply the *type 2 conditional bias penalty.* The term DIS/MSE_r measures discrimination, and this term generally makes a positive contribution to skill. It will be referred to here as the reward given to forecasts for which $\langle f_x \rangle \neq \langle f \rangle$ for some x, or simply, the *discrimination reward.*

b. Skill scores based on climatology

1) COF DECOMPOSITION

Let SS_{cf} denote the decomposed skill score associated with climatological reference forecasts and the cof factorization. Then, replacing MSE_r in (21) by the expression for MSE_c in Table 1, it follows that

$$\text{SS}_{cf} = \frac{d^2 s_x^2 + \text{RES} - \text{CB}_f}{(d^2 + 1)s_x^2}. \tag{23}$$

Thus, $\text{SS}_{cf} > (=, <)\ 0$ when $\text{RES} > (=, <)\ \text{CB}_f - d^2 s_x^2$. That is, skill is positive in this case when the resolution reward exceeds the difference between the type 1 conditional bias penalty and the SR term [the latter is defined here as $d^2 s_x^2 = (\mu - \langle x \rangle)^2$]. The contributions to SS_{cf} due to the SR term, resolution reward, and type 1 conditional bias penalty are defined in terms of basic sample quantities in Table 3.

Note that the SR term in (23) makes a positive contribution to overall skill, as measured by SS_{cf}, when $d \neq 0$. In this regard, $\text{SS}_{cf} > 0$ when $\text{RES} = 0$ if $\text{CB}_f < d^2 s_x^2$. That is, completely unresolved forecasts will exhibit positive skill if the type 1 conditional bias penalty is less than the SR term. Moreover, completely unresolved but perfectly reliable forecasts (i.e., $f = \langle x \rangle$ for all f) will exhibit positive skill if the SR term is positive (i.e., $\langle x \rangle \neq \mu$).

In the special case in which $d = 0$ (i.e., complete SR), SS_{cf} is denoted here by SS_{cf}^*, where

$$\text{SS}_{cf}^* = \frac{\text{RES} - \text{CB}_f}{s_x^2}. \tag{24}$$

In this case, $\text{SS}_{cf}^* > (=, <)\ 0$ when $\text{RES} > (=, <)\text{CB}_f$. Thus, skill is positive in this special case if the resolution reward exceeds the type 1 conditional bias (i.e., reliability) penalty. The contributions to skill from the resolution reward and the type 1 conditional bias penalty are RES/s_x^2 and $-\text{CB}_f/s_x^2$, respectively.

From (23) and (24), it follows that

$$\text{SS}_{cf} = \frac{d^2 + \text{SS}_{cf}^*}{d^2 + 1}. \tag{25}$$

This expression implies that $\text{SS}_{cf} \geqslant \text{SS}_{cf}^*$, with equality only when $d = 0$ (complete SR). Therefore, when SR is incomplete ($d \neq 0$), skill measured relative to long-term climatology equals or exceeds skill measured relative to sample climatology.

2) COX DECOMPOSITION

Let SS_{cx} denote the decomposed skill score associated with climatological reference forecasts and the cox factorization. Then, replacing MSE_r in (22) by the expression for MSE_c in Table 1, it follows that

$$\text{SS}_{cx} = \frac{(d^2 + 1 - v^2)s_x^2 + \text{DIS} - \text{CB}_x}{(d^2 + 1)s_x^2}, \tag{26}$$

where $v^2 = (s_f/s_x)^2$. Thus, $\text{SS}_{cx} > (=, <)\ 0$ when $\text{DIS} > (=, <)\text{CB}_x - (d^2 + 1 - v^2)s_x^2$. That is, skill is positive in this case when the discrimination reward exceeds the difference between the type 2 conditional bias penalty and the SR term [here the latter is defined as $(d^2 + 1 - v^2)s_x^2 = (\mu - \langle x \rangle)^2 + (s_x^2 - s_f^2)$]. The contributions to SS_{cx} due to the SR term, discrimination reward, and type 2 conditional bias penalty are defined in terms of basic sample quantities in Table 4.

Note that the SR term in (26) makes a positive contribution to overall skill, as measured by SS_{cx}, when $(d^2 + 1)s_x^2 > s_f^2$. In this regard, forecasts for which $\text{DIS} = 0$ are of positive skill if $\text{CB}_x < (\mu - \langle x \rangle)^2 + (s_x^2 - s_f^2)$. That is, forecasts possessing no discrimination reward will exhibit positive skill if the type 2 conditional bias penalty is less than the SR term. More-

TABLE 3. Expressions for the various contributions to the skill score associated the cof factorization, SS_{rf}, in the cases of climatological reference forecasts ($\text{SS}_{rf} = \text{SS}_{cf}$), persistence reference forecasts ($\text{SS}_{rf} = \text{SS}_{pf}$), and combined climatological–persistence reference forecasts ($\text{SS}_{rf} = \text{SS}_{cpf}$).

Skill score	Contribution due to SR and/or AC		Contribution due to RES		Contribution due to CB_f
SS_{cf}	$d^2/(d^2 + 1)$	+	$\text{RES}/(d^2 + 1)s_x^2$	−	$\text{CB}_f/(d^2 + 1)s_x^2$
SS_{pf}	$(1 - 2r)/2(1 - r)$	+	$\text{RES}/2(1 - r)s_x^2$	−	$\text{CB}_f/2(1 - r)s_x^2$
SS_{cpf}	$[d^2(1 - k)^2 + k(k - 2r)]s_x^2/\text{DEN}$	+	RES/DEN	−	CB_f/DEN

Key:

$s_x^2 = \Sigma_x\, p(x)(x - \langle x \rangle)^2$, $\text{RES} = \Sigma_f\, p(f)(\langle x_f \rangle - \langle x \rangle)^2$, $\text{CB}_f = \Sigma_f\, p(f)(f - \langle x_f \rangle)^2$

$\text{DEN} = \text{MSE}_{cp} = [(d^2 + 1)(1 - k)^2 + 2k(1 - r)]s_x^2$

$d^2 = [(\mu - \langle x \rangle)/s_x]^2$, $k = (d^2 + r)/(d^2 + 1)$, $r = s_{x_0x}/s_x^2$, $s_{x_0x} = \Sigma_{x_0} \Sigma_x\, p(x_0, x)(x_0 - \langle x \rangle)(x - \langle x \rangle)$

TABLE 4. Expressions for the various contributions to the skill score associated with the cox factorization, SS_{rx}, in the cases of climatological reference forecasts ($SS_{rx} = SS_{cx}$), persistence reference forecasts ($SS_{rx} = SS_{px}$), and combined climatological–persistence reference forecasts ($SS_{rx} = SS_{cpx}$).

