Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis, Oregon

(Manuscript received 1 February 1988, in final form 11 April 1988)

ABSTRACT

Several skill scores are defined, based on the mean-square-error measure of accuracy and alternative climatological standards of reference. Decompositions of these skill scores are formulated, each of which is shown to possess terms involving 1) the coefficient of correlation between the forecasts and observations, 2) a measure of the nonsystematic (i.e., conditional) bias in the forecasts, and 3) a measure of the systematic (i.e., unconditional) bias in the forecasts. Depending on the choice of standard of reference, a particular decomposition may also contain terms relating to the degree of association between the reference forecasts and the observations. These decompositions yield analytical relationships between the respective skill scores and the correlation coefficient, document fundamental deficiencies in the correlation coefficient as a measure of performance, and provide additional insight into basic characteristics of forecasting performance. Samples of operational precipitation probability and minimum temperature forecasts are used to investigate the typical magnitudes of the terms in the decompositions. Some implications of the results for the practice of forecast verification are discussed.

1. Introduction

Skill scores are generally defined as measures of the relative accuracy of forecasts produced by two forecasting systems, one of which is a "reference system" (e.g., see Murphy and Daan 1985). Positive skill (i.e., a favorable difference in accuracy) is usually considered to represent a minimal level of acceptable performance for a set of forecasts. To the extent that the difficulty inherent in forecasting situations is reflected in the level of accuracy of the reference forecasts, skill scores also take difficulty into account. As a result, they can be used (with appropriate caveats) to compare forecasting performance across different locations or time periods. Thus, it is not surprising that skill scores are widely used in evaluating the performance of operational and experimental forecasts (e.g., see Dagostaro et al. 1988; Murphy and Daan 1985).

In the context of forecast verification, correlation coefficients are measures of the degree of linear association between the forecasts of interest and the relevant observations. It has been recognized for many years (e.g., see Brier and Allen 1951) that correlation coefficients suffer from certain deficiencies as verification measures. Nevertheless, they *are* employed from time to time in forecast verification programs and several different correlation coefficients are currently used in conjunction with model verification studies (e.g., see Arpe et al. 1985; Miyakoda et al. 1972; Sanders 1987).

Despite the rather widespread use of both skill scores and correlation coefficients, the relationships between these two common types of verification measures have evidently not been explored. In addition, little if any attention has been devoted to the problem of obtaining a quantitative appreciation of the deficiencies in the correlation coefficient as a measure of forecasting performance. The primary purpose of this paper is to describe decompositions of a family of climatological skill scores that yield insight into (i) the relationships between these measures and the (product moment) correlation coefficient and (ii) the deficiencies in the latter as a performance measure.

In section 2, we define the terms "accuracy" and "skill" and identify the basic measures of these attributes—namely, the mean-square-error measure of accuracy and the mean-square-error skill score—employed in this paper. This section also describes alternative climatological standards of reference and defines a mean-square-error skill score based on each of the reference procedures. Decompositions of the skill scores are formulated in section 3, and these decompositions—and the skill score/correlation coefficient relationships—are discussed and interpreted in section 4. Operational precipitation probability and minimum temperature forecasts are used to investigate the typical magnitudes of the terms in these decompositions in section 5. Section 6 consists of a brief summary and some discussion of the implications of these results for the practice of forecast verification.

Corresponding author address: Prof. Allan H. Murphy, Dept. of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331-2209.

2. Mean-square-error skill scores

a. Basic measures of accuracy and skill

Accuracy is usually defined as the average degree of correspondence between individual forecasts and observations (e.g., see Murphy and Daan 1985). Thus, the mean absolute error represents a prototypical measure of accuracy. Skill, on the other hand, is generally defined as the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by some reference procedure—or *standard of reference*—such as climatology or persistence. To measure skill, we might compute the improvement in the mean absolute error of the forecasts over the mean absolute error of climatological forecasts [this improvement is usually compared with the total possible improvement—see (2) below].