Skill score	Contribution due to SR and/or AC		Contribution due to DIS		Contribution due to CB_x
SS_{cx}	$(d^2 + 1 - v^2)/(d^2 + 1)$	+	$DIS/(d^2 + 1)s_x^2$	−	$CB_x/(d^2 + 1)s_x^2$
SS_{px}	$[2(1 - r) - v^2]/2(1 - r)$	+	$DIS/2(1 - r)s_x^2$	−	$CB_x/2(1 - r)s_x^2$
SS_{cpx}	$[(d^2 + 1)(1 - k)^2 + 2k(1 - r) - v^2]s_x^2/DEN$	+	DIS/DEN	−	CB_x/DEN

Key:

$s_f^2 = \Sigma_f\, p(f)(f - \langle f\rangle)^2$, $DIS = \Sigma_x\, p(x)(\langle f_x\rangle - \langle f\rangle)^2$, $CB_x = \Sigma_x\, p(x)(\langle f_x\rangle - x)^2$

$s_x^2 = \Sigma_x\, p(x)(x - \langle x\rangle)^2$, $DEN = MSE_{cp} = [(d^2 + 1)(1 - k)^2 + 2k(1 - r)]s_x^2$

$d^2 = [(\mu - \langle x\rangle)/s_x]^2$, $v^2 = (s_f/s_x)^2$, $k = (d^2 + r)/(d^2 + 1)$, $r = s_{x_0x}/s_x^2$,

$s_{x_0x} = \Sigma_{x_0}\Sigma_x\, p(x_0, x)(x_0 - \langle x\rangle)(x - \langle x\rangle)$

over, forecasts for which DIS = 0 and $CB_x = 0$ will exhibit positive skill if the SR term is positive [i.e., if $(\mu - \langle x\rangle)^2 > s_f^2 - s_x^2$].

When $d = 0$ (i.e., complete SR), SS_{cx} is denoted here by SS_{cx}^*, where

$$SS_{cx}^* = \frac{(s_x^2 - s_f^2) + DIS - CB_x}{s_x^2}. \tag{27}$$

In this case, $SS_{cx}^* > (=, <)\ 0$ when $DIS > (=, <)\ CB_x + (s_f^2 - s_x^2)$. Thus, skill is positive in this special case if the DIS reward is greater than the sum of the CB_x penalty and the difference between the sample variances of F and X. Moreover, for those verification data samples for which $s_f^2 < s_x^2$ (an inequality that holds for most such samples), the condition $DIS \geq CB_x$ is sufficient to ensure positive skill.

Note that, from (26) and (27),

$$SS_{cx} = \frac{d^2 + SS_{cx}^*}{d^2 + 1}. \tag{28}$$

Since $SS_{cx}^* \leq 1$, it follows that $SS_{cx} \geq SS_{cx}^*$, with equality only when $d = 0$ (complete SR). Thus, if SR is incomplete ($d \neq 0$), skill measured relative to long-term climatology equals or exceeds skill measured relative to sample climatology.

c. *Skill scores based on persistence*

1) COF DECOMPOSITION

Let SS_{pf} denote the decomposed skill score associated with persistence reference forecasts and the cof factorization. Then, replacing MSE_r in (21) by the expression for MSE_p in Table 1, it follows that

$$SS_{pf} = \frac{(1 - 2r)s_x^2 + RES - CB_f}{2(1 - r)s_x^2}. \tag{29}$$

Thus, $SS_{pf} > (=, <)\ 0$ when $RES > (=, <)\ CB_f - (1 - 2r)s_x^2$. That is, skill is positive in this case when the resolution reward exceeds the difference between the type 1 conditional bias penalty and the autocorrelation coefficient (AC) term [the latter is defined here as $(1 - 2r)s_x^2$]. The contributions to SS_{pf} due to the AC term, resolution reward, and type 1 conditional bias penalty are defined in terms of basic sample quantities in Table 3.

It is interesting to note that $SS_{pf} > 0$ when RES = 0 if $CB_f < (1 - 2r)s_x^2$. That is, completely unresolved forecasts will exhibit positive skill if the type 1 conditional bias penalty is less than the AC term. Moreover, completely unresolved but type 1 conditionally unbiased forecasts (i.e., $f = \langle x\rangle$ for all f) will exhibit positive skill if $r < 1/2$.

2) COX DECOMPOSITION

Let SS_{px} denote the decomposed skill score associated with persistence reference forecasts and the cox factorization. Then, replacing MSE_r in (22) by the expression for MSE_p in Table 1, it follows that

$$SS_{px} = \frac{[2(1 - r) - v^2]s_x^2 + DIS - CB_x}{2(1 - r)s_x^2}. \tag{30}$$

Thus, $SS_{px} > (=, <)\ 0$ when $DIS > (=, <)\ CB_x - [2(1 - r) - v^2]s_x^2$. That is, skill is positive in this case when the discrimination reward exceeds the difference between the type 2 conditional bias penalty and the AC term {the latter is defined here as $[2(1 - r) - v^2]s_x^2$}. The contributions to SS_{px} due to the AC term, discrimination reward, and type 2 conditional bias penalty are defined in terms of basic sample quantities in Table 4.

It is interesting to note that $SS_{px} > 0$ when DIS = 0 if $CB_x < [2(1 - r) - v^2]s_x^2$. That is, forecasts possessing no discrimination reward will exhibit positive skill if the type 2 conditional bias penalty is less than the AC term. Moreover, forecasts possessing no discrimination reward and no type 2 conditional bias pen-

alty (i.e., $\langle f \rangle = x$ for all x) will exhibit positive skill if $r < 1 - (1/2)v^2$.

d. Skill scores based on combined climatology and persistence

1) COF DECOMPOSITION

Let SS_{cpf} denote the decomposed skill score associated with combined climatological–persistence reference forecasts and the cof factorization. Then, replacing MSE_r in (21) by the expression for MSE_{cp} in Table 1, it follows that

$$SS_{cpf} = \frac{[d^2(1-k)^2 + k(k-2r)]s_x^2 + RES - CB_f}{MSE_{cp}}, \tag{31}$$

where $MSE_{cp} = [(d^2+1)(1-k)^2 + 2k(1-r)]s_x^2$ (see Table 1). Thus, $SS_{cpf} > (=, <)\ 0$ when $RES > (=, <)\ CB_f - [d^2(1-k)^2 + k(k-2r)]s_x^2$. That is, skill is positive in this case when the resolution reward exceeds the difference between the type 1 conditional bias penalty and the SR/AC term {the latter is defined here as $[d^2(1-k)^2 + k(k-2r)]s_x^2$}. The contributions to SS_{cpf} due to the SR/AC term, resolution reward, and type 1 conditional bias penalty are defined in terms of basic sample quantities in Table 3.