The basic measure of accuracy employed in this paper is the mean square error (MSE). Consider a sample of n (matched) forecasts and observations, in which f_i and x_i denote the ith forecast and ith observation, respectively. Then, MSE(f, x) can be expressed as follows:

$$\mathrm{MSE}(f, x) = \frac{1}{n} \sum_{i=1}^{n} (f_i - x_i)^2. \quad (1)$$

Note that MSE$(f, x) \geqslant 0$, with equality only for perfect forecasts ($f_i = x_i$ for all i).

With regard to skill, it is reasonable—and traditional—practice to define a generic skill score S in terms of generic measure of accuracy A in the following manner:

$$S = (A_f - A_r)/(A_p - A_r), \quad (2)$$

where A_f, A_p, and A_r denote the accuracy of the forecasts of interest, the accuracy of perfect forecasts, and the accuracy of the reference forecasts, respectively (see Murphy and Daan 1985). Note that S represents the improvement in accuracy of the forecasts over the reference forecasts relative to the total possible improvement in accuracy.

In view of the definition of the generic skill score in (2), a skill score SS based on the mean-square-error measure of accuracy can be expressed as follows:

$$\mathrm{SS}(f, r, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(r, x)], \quad (3)$$

since MSE$(p, x) = 0$. Note that SS in (3) is a function of the forecasts of interest (f), the reference forecasts (r), and the observations (x). The skill score SS is positive (negative) when the accuracy of the forecasts is greater (less) than the accuracy of the reference forecasts. Moreover, SS = 1 when MSE$(f, x) = 0$ (perfect forecasts) and SS = 0 when MSE$(f, x) = $ MSE(r, x). It can be translated into a measure of percentage improvement in accuracy simply by multiplying the right-hand side (rhs) of (3) by 100.

b. Climatological standards of reference

As noted previously, climatology is taken to be the standard of reference in this paper. Thus, the reference forecasts are assumed to be based solely on a relevant set of *observations* of the variable or event of interest. Several alternative definitions of these climatological reference forecasts are possible, depending on the particular set of observations employed and the way in which the observations are used to create the forecasts. First, the climatological forecasts could be based on observations from some historical period or they could (at least, hypothetically) be based on the sample of observations from the experimental period. We will refer to reference forecasts derived from historical and experimental periods as *external* climatological forecasts and *internal* climatological forecasts, respectively (in the sense that the respective climatologies are external and internal to the matched sample of n forecasts and observations). External climatology is generally called long-term or historical climatology, whereas internal climatology is usually referred to as short-term or sample climatology. For a discussion of the relative merits of using internal and external climatologies as standards of reference in formulating skill scores, see Murphy (1974).

Second, the reference forecasts could consist of a single constant forecast applicable to all forecasting occasions (and based on the entire sample of observations) or they could consist of different forecasts for different occasions (based on appropriate subsamples of the observations). In this regard, for example, it might be argued that the use of a single climatological reference forecast would be inappropriate when the forecasts of interest are to be evaluated over time periods in excess of a month or season. We will refer to such forecasts as *single-valued* and *multiple-valued* climatological forecasts.

These considerations lead to the identification of four types of climatological reference forecasts: 1) single-valued internal reference forecasts; 2) multiple-valued internal reference forecasts; 3) single-valued external reference forecasts; and 4) multiple-valued external reference forecasts. For convenience, we will sometimes refer to situations involving the use of these various reference forecasts as cases I, II, III, and IV, respectively.

c. Mean-square-error skill scores based on climatology

To define skill scores based on the climatological standards of reference described in section 2b, it is necessary to introduce some additional notation. In the case of internal climatology, let x_i^* denote the sample climatology relevant to the ith matched pair of forecasts and observations and let $\bar{x}$, where $\bar{x} = (1/n) \sum_{i=1,n} x_i$, denote the mean sample climatology. (An asterisk has been included to distinguish between the ith observation in the sample and the corresponding internal

climatological standard of reference. The latter is presumably based on a *subsample* of the n observations in the sample. As a result, it is reasonable to assume that $\bar{x}^* = \bar{x}$.) Then $\bar{x}$ represents the single-valued internal climatological standard of reference (Case I) and the x_i^* ($i = 1, \cdots, n$) represent the multiple-valued internal climatological standard of reference (Case II).

With regard to external climatology, let μ_i denote the long-term climatology relevant to the ith pair of forecasts and observations and let $\bar{\mu}$, where $\bar{\mu} = (1/n) \sum_{i=1,n} \mu_i$, denote the mean long-term climatology. Then $\bar{\mu}$ represents the single-valued external climatological standard of reference (Case III) and the $\mu_i (i = 1, \cdots, n)$ represent the multiple-valued external climatological standard of reference (Case IV). It should be noted that the μ_i associated with various subsamples of forecasts may be identical. For example, in evaluating the skill of day-to-day forecasts produced over a one-year period, it might be appropriate to use long-term *monthly* climatological values as the standard of reference.

Now we can define skill scores based on the four possible types of climatological reference forecasts. In the case of single-valued internal climatology (Case I), $\mathrm{SS}(f, r, x)$ becomes $\mathrm{SS}(f, \bar{x}, x)$, where

$$\mathrm{SS}(f, \bar{x}, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(\bar{x}, x)]. \quad (4)$$

For multiple-valued internal climatology (Case II), $\mathrm{SS}(f, r, x)$ becomes $\mathrm{SS}(f, x^*, x)$, where

$$\mathrm{SS}(f, x^*, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(x^*, x)]. \quad (5)$$

In the case of single-valued external climatology (Case III), $\mathrm{SS}(f, r, x)$ becomes $\mathrm{SS}(f, \bar{\mu}, x)$, where

$$\mathrm{SS}(f, \bar{\mu}, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(\bar{\mu}, x)]. \quad (6)$$

For multiple-valued external climatology (Case IV), $\mathrm{SS}(f, r, x)$ becomes $\mathrm{SS}(f, \mu, x)$, where

$$\mathrm{SS}(f, \mu, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(\mu, x)]. \quad (7)$$

3. Decompositions of skill scores

a. Decomposition of mean-square-error measure of accuracy

To facilitate the decomposition of the various skill scores, we first decompose $\mathrm{MSE}(f, x)$. The decomposition of MSE of interest here can be derived by adding and subtracting the mean forecast $\bar{f}$ [$=(1/n) \times \sum_{i=1,n} f_i$] and the mean observation $\bar{x}$ [$=(1/n) \times \sum_{i=1,n} x_i$] within the parentheses on the rhs of (1). Initially, we obtain

$$\mathrm{MSE}(f, x) = \frac{1}{n} \sum_{i=1}^{n} [(f_i - \bar{f}) - (x_i - \bar{x}) + (\bar{f} - \bar{x})]^2. \quad (8)$$

Completing the squaring process within the brackets on the rhs of (8) and averaging over the sample of n forecasts and observations yields

$$\mathrm{MSE}(f, x) = (\bar{f} - \bar{x})^2 + s_f^2 + s_x^2 - 2s_{fx}, \quad (9)$$

where $s_f^2 = (1/n) \sum_{i=1,n} (f_i - \bar{f})^2$ is the sample variance of the forecasts, $s_x^2 = (1/n) \sum_{i=1,n} (x_i - \bar{x})^2$ is the sample variance of the observations, and $s_{fx} = (1/n) \times \sum_{i=1,n} (f_i - \bar{f})(x_i - \bar{x})$ is the sample covariance of the forecasts and observations. [The decomposition of MSE in (9) is similar to a decomposition of the mean probability score described by Yates (1982).] Moreover, since $s_{fx} = s_f s_x r_{fx}$, $\mathrm{MSE}(f, x)$ in (9) can also be expressed as

$$\mathrm{MSE}(f, x) = (\bar{f} - \bar{x})^2 + s_f^2 + s_x^2 - 2s_f s_x r_{fx}, \quad (10)$$

where r_{fx} is the sample (product moment) coefficient of correlation between the forecasts and observations. The decomposition in (10) is the basic expression for $\mathrm{MSE}(f, x)$ employed in this paper.