Note that $SS_{cpf} > 0$ when $RES = 0$ if $CB_f < [d^2(1-k)^2 + k(k-2r)]s_x^2$. That is, completely unresolved forecasts will exhibit positive skill if the type 1 conditional bias penalty is less than the SR/AC term. Moreover, completely unresolved but type 1 conditionally unbiased forecasts (i.e., $f = \langle x \rangle$ for all f) will exhibit positive skill if $d^2(1-k)^2 + k(k-2r) > 0$.

In the special case in which $d = 0$ (complete SR), SS_{cpf} becomes SS^*_{cpf}, where

$$SS^*_{cpf} = \frac{RES - CB_f - r^2 s_x^2}{(1-r^2)s_x^2}. \tag{32}$$

In this case, $SS^*_{cpf} > (=, <)\ 0$ when $RES > (=, <)\ CB_f + r^2 s_x^2$.

2) COX DECOMPOSITION

Let SS_{cpx} denote the decomposed skill score associated with combined climatological–persistence reference forecasts and the cox factorization. Then, replacing MSE_r in (22) by the expression for MSE_{cp} in Table 1, it follows that

$$SS_{cpx} = \frac{[(d^2+1)(1-k)^2 + 2k(1-r) - v^2]s_x^2 + DIS - CB_x}{MSE_{cp}}. \tag{33}$$

Thus, $SS_{cpx} > (=, <)\ 0$ when $DIS > (=, <)\ CB_x - MSE_{cp} + s_f^2$. That is, skill is positive in this case when the discrimination reward exceeds the difference between the type 2 conditional bias penalty and the SR/AC term (here the latter is defined as $MSE_{cp} - s_f^2$). The contributions to SS_{cpx} due to the SR/AC term, discrimination reward, and type 2 conditional bias penalty are defined in terms of basic sample quantities in Table 4.

Note that forecasts for which $DIS = 0$ are of positive skill if $CB_x < MSE_{cp} - s_f^2$. That is, forecasts possessing no discrimination reward will exhibit positive skill if the type 2 conditional bias penalty is less than the SR/AC term. Moreover, forecasts for which $DIS = 0$ and $CB_x = 0$ will exhibit positive skill if the SR/AC term is positive [i.e., if $(d^2+1)(1-k)^2 + 2k(1-r) > v^2$].

In the special case in which $d = 0$ (complete SR), SS_{cpx} becomes SS^*_{cpx}, where

$$SS^*_{cpx} = \frac{(1-r^2-v^2)s_x^2 + DIS - CB_x}{(1-r^2)s_x^2}. \tag{34}$$

In this case, $SS^*_{cpx} > (=, <)\ 0$ if $DIS > (=, <) CB_x - (1-r^2-v^2)s_x^2$.

5. Comparison of skill scores

The magnitudes of the skill scores based on the climatological, persistence, and combined climatological–persistence reference forecasts are compared in this section. In the cases of the decomposed skill scores derived from the cof and cox factorizations, these comparisons are concerned with the relative magnitudes of the corresponding terms in the respective decompositions.

a. Overall skill scores

Since the overall skill scores of interest here are defined as the fractional decrease (or increase) in the MSE of the forecasts over the MSE of the respective reference forecasts (see section 4), the relationships among these skill scores are analogous to the relationships among the mean square errors of the reference forecasts. That is, 1) $SS_c < (=, >)\ SS_p$ if $r < (=, >)(1/2)(1-d^2)$ and 2) $SS_{cp} \leq \min(SS_c, SS_p)$, with equality between SS_{cp} and SS_c only when $r = 0$ and with equality between SS_{cp} and SS_p only when $r = 1$ (see section 2b). Thus,

$$SS_{cp} \leq SS_c \leq SS_p \quad \text{if} \quad r \leq \left(\frac{1}{2}\right)(1-d^2) \tag{35}$$

and

$$SS_{cp} \leq SS_p \leq SS_c \quad \text{if} \quad r \geq \left(\frac{1}{2}\right)(1-d^2). \tag{36}$$

In the special case of complete SR (i.e., $d = 0$), the inequalities in (35) and (36) hold under the conditions $r \leq 1/2$ and $r \geq 1/2$, respectively.

TABLE 5. (a) Augmented 2 × 2 contingency table depicting the joint and marginal distributions of forecasts and/or observations in a binary situation. (b) Definitions of joint, conditional, and marginal probabilities of forecasts and/or observations in a binary situation.

(a) Joint and marginal distributions

		Observations		
Forecasts	$p(f, x)$	$x = 1$	$x = 0$	$p(f)$
	$f = 1$	p_{11}	p_{10}	$p_1(f)$
	$f = 0$	p_{01}	p_{00}	$p_0(f)$
	$p(x)$	$p_1(x)$	$p_0(x)$	1

(b) Joint, conditional, and marginal probabilities

Joint probabilities: $p(f, x)$

$p_{11} = \Pr(f = 1, x = 1)$
$p_{10} = \Pr(f = 1, x = 0)$
$p_{01} = \Pr(f = 0, x = 1)$
$p_{00} = \Pr(f = 0, x = 0)$
$p_{11} + p_{10} + p_{01} + p_{00} = 1$

Marginal probabilities: $p(f)$ and $p(x)$

$p_1(f) = \Pr(f = 1) = p_{11} + p_{10}$
$p_0(f) = \Pr(f = 0) = p_{01} + p_{00}$
$p_1(f) + p_0(f) = 1$
$p_1(x) = \Pr(x = 1) = p_{11} + p_{01}$
$p_0(x) = \Pr(x = 0) = p_{10} + p_{00}$
$p_1(x) + p_0(x) = 1$

Conditional probabilities: $p(x \mid f)$ and $p(f \mid x)$

$p_{11}(f) = \Pr(x = 1 \mid f = 1) = p_{11}/p_1(f)$
$p_{10}(f) = \Pr(x = 0 \mid f = 1) = p_{10}/p_1(f), p_{11}(f) + p_{10}(f) = 1$
$p_{01}(f) = \Pr(x = 1 \mid f = 0) = p_{01}/p_0(f)$
$p_{00}(f) = \Pr(x = 0 \mid f = 0) = p_{00}/p_0(f), p_{01}(f) + p_{00}(f) = 1$
$p_{11}(x) = \Pr(f = 1 \mid x = 1) = p_{11}/p_1(x)$
$p_{01}(x) = \Pr(f = 0 \mid x = 1) = p_{01}/p_1(x), p_{11}(x) + p_{01}(x) = 1$
$sp_{10}(x) = \Pr(f = 1 \mid x = 0) = p_{10}/p_0(x)$
$p_{00}(x) = \Pr(f = 0 \mid x = 0) = p_{00}/p_0(x), p_{10}(x) + p_{00}(x) = 1$

b. Terms in decomposed skill scores

Comparisons of the terms in the decomposed skill scores associated with a particular factorization yield relationships analogous to those reported in section 5a for the overall skill scores. For example, let CB_{cf}, CB_{pf}, and CB_{cpf} denote the contributions to overall skill due to the conditional bias terms in the respective skill scores associated with the cof factorization. Then, it follows from the relationships among the MSEs of the reference forecasts that

$$CB_{cpf} \geqslant CB_{cf} \geqslant CB_{pf} \quad \text{if} \quad r \leqslant \left(\frac{1}{2}\right)(1 - d^2) \quad (37)$$

and

$$CB_{cpf} \geqslant CB_{pf} \geqslant CB_{cf} \quad \text{if} \quad r \geqslant \left(\frac{1}{2}\right)(1 - d^2). \quad (38)$$