b. Decompositions of mean-square-error skill scores

Decompositions of the four skill scores can now be obtained by substituting the decomposition of $\mathrm{MSE}(f, x)$ in (10)—and the appropriate expression for $\mathrm{MSE}(r, x)$—into the respective skill score formula. In the case of single-valued internal climatology (Case I, for which $r = \bar{x}$), it is evident that $\mathrm{MSE}(\bar{x}, x) = s_x^2$ [see (10)]. Thus, from (4) and (10), it follows that

$$\mathrm{SS}(f, \bar{x}, x) = 2(s_f/s_x) r_{fx} - (s_f/s_x)^2 - [(\bar{f} - \bar{x})/s_x]^2, \quad (11)$$

or

$$\mathrm{SS}(f, \bar{x}, x) = r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\bar{f} - \bar{x})/s_x]^2. \quad (12)$$

Note that $\mathrm{SS}(f, \bar{x}, x)$ in (12) is expressed as the "sum" of three nonnegative terms. For convenience, we will refer to the terms on the rhs of (12) as IA, IB, and IC, respectively.

In the case of multiple-valued internal climatology (Case II, for which $r = x_i^*$), $\mathrm{MSE}(x^*, x) = s_x^2 + s_{x^*}^2 - 2s_{x^*} s_x r_{x^* x}$ [see (10) and note that $\bar{x}^* = \bar{x}$], where $s_{x^*}^2$ is the variance of the x_i^* and $r_{x^* x}$ is the coefficient of correlation between the x_i^* and the observations. Thus, from (5) and (10),

$$\mathrm{SS}(f, x^*, x) = \{r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\bar{f} - \bar{x})/s_x]^2 - r_{x^* x}^2 + [r_{x^* x} - (s_{x^*}/s_x)]^2\} / \{1 - r_{x^* x}^2 + [r_{x^* x} - (s_{x^*}/s_x)]^2\}. \quad (13)$$

The decomposition of $\mathrm{SS}(f, x^*, x)$ in (13) contains the terms IA, IB, and IC, as well as two terms (in both numerator and denominator) relating to the degree of correspondence between the internal climatological values (i.e., the x_i^*) and the observations. For conve-

nience, we will denote these latter terms by IIA and IIB, respectively.

With regard to single-valued external climatology (Case III, for which $r = \bar{\mu}$), $\mathrm{MSE}(\bar{\mu}, x) = s_x^2 + (\bar{\mu} - \bar{x})^2$ [see (10)] and, from (6) and (10),

$$\mathrm{SS}(f, \bar{\mu}, x) = \{r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\bar{f} - \bar{x})/s_x]^2 + [(\bar{\mu} - \bar{x})/s_x]^2\}/\{1 + [(\bar{\mu} - \bar{x})/s_x]^2\}. \quad (14)$$

In addition to the terms IA, IB, and IC, the decomposition of $\mathrm{SS}(f, \bar{\mu}, x)$ in (14) contains a term (in both numerator and denominator) related to the degree of correspondence between the mean external and internal climatologies. This term will be denoted by IIIA.

Finally, in the case of multiple-valued external climatology (Case IV, for which $r = \mu_i$), $\mathrm{MSE}(\mu, x) = s_x^2 + s_\mu^2 + (\bar{\mu} - \bar{x})^2 - 2s_\mu s_x r_{\mu x}$ [see (10)], where s_μ^2 is the variance of the μ_i and $r_{\mu x}$ is the coefficient of correlation between the μ_i and the observations. Thus, from (7) and (10),

$$\mathrm{SS}(f, \mu, x) = \{r_{fx}^2 - [r_{fx} - (s_f/s_x)]^2 - [(\bar{f} - \bar{x})/s_x]^2 - r_{\mu x}^2 + [r_{\mu x} - (s_\mu/s_x)]^2 + [(\bar{\mu} - \bar{x})/s_x]^2\}/\{1 - r_{\mu x}^2 + [r_{\mu x} - (s_\mu/s_x)]^2 + [(\bar{\mu} - \bar{x})/s_x]^2\}. \quad (15)$$

In addition to the terms IA, IB, and IC, the decomposition of $\mathrm{SS}(f, \mu, x)$ in (15) contains three terms (in both numerator and denominator) relating to the degree of association between the external climatological values (i.e., the μ_i) and the observations. We will denote these three terms by IVA, IVB, and IVC, respectively.