Thus, the type 1 conditional bias penalty in the case of a MSE-based skill score employing combined climatological–persistence reference forecasts equals or exceeds the type 1 conditional bias penalty in the case of a MSE-based skill score employing either climatological or persistence reference forecasts. Similar relationships hold for the other terms in the decomposition associated with the cof factorization (i.e., the SR or AC term, the RES term), as well as for the terms in the decomposition associated with the cox factorization (i.e., the SR and/or AC term, the CB_x term, the DIS term).

6. Application: Binary forecasts and observations

Computation and interpretation of the skill-score decompositions introduced in section 4 are illustrated here by evaluating hypothetical samples of binary forecasts and observations. Specifically, the relative quality of the forecasts is assessed by comparing the magnitudes of the terms in these decompositions. Discussion of the inferences that can be drawn from these—and other—methods of assessing relative forecasting performance in this context is postponed until section 7. An application of the cof and cox decompositions of skill scores based on climatological reference forecasts to verification data samples consisting of precipitation probability forecasts and binary observations has been reported by Murphy and Winkler (1992).

a. Basic definitions and expressions

The forecasts and observations considered here are binary in the sense that $F = 0$ or 1 and $X = 0$ or 1 only. In this situation the joint distribution of forecasts and observations, $p(f, x)$, can be summarized in the form of a 2 × 2 contingency table. Table 5a represents an augmented 2 × 2 contingency table that identifies the probabilities that constitute this joint distribution, as well as the probabilities that constitute the marginal distributions of the forecasts and observations, $p(f)$ and $p(x)$, respectively. Table 5b defines these joint and marginal probabilities, as well as the probabilities that constitute the conditional distributions of the observations given the forecasts, $p(x|f)$, and the conditional distributions of the forecasts given the observations, $p(f|x)$.

Measures of the various aspects of quality of interest here are expressed in terms of these joint, conditional, and marginal probabilities in Table 6. Specifically, Table 6a contains expressions for the basic MSE as well as the overall MSEs and SSs associated with the three types of reference forecasts. For convenience, it has been assumed that the sample and long-term climatological means of the underlying variable are equal (i.e., SR is assumed to be complete). Tables 6b and 6c contain expressions for the MSEs, SSs, and terms in the decompositions of these measures in the cases of the cof and cox factorizations, respectively.

b. Numerical results: Computation and interpretation

The joint and marginal distributions for hypothetical verification data samples associated with three forecasting methods—denoted here by A, B, and C—are depicted in Table 7 in the form of augmented 2 × 2

TABLE 6. Expressions for MSEs and SSs, and terms in their respective decompositions, based on climatological, persistence, and combined climatological–persistence reference forecasts, in a situation involving binary forecasts and observations with complete sample representativeness in the mean (i.e., $\langle x \rangle = \mu$). (a) Overall measures. (b) Measures associated with the cof factorization. (c) Measures associated with the cox factorization.

(a) Overall measures
$\text{MSE} = p_{10} + p_{01}$
$\text{MSE}_c^* = s_x^2 = p_1(x)p_0(x)$, $\text{SS}_c^* = 1 - (\text{MSE}/\text{MSE}_c^*)$
$\text{MSE}_p = 2(1 - r)s_x^2 = 2(1 - r)p_1(x)p_0(x)$,
$\text{SS}_p = 1 - (\text{MSE}/\text{MSE}_p)$
$\text{MSE}_{cp}^* = (1 - r^2)s_x^2 = (1 - r^2)p_1(x)p_0(x)$,
$\text{SS}_{cp}^* = 1 - (\text{MSE}/\text{MSE}_{cp}^*)$

(b) Measures associated with cof factorization
$\text{MSE}_f = s_x^2 + \text{CB}_f - \text{RES}$
$s_x^2 = p_1(x)p_0(x)$
$\text{CB}_f = p_{10}p_{10}(f) + p_{01}p_{01}(f)$
$\text{RES} = p_{10}[p_{10}(f) - p_0(x)] + p_{01}[p_{01}(f) - p_1(x)]$
$+ (p_{11}p_{00} - p_{10}p_{01})$
$\text{SS}_{cf}^* = 1 - (\text{MSE}_f/\text{MSE}_c^*)$
$\text{SS}_{pf} = 1 - (\text{MSE}_f/\text{MSE}_p)$
$\text{SS}_{cpf}^* = 1 - (\text{MSE}_f/\text{MSE}_{cp}^*)$

(c) Measures associated with cox factorization
$\text{MSE}_x = s_f^2 + \text{CB}_x - \text{DIS}$
$s_f^2 = p_1(f)p_0(f)$
$\text{CB}_x = p_{10}p_{10}(x) + p_{01}p_{01}(x)$
$\text{DIS} = p_{10}[p_{10}(x) - p_1(f)] + p_{01}[p_{01}(x) - p_0(f)]$
$+ (p_{11}p_{00} - p_{10}p_{01})$
$\text{SS}_{cx}^* = 1 - (\text{MSE}_x/\text{MSE}_c^*)$
$\text{SS}_{px} = 1 - (\text{MSE}_x/\text{MSE}_p)$
$\text{SS}_{cpx}^* = 1 - (\text{MSE}_x/\text{MSE}_{cp}^*)$

contingency tables. It has been assumed that the forecasts produced by these methods have been made on the same set of forecasting occasions. Thus, the distribution of observations—that is, $p(x)$—is the same for all three verification data samples. The probabilities that constitute this distribution, $p_1(x) = 0.25$ and $p_0(x) = 0.75$, represent the sample climatological probabilities of the events of interest on these occasions. The relevant conditional distributions, $p(x|f)$ and $p(f|x)$, can be calculated from the joint and marginal distributions (see Table 5b).

Numerical values of the MSE—and of the terms in the decompositions of the MSE associated with the cof and cox factorizations—for these data samples are recorded in Table 8a. According to the MSE, method B's forecasts are the most accurate and method A's forecasts are the least accurate. The other quantities in Table 8a are numerical values of the terms in the decompositions of MSE_f and MSE_x [see (14) and (19), respectively]. For example, $\text{MSE}_f = 0.19 = \text{VARX} + \text{CB}_f - \text{RES} = 0.1875 + 0.0550 - 0.0525$ and $\text{MSE}_x = 0.19 = \text{VARF} + \text{CB}_x - \text{DIS} = 0.2100 + 0.0388 - 0.0588$ in the case of A's forecasts. The numerical values of the terms in the decomposition of MSE_f indicate that B's forecasts possess less conditional bias (in the type 1 sense) and exhibit greater resolution than A's and C's forecasts [note that $\text{VARX} = p_1(x)p_0(x) = 0.1875$ for all three hypothetical data samples]. Comparison of A's and C's conditional bias and resolution terms indicates that C's forecasts possess less conditional bias than A's forecasts but that A's forecasts exhibit greater resolution than C's forecasts.

Evaluation of relative forecasting performance using terms in the decomposition of MSE_x is complicated by the fact that *three* forecast-dependent quantities contribute to overall accuracy in this framework. That is, VARF—unlike VARX in the decomposition of MSE_f—generally varies from forecasting method to forecasting method. In this regard, method A possesses a smaller conditional bias penalty (in the type 2 sense) and a larger discrimination reward than method B, but it must be kept in mind that A's forecasts exhibit considerably greater variability than B's forecasts.

Tables 8b, 8c, and 8d contain numerical values of SS_c^*, SS_p, and SS_{cp}^*, respectively, as well as the terms in the cof and cox decompositions of these skill scores, for the three verification data samples. In calculating these quantities, it has been assumed that $\langle x \rangle = \mu$ ($d = 0$) and $r = 0.4$. Under these assumptions, and given the sample climatological probabilities in Table 7, it follows that $\text{MSE}_c = \text{MSE}_c^* = 0.1875$, $\text{MSE}_p = 0.2250$, and $\text{MSE}_{cp} = \text{MSE}_{cp}^* = 0.1575$ (see Table 1).

The overall skill scores indicate that B's forecasts are most skillful and A's forecasts are least skillful. As expected [see (35)], $\text{SS}_{cp}^* \leq \text{SS}_c^* \leq \text{SS}_p$ for all three forecasting methods. Note that the skill scores are negative for methods A and C when more accurate reference forecasts are used to define the zero point on the

TABLE 7. Joint and marginal distributions of forecasts and/or observations for hypothetical verification data samples for alternative forecasting methods. (a) Method A. (b) Method B. (c) Method C.

(a) Method A

	Observations			
Forecasts	$p(f, x)$	$x = 1$	$x = 0$	$p(f)$
	$f = 1$	0.18	0.12	0.30
	$f = 0$	0.07	0.63	0.70
	$p(x)$	0.25	0.75	1

(b) Method B

	Observations			
Forecasts	$p(f, x)$	$x = 1$	$x = 0$	$p(f)$
	$f = 1$	0.15	0.05	0.20
	$f = 0$	0.10	0.70	0.80
	$p(x)$	0.25	0.75	1

(c) Method C

	Observations			
Forecasts	$p(f, x)$	$x = 1$	$x = 0$	$p(f)$
	$f = 1$	0.15	0.08	0.23
	$f = 0$	0.10	0.67	0.77
	$p(x)$	0.25	0.75	1

TABLE 8. Terms in decompositions of (a) MSE, (b) SS_c^*, (c) SS_p, and (d) SS_{cp}^* for hypothetical forecasts produced by forecasting methods A, B, and C.

(a) MSE

Method	MSE	VARX	CB_f	RES	VARF	CB_x	DIS
A	0.19	0.1875	0.0550	0.0525	0.2100	0.0388	0.0588
B	0.15	0.1875	0.0250	0.0625	0.1600	0.0433	0.0533
C	0.18	0.1875	0.0408	0.0483	0.1771	0.0485	0.0456

(b) SS_c^* ($MSE_c^* = 0.1875$, with $\langle x \rangle = \mu = 0.25$)

Method	SS_c^*	$1 - VARX_c^*$	RES_c^*	CB_{cf}^*	$1 - VARF_c^*$	DIS_c^*	CB_{cx}^*
A	−0.0133	0	0.2800	0.2933	−0.1200	0.3136	0.2069
B	0.2000	0	0.3333	0.1333	0.1467	0.2843	0.2309
C	0.0400	0	0.2576	0.2176	0.0555	0.2432	0.2587

(c) SS_p ($MSE_p = 0.2250$, with $r = 0.4$)

Method	SS_p	$1 - VARX_p$	RES_p	CB_{pf}	$1 - VARF_p$	DIS_p	CB_{px}
A	0.1556	0.1667	0.2333	0.2444	0.0667	0.2613	0.1724
B	0.3333	0.1667	0.2778	0.1111	0.2889	0.2369	0.1924
C	0.2000	0.1667	0.2147	0.1813	0.2129	0.2027	0.2156

(d) SS_{cp}^* ($MSE_{cp}^* = 0.1575$, with $\langle x \rangle = \mu = 0.25$ and $r = 0.4$)

Method	SS_{cp}^*	$1 - VARX_{cp}^*$	RES_{cp}^*	CB_{cpf}^*	$1 - VARF_{cp}^*$	DIS_{cp}^*	CB_{cpx}^*
A	−0.2063	−0.1905	0.3333	0.3492	−0.3333	0.3733	0.2463
B	0.0476	−0.1905	0.3968	0.1587	−0.0159	0.3384	0.2749
C	−0.1429	−0.1905	0.3067	0.2590	−0.1244	0.2895	0.3079

Key: Part (b)

$VARX_c^* = VARX/MSE_c^*$, $RES_c^* = RES/MSE_c^*$, $CB_{cf}^* = CB_f/MSE_c^*$

$VARF_c^* = VARF/MSE_c^*$, $DIS_c^* = DIS/MSE_c^*$, $CB_{cx}^* = CB_x/MSE_c^*$

Key: Part (c)

$VARX_p = VARX/MSE_p$, $RES_p = RES/MSE_p$, $CB_{pf} = CB_f/MSE_p$

$VARF_p = VARF/MSE_p$, $DIS_p = DIS/MSE_p$, $CB_{px} = CB_x/MSE_p$

Key: Part (d)

$VARX_{cp}^* = VARX/MSE_{cp}^*$, $RES_{cp}^* = RES/MSE_{cp}^*$, $CB_{cpf}^* = CB_f/MSE_{cp}^*$

$VARF_{cp}^* = VARF/MSE_{cp}^*$, $DIS_{cp}^* = DIS/MSE_{cp}^*$, $CB_{cpx}^* = CB_x/MSE_{cp}^*$

scale on which skill is measured. Based on overall skill scores alone, method B would be judged to be superior to methods A and C, with method C being judged superior to method A.