4. Discussion and interpretation

As indicated in section 3b, the decompositions of all four skill scores contain the terms IA, IB, and IC. In fact, the decomposition of $\mathrm{SS}(f, \bar{x}, x)$ (Case I) consists solely of these three terms [see (12)]. The decompositions of the other skill scores include additional terms relating to the degree of association between the reference forecasts and the observations. We focus initially on the three common terms and then discuss the other terms that arise in Cases II, III, and IV.

In developing interpretations for the three common terms, it is useful to recall that all of the non-time-dependent information relevant to verification is contained in the joint distribution of forecasts and observations (Murphy and Winkler 1987). In this regard, the three terms of interest here are defined in terms of summary measures of the (empirical) joint and marginal distributions of forecasts and observations [i.e., in terms of means, variances, and a covariance or correlation—see (12)]. Furthermore, this joint distribution can be described in terms of a linear regression model in which the observations are regressed on the forecasts.

With this model in mind, the term IA—the square of the sample correlation coefficient—obviously represents an overall, nondimensional measure of the strength of the linear relationship between the forecasts and observations. This measure ranges from zero, corresponding to no linear relationship, to one, corresponding to a perfect linear relationship. In general, skill increases as the strength of this linear relationship increases.

The term IB represents the square of the difference between the sample correlation coefficient and the ratio of the standard deviation of the forecasts to the standard deviation of the observations. In this regard, it should be noted that the slope of the regression line, which describes the relationship between the expected values of the observations (given the forecasts) and the forecasts, can be expressed as $(s_x/s_f)r_{fx}$. It can then be seen that the term of interest here vanishes (only) when the slope of the regression line is equal to one, an obviously desirable characteristic of such a line in the context of forecast verification. When the slope of the line is *not* equal to one, it implies that the conditional expected values of the observations are not equal to the corresponding forecasts and, as a result, that the latter are biased. Thus, this term represents a nondimensional measure of the conditional bias in the forecasts. In view of the fact that the term is nonnegative and is preceded by a minus sign, linear relationships between forecasts and observations characterized by slopes that differ from unity will tend to decrease skill.

Finally, the term IC is the square of the difference between the mean forecast and the mean observation, divided by the variance of the observations. It is immediately evident that this term is a nondimensional measure of the unconditional (i.e., overall) bias in the forecasts and that it vanishes only for unbiased forecasts ($\bar{f} = \bar{x}$). Moreover, the term IC is related to the constant (intercept) term in the linear regression model. However, it should be noted that the intercept in the regression model is identically equal to zero only when the forecasts are unconditionally *and* conditionally unbiased. In view of the fact that this term is nonnegative and is preceded by a minus sign, the skill of the forecasts tends to decrease as the unconditional bias increases (and vice versa).

Thus, it is evident that the decomposition of the skill scores into expressions containing the terms IA, IB, and IC yields analytical relationships between the respective skill scores and the square of the correlation coefficient (i.e., r_{fx}^2). The relationship is simplest in the case of single-valued internal climatology (Case I), because only the three common terms are involved. In this case, the overall skill of the forecasts can be seen to consist of three parts: (i) the strength of the linear relationship between the forecasts and observations, as reflected by the square of the correlation coefficient;

(ii) the conditional bias in the forecasts, as reflected by the extent to which the (square of the) slope of the regression line differs from unity; and (iii) the unconditional bias in the forecasts, as reflected by the square of a nondimensional measure of the difference between the mean forecast and the mean observation. These statements apply, as well, to the common terms in the decompositions of the other skill scores, but in these cases additional terms involving the degree of association between the reference forecasts and the observations also influence forecast skill.

Before discussing the other terms that arise in Cases II, III, and IV, the implications of this skill score/correlation coefficient relationship for the correlation coefficient as a measure of forecasting performance are briefly considered. For convenience, we once again refer to the terms IA, IB, and IC in (12). First, it should be noted that this relationship suggests that the square of the correlation coefficient (i.e., Term IA) is itself a kind of skill score. [In the context of the linear regression model, the square of the correlation coefficient represents the fraction of the variance of the observations "explained" (or accounted for) by the forecasts]. Second, the existence of the other terms in the decomposition of $\mathrm{SS}(f, \bar{x}, x)$ (i.e., Terms IB and IC) indicates that, as a skill score, the (square of the) correlation coefficient suffers from two fundamental deficiencies—it ignores both the conditional and unconditional biases in the forecasts. Thus, it is reasonable to consider the square of the correlation coefficient as a measure of *potential* skill (i.e., the level of skill attainable when any conditional and unconditional biases are eliminated), whereas the skill score $\mathrm{SS}(f, \bar{x}, x)$ is a measure of *actual* skill in the case of single-valued internal climatology.

When standards of reference other than single-valued internal climatology are employed, other terms also appear in the decompositions. In the following discussion, we provide interpretations of these additional terms. It should first be noted that these terms bear a strong "pairwise" resemblance to the three common terms, except that the former relate to the association between the *reference* forecasts and the observations. In fact, these additional terms can also generally be interpreted in terms of a linear regression model—in this context, a model in which the observations are regressed on the reference forecasts. The presence or absence of particular (additional) terms depends on the nature of the standard of reference.

For example, in the case of multiple-valued internal climatology (Case II, for which $r = x_i^*$), the reference forecasts are not constant. The accuracy of such *reference* forecasts depends on the variance of the forecasts and the covariance (or correlation) between the forecasts and observations, in addition to the variance of the observations. No term involving the difference between the mean reference forecast and mean observation appears in this (accuracy) expression, because it has been assumed that $\bar{x}^* = \bar{x}$ (see section 2c). As a result, the decomposition of $\mathrm{SS}(f, x^*, x)$ [see (13)] involves two additional terms (IIA and IIB) that appear in both numerator and denominator. These terms vanish only when $2r_{x^*x} = s_{x^*}/s_x$. Moreover, under the assumption that $2r_{x^*x} > s_{x^*}/s_x$ (a reasonable assumption in this context), it can be shown that $\mathrm{SS}(f, x^*, x) < \mathrm{SS}(f, \bar{x}, x)$. That is, the use of multiple-valued internal climatology as a standard of reference instead of single-valued internal climatology will, in general, be associated with a decrease in skill.

In the case of single-valued external climatology (Case III, for which $r = \bar{\mu}$), the reference forecasts are constant (i.e., $s_\mu = 0$) and the regression model fails (infinite slope). The accuracy of such forecasts depends only on the squared difference between the mean external and internal climatologies, in addition to the variance of the observations. As a result, the decomposition of $\mathrm{SS}(f, \bar{\mu}, x)$ [see (14)] involves only a single additional nondimensional term—a term that appears in both numerator and denominator—involving the difference between $\bar{\mu}$ and $\bar{x}$. Note that this term vanishes only when the two climatological means are equal (i.e., $\bar{\mu} = \bar{x}$). Moreover, it is immediately evident that $\mathrm{SS}(f, \bar{\mu}, x) \geq \mathrm{SS}(f, \bar{x}, x)$, with equality only when $\bar{\mu} = \bar{x}$. In effect, use of external climatology "rewards" the forecasting system (in terms of a higher skill score) for any difference between $\bar{x}$ and $\bar{\mu}$, whereas use of internal climatology assumes that this information (i.e., $\bar{x}$) is already available to the forecasting system.