Comparison of the terms in the decompositions of the skill scores associated with the cof factorization indicates that the contributions to skill from the RES (reward) and CB_f (penalty) terms are larger and smaller, respectively, for B's forecasts than for A's and C's forecasts. Thus, these ordinal relationships are consistent with the overall skill scores. However, comparison of these same terms for methods A and C reveals that A's contribution to skill from the RES (reward) term is larger than C's contribution, whereas C's contribution to skill from the CB_f (penalty) term is smaller than A's contribution. Thus, although C's forecasts are more skillful (and less conditionally biased, in the type 1 sense) than A's forecasts, A's forecasts exhibit greater resolution than C's forecasts. As expected, the magnitudes of the RES and CB_f terms increase as the accuracy of the reference forecasts increases [i.e., from SS_p to SS_c^* to SS_{cp}^*; see (37)].

When the terms in the decomposed skill scores associated with the cox factorization are examined, the

picture becomes somewhat more complicated. As noted previously, all three terms in these decompositions vary from forecasting method to forecasting method. In this regard, the contribution to skill associated with the VARF term is larger for method B than for methods A and C. However, the DIS (reward) term and the CB_x (penalty) term are larger and smaller, respectively, for method A than for method B. Thus, according to the terms in the cox decomposition, method A's forecasts are more discriminatory and less conditionally biased (in the type 2 sense) than method B's forecasts. On the other hand, comparison of methods B and C on the basis of these same terms indicates that B's forecasts possess better scores than C's forecasts for all three aspects—or dimensions—of forecasting performance. These ordinal relationships among the forecasting methods hold for all skill scores (and terms in the respective decompositions); that is, the scaling associated with the different reference forecasts does not alter the relative position of the methods on the various dimensions of forecasting performance.

7. Discussion

Some issues related to the use of the general decompositions of MSE-based skill scores in forecast-verification studies are briefly discussed in this section. Attention is focused on two issues: (a) The relative merits of these decompositions and traditional measures of overall forecasting performance in the process/practice of forecast verification. (b) The relationship between relative performance as determined by measures of one or more aspects of quality and unambiguous superiority as determined by the sufficiency relation. To facilitate this discussion, and to enhance its specificity, various measures of aspects of forecasting performance have been computed for the three hypothetical samples of data introduced in section 6.

a. Skill-score decompositions vis-a-vis traditional measures

The numerical values of several traditional measures of overall forecasting performance in 2×2 verification problems are reported in Table 9 for the hypothetical data samples considered in section 6. These measures include the fraction correct (FC), the critical success index (CSI), the Heidke skill score (HSS), and the Hanssen–Kuipers index (HKI) (see Wilks 1995, 238–250). According to the values of FC, CSI, and HSS, forecasting method B is superior to forecasting methods A and C (note that FC = 1 − MSE in this 2×2 problem). On the other hand, the values of HKI indicate that A's forecasts are superior to B's and C's forecasts. Moreover, method A is inferior to the other two methods according to FC, whereas method C is inferior to the other two methods according to CSI, HSS, and HKI.

Two points are worthy of note here. First, FC, CSI, HSS, and HKI are all measures of accuracy in an absolute or relative sense. Specifically, FC and CSI measure absolute accuracy (CSI measures accuracy over the subsample of the verification data sample for which $F = 1$ and/or $X = 1$), and HSS and HKI are measures of relative accuracy (i.e., skill). Second, the results are in conflict, in the sense that the relative performance of the three forecasting methods varies among the measures. Only CSI and HSS rank the forecasting methods in the same order. Evidently, focusing attention on a single aspect of quality—in this case, absolute and relative accuracy—is no guarantee that all ambiguity regarding relative performance can be avoided.

The terms in the skill-score decompositions relate to several different aspects or dimensions of forecasting performance (as well as the characteristics of the forecasting situations), thereby providing information that is not accessible when the process/practice of verification is restricted to traditional measures. Although the terms in these decompositions may indicate that (for example) A is superior to C on one dimension and C is superior to A on another dimension (see section 6), it is important to recognize that these results are *not* in conflict because different dimensions of quality are involved. Since forecast quality is multifaceted, an approach to verification problems involving individual measures of multiple aspects of quality is clearly more informative than an approach involving multiple measures of a single aspect of quality.

The benefits of assessing multiple dimensions of forecasting performance are at least twofold. 1) As noted previously, such an approach is consistent with the multifaceted nature of forecast quality. In particular, it is less likely than the traditional approach to yield seriously incomplete or misleading results. 2) Deficiencies in some dimensions of quality may be more amenable to reduction or elimination than deficiencies in other dimensions of quality. Conditional biases—the aspects of forecasting performance measured by the terms CB_f and CB_x in the skill-score decompositions—represent a case in point. By definition, such biases (i.e., average errors that occur over subsamples of verification data samples) should be relatively easy to identify. Moreover, measures of conditional biases should provide users of the output of verification systems with reliable—and potentially insightful—information on which to base efforts to improve forecasting performance.

The perceived disadvantages of employing multiple measures (e.g., the terms in the skill-score decompositions) in comparative verification include the added computational burden and the increased likelihood of conflicting or ambiguous results. In view of the remarkable increase in recent years in the ease and speed with which calculations of this type can be made, the burden imposed by computing multiple verification measures is no longer a serious issue. With regard to

TABLE 9. Other measures of aspects of forecast quality for forecasting methods A, B, and C. See text for additional details.

Method	FC	CSI	HSS	HKI	RK1	RK0	POD	FAR	BR
A	0.81	0.4865	0.5250	0.560	0.600	0.100	0.7200	0.4000	1.20
B	0.85	0.5000	0.5714	0.533	0.750	0.125	0.6000	0.2500	0.80
C	0.82	0.4545	0.5068	0.493	0.720	0.160	0.6000	0.3478	0.92

Key:
Fraction correct: $FC = p_{11} + p_{00}$
Critical success index: $CSI = p_{11}/(p_{11} + p_{10} + p_{01})$
Heidke skill score: $HSS = (FC - FC_c)/(1 - FC_c)$, where $FC_c = p_1(f)p_1(x) + p_0(f)p_0(x)$
Hanssen–Kuipers index: $HKI = (p_{11}p_{00} - p_{10}p_{01})/p_1(x)p_0(x)$
Risk 1: $RK1 = p_{11}(f) = p_{11}/p_1(f)$
Risk 0: $RK0 = p_{01}(f) = p_{01}/p_0(f)$
Probability of detection: $POD = p_{11}(x) = p_{11}/p_1(x)$
False-alarm ratio: $FAR = p_{10}(f) = p_{10}/p_1(f)$
Bias ratio: $BR = p_1(f)/p_1(x)$

perceptions of conflicting results, recall that the various terms in the skill-score decompositions are concerned with *different* aspects of quality. Thus, the fact that method S is superior to method T according to one term and method T is superior to method S according to another term is *not* indicative of a conflict. It is simply a reflection of differences in relative performance across different dimensions of quality. The availability of a set of measures that can provide information of this type should be viewed as an advantage—rather than a disadvantage—of the verification methodology introduced here.