Finally, in the case of multiple-valued external climatology (Case IV, for which $r = \mu_i$), the accuracy of the reference forecasts (which are not constant) depends on the variance of the forecasts, the covariance (or correlation) between the forecasts and observations, and the difference between the mean forecast and mean observation (in addition to the variance of the observations). As a result, the decomposition of $\mathrm{SS}(f, \mu, x)$ [see (15)] involves three additional terms (IVA, IVB, and IVC) that appear in both numerator and denominator. These terms vanish only when $2r_{\mu x} = s_\mu/s_x$ and $\bar{\mu} = \bar{x}$. Moreover, under the assumption that $2r_{\mu x} > s_\mu/s_x$ (once again, a reasonable assumption in this context), it can be shown that $\mathrm{SS}(f, \mu, x) < \mathrm{SS}(f, \bar{\mu}, x)$ for unbiased forecasts. That is, the use of multiple-valued external climatology as a standard of reference instead of single-valued external climatology will also generally be associated with a decrease in skill.

5. Some numerical results

To investigate the typical magnitudes of the terms in the four skill-score decompositions, we examine samples of operational National Weather Service (NWS) forecasts for Portland, Oregon, formulated during the warm season (April–September) for the period 1980–85. The forecasts of interest are the so-called "objective" probability of precipitation (PoP) and

minimum temperature (T_{MIN}) forecasts produced by the numerical-statistical forecasting system (Glahn 1985). The PoP forecasts considered here relate to the 0000 UTC cycle time.

Table 1 contains the numerical results for Case I—that is, SS($f, \bar{x}, x$) and the terms in its decomposition—for various lead times for both the PoP and T_{MIN} forecasts. Here we can examine the magnitudes of the three terms that are common to all four decompositions and compare SS($f, \bar{x}, x$) with the square of the correlation coefficient (r^2_{fx}—Term IA). Table 2 presents the numerical values of all four skill scores—SS($f, \bar{x}, x$) (Case I), SS(f, x^*, x) (Case II), SS($f, \bar{\mu}, x$) (Case III), and SS(f, μ, x) (Case IV)—as well as the value of r^2_{fx}. These results make it possible to compare all of the skill scores with r^2_{fx}, as well as to examine the relative magnitudes of the various skill scores. The numerical values of the terms (other than the three common terms) associated with the decompositions in Cases II, III, and IV are presented in Table 3. In computing all of these quantities, six-month (i.e., warm season) averages were used in the cases of single-valued climatology (Cases I and III) and monthly averages were used in the cases of multiple-valued climatology (Cases II and IV).

In Case I [S($f, \bar{x}, x$)—single-valued internal climatology], the results in Table 1 indicate that the terms relating to the conditional and unconditional biases (Terms IB and IC) in the PoP forecasts are quite small. Thus, SS($f, \bar{x}, x$) is only slightly smaller than r^2_{fx} (Term IA) (at most about 0.010, or 1% when multiplied by 100). On the other hand, very much larger unconditional biases (Term IC) exist in the case of the T_{MIN} forecasts, and the conditional biases (Term IB) for these forecasts are also larger for the 48-hour and 60-hour lead times. As a consequence, SS($f, \bar{x}, x$) is roughly 10–13% smaller than r^2_{fx} for the T_{MIN} forecasts. These results appear to be typical of the magnitudes of the respective quantities for such forecasts at other NWS offices (results omitted to conserve space). In general, the conditional bias term for these forecasts is less than 0.05 and the unconditional bias term ranges from 0.00 to 0.15. It is evident that use of the square of the correlation coefficient as a measure of skill may substantially overestimate actual forecasting performance in some situations.

Comparison of the numerical values of the four skill scores with the value of r^2_{fx} (see Table 2) reveals that the latter always exceeds the former. Of course, these ordinal relationships are strongly "influenced" by the ordinal relationships among the skill scores themselves (see section 4). In this regard, recall that SS($f, \bar{\mu}, x$) $\geq$ SS($f, \bar{x}, x$) and that, in general, SS(f, x^*, x) $<$ SS($f, \bar{x}, x$) and SS(f, μ, x) $<$ SS($f, \bar{\mu}, x$).

TABLE 1. Numerical values of SS($f, \bar{x}, x