b. *Measures of aspects of quality and the sufficiency relation*

The numerical results presented in section 6 (terms in skill-score decompositions) and section 7a (traditional measures) raise the following basic question: what conditions must be satisfied in 2 × 2 verification problems for the forecasts produced by forecasting method S to be judged *unambiguously* superior (or inferior) to the forecasts produced by forecasting method T? To address this question, it is first necessary to define the concept of unambiguous superiority (inferiority). Given such a definition, other questions arise. For example, what inferences regarding the relative magnitudes of measures of one or more aspects of forecast quality can be drawn from knowledge that S's forecasts are unambiguously superior to T's forecasts? Is it possible to infer unambiguous superiority from the relative magnitudes of measures of one or more aspects of forecast quality? The purpose of this discussion is simply to introduce the reader to some of the issues involved in any attempt to answer such questions in 2 × 2 problems—the simplest possible verification problems. An in-depth treatment of these issues—or the consideration of such issues in general $k \times k$ ($k \geq 2$) verification problems—are beyond the scope of the present paper.

The concept of unambiguous superiority considered here is embodied in the *sufficiency relation,* first described and applied in a meteorological context by Ehrendorfer and Murphy (1988). This relation identifies the conditions that must be satisfied for S's forecasts to be superior in all respects to T's forecasts. The phrase "superior in all respects" should be understood to mean that if (for example) S's forecasts are sufficient for T's forecasts, then T's forecasts contain greater uncertainty than S's forecasts and *all* potential users of the forecasts would prefer to base their decisions on S's forecasts rather than on T's forecasts (since their welfare based on S's forecasts would equal or exceed their welfare based on T's forecasts).

It is relatively easy to identify the conditions that must be satisfied for the sufficiency relation to hold in 2 × 2 problems. For example, the measures risk 1 (RK1) and risk 0 (RK0)—defined in Table 9—can be used to determine whether or not unambiguous superiority holds in such problems. These quantities are conditional probabilities derived from an augmented 2 × 2 contingency table.

In 2 × 2 problems, method S is sufficient for method T if $RK1(S) \geq RK1(T)$ *and* $RK0(S) \leq RK0(T)$. If these conditions (or their converse) are not satisfied, then methods S and T are said to be insufficient for each other. In this latter case, some users would prefer S's forecasts to T's forecasts and other users would prefer T's forecasts to S's forecasts.

Comparison of forecasting methods A, B, and C using these risk measures (see columns 6 and 7 of Table 9) reveals that method B is sufficient for method C. Thus, all users should prefer B's forecasts to C's forecasts. On the other hand, methods A and B and methods A and C are insufficient for each other, implying that some users would prefer A to B (or A to C) and other users would prefer B to A (or C to A).

The conditions for sufficiency can also be defined in terms of other conditional probabilities derived from augmented 2 × 2 contingency tables. In this regard, the probability of detection (POD) and the false-alarm ratio (FAR), which form the basis of the signal detection theory approach to forecast verification (see Mason

1982), can be used for this purpose. Specifically, method S is sufficient for method T if $POD(S) \geq POD(T)$ *and* $FAR(S) \leq FAR(T)$. The values of POD and FAR for the forecasts produced by methods A, B, and C are included as columns 8 and 9 in Table 9. Use of these measures as a basis for determining sufficiency leads to the same results as those that were obtained using the risk measures RK1 and RK0.

Given that forecasting method S is sufficient for forecasting method T, what can be inferred about the relative performance of these methods according to measures of various dimensions of quality? Under this condition, it seems reasonable to conclude that such measures would indicate that S's forecasts were superior to T's forecasts across every dimension of quality. This conclusion is apparently supported in the case of forecasting methods B and C (recall that B's forecasts are sufficient for C's forecasts) by the results presented in Tables 8 and 9. Specifically, the expected ordinal relationships hold for the four traditional measures of accuracy or skill (Table 9, columns 2–5), the measures MSE and SS (Table 8, column 2), and the terms in the various decompositions of MSE and SS (Table 8, columns 3–8). On the other hand, such ordinal relationships do not hold for all of these measures when analogous pairwise comparisons of measures and terms are made for forecasting methods A and B or for forecasting methods A and C (recall that methods A and B and methods A and C are insufficient for each other). For example, $CB_f(B) < CB_f(A)$ and $RES(B) > RES(A)$, but $CB_x(A) < CB_x(B)$ and $DIS(A) > DIS(B)$ (see Table 8).

Notwithstanding its apparent reasonableness, the above-mentioned conclusion may *not* be warranted for all dimensions of quality. The bias ratio (BR), frequently used to measure unconditional (or systematic) bias in 2×2 problems, provides a case in point. The values of this measure for the three sets of hypothetical forecasts are included in the tenth column of Table 9. Note that despite the fact that B's forecasts are sufficient for C's forecasts, BR(C) is closer than BR(B) to the ideal value of $BR = 1$ (i.e., unconditionally unbiased forecasts).

Another question that arises in this context relates to what can be inferred concerning unambiguous superiority from knowledge of the relative magnitudes of measures of one or more aspects of forecast quality. It is already clear that $HSS(S) > HSS(T)$, for example, does not imply that method S is sufficient for method T. From such an ordinal relationship on the values of HSS, it would be more appropriate to conclude that method T is *not* sufficient for method S. However, in view of the fact that method B is sufficient for method C and yet BR(C) is closer than BR(B) to the ideal value of $BR = 1$, care obviously must be exercised in drawing even such "negative" inferences.

Likewise, superiority in terms of multiple aspects of forecasting performance may not in itself ensure superiority over all potentially relevant dimensions of forecast quality. Thus, despite the fact that comparative verification based on the terms in the skill-score decompositions explicitly considers several aspects of performance, this approach may not be adequate to determine sufficiency (or insufficiency). For example, the fact that $CB_f(S) < CB_f(T)$ and $RES(S) > RES(T)$ does not guarantee that method S is sufficient for method T. At most, all that can be legitimately inferred from these ordinal relationships (regarding sufficiency) is that method T is *not* sufficient for method S.

Since it is relatively easy to determine whether or not sufficiency holds in 2×2 verification problems, it might be argued that the use of a set of measures of various aspects of quality—such as the terms in the skill-score decompositions—is unnecessary in such problems. However, this argument overlooks the essential fact that such measures may provide insights into basic aspects of forecast quality that can help to guide efforts to improve forecasting performance. Moreover, sufficiency is much more difficult to assess in situations involving forecasts and observations for multiple (i.e., three or more) categories and/or in situations involving probabilistic forecasts, because of the substantially greater dimensionality of verification problems in these situations (e.g., see Ehrendorfer and Murphy 1988, 1992). In these contexts comparative verification based on a set of measures of various aspects of forecasting performance appears to offer important advantages over comparative verification based solely on one or two overall measures of accuracy or skill. As the verification data samples considered here have illustrated, inferences regarding superiority or inferiority based solely on a skill score (or any other one-dimensional measure) would be particularly problematic in these situations. The use of a set of measures of basic aspects of forecasting performance, such as the terms in the skill-score decompositions described here, represents a reasonable compromise betweeen the traditional approach that generally overlooks the multidimensional structure of verification problems and an approach based on the sufficiency relation that may be difficult (if not impossible) to apply in practice. In essence, it is clearly more appropriate—and less likely to lead to erroneous conclusions—to judge the relative merits of the forecasts produced by two (or more) forecasting methods on the basis of a comparison across a set of measures of different aspects of performance than on the basis of a single measure of overall accuracy or skill.

8. Summary and conclusions

Two general types of decompositions of skill scores, in which the underlying measure of accuracy is the MSE, have been described in this paper. One type of decomposition is based on conditioning on the forecasts, whereas the other type of decomposition is based

on conditioning the observations. These general decompositions were then applied to skill scores in which climatology, persistence, or a linear combination of climatology and persistence served as the underlying standard of reference. Each of the six specific skill-score decompositions considered here contains measures of basic statistical characteristics of the forecasts and/or the observations, as well as measures of aspects of the relationship between the forecasts and observations.

In summarizing the results presented here—and their implications for the practice of forecast verification—it is useful to distinguish between absolute verification and comparative verification. In the context of absolute verification, these decompositions provide quantitative measures of basic characteristics of the forecasts, the observations, and their relationship. Thus, forecast verification based on skill-score decompositions explicitly recognizes the multidimensional structure of verification problems. On the other hand, the traditional approach to verification problems usually places primary if not exclusive emphasis on assessing overall accuracy or skill, thereby largely overlooking the multifaceted nature of forecast quality. Moreover, it should be noted that the terms in the skill-score decompositions also can be interpreted as positive or negative contributions to skill.

In the context of comparative verification, in which emphasis is naturally placed on assessing relative forecasting performance, an approach to verification problems based on skill-score decompositions also appears to offer important advantages over the traditional approach. As the analysis of the verification data samples considered here indicates, superiority in terms of (for example) skill is no guarantor either of unambiguous superiority (i.e., sufficiency) or of superiority with respect to individual aspects of quality such as reliability, resolution, or discrimination. Thus, to judge relative performance in a coherent manner—and to provide a sound basis for choosing among alternative forecasting methods—it is necessary to examine multiple aspects of quality as well as quantitative measures of these characteristics of forecasting performance. Although superiority over a relatively small set of aspects of performance, such as might be provided by the terms in the skill-score decompositions, is no guarantor of sufficiency, this approach is clearly to be preferred to an approach based solely on superiority with respect to overall skill. In this sense, an approach to comparative verification based on skill-score decompositions represents a reasonably sound compromise between the traditional approach, which is relatively easy to apply but may produce misleading results, and an approach based on the sufficiency relation, which provides a conceptually sound basis for drawing conclusions but may be extremely difficult to apply in practice.

It is relatively easy to show that traditional practices in forecast verification are fundamentally incommensurate with both the multidimensional structure of verification problems and the multifaceted nature of forecast quality. In effect, traditional practices overlook basic aspects of forecast quality. The "negative" implications of following traditional practices are at at least twofold. First, the information that can be gleaned from the process of absolute verification is limited, thereby undermining efforts to properly assess and subsequently improve forecasting performance. Second, the information obtained from the process of comparative verification may be misleading, thereby adversely affecting efforts to identify and implement forecasting methods whose forecasts possess desirable properties. The skill-score decompositions introduced and applied in this paper are proposed as a practical—and potentially useful—solution to these problems. The terms in these decompositions include quantitative measures of basic aspects of forecast quality, such as reliability, resolution, and discrimination. Knowledge of the extent to which the forecasts of interest do or do not possess these characteristics should be useful in the processes of assessing, comparing, and improving forecasting performance.

Acknowledgments. This work was initiated during the first half of 1994, at which time the author was a visiting scientist at the Max Planck Institute for Meteorology in Hamburg, Germany. Two anonymous reviewers provided helpful comments on an earlier version of the manuscript.

APPENDIX

MSE for Combined Climatological–Persistence Forecasts

The combined climatological–persistence forecasts are based on an optimal linear combination of climatological forecasts ($f = \mu$) and persistence forecasts ($f = x_0$). Let the combined forecast be denoted by $f = hx_0 + (1 - h)\mu$, where h is a constant ($0 \leq h \leq 1$). If the mean square error of the combined forecasts is denoted by MSE_{cp}, then

$$\mathrm{MSE}_{cp} = \sum_f \sum_x p(f, x)[hx_0 + (1 - h)\mu - x]^2, \quad \text{(A1)}$$

in which the argument f in the joint distribution and summation involves only the persistence forecasts $f = x_0$ (the climatological forecasts are constant). The value of h that minimizes MSE_{cp}—and defines the optimal linear combination of x_0 and μ—can be found by taking the partial derivative of MSE_{cp} with respect to h and setting the resulting expression equal to zero. Under the assumption of negligible end effects (see section 2b), it can be shown that the optimal value of h is k, where $k = (r + d^2)(1 + d^2)^{-1}$, in which $d^2 = [(\mu - \langle x \rangle)/s_x]^2$. Substitution of k into (A1) yields the following expression for the MSE of the combined forecasts:

$$\mathrm{MSE}_{cp} = [(1 - k)^2(1 + d^2) + 2k(1 - r)]s_x^2. \quad \text{(A2)}$$

REFERENCES

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

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

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

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12,** 595–600.

——, 1988: Skill scores based on the mean square error and their relationship to the correlation coefficient. *Mon. Wea. Rev.*, **116,** 2417–2424.

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

——, 1992: Climatology, persistence, and their linear combination as standards of reference in skill scores. *Wea. Forecasting,* **7,** 692–698.

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

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

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117,** 572–581.

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

Rice, J. A., 1988: *Mathematical Statistics and Data Analysis.* Wadsworth and Brooks, 595 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 464 pp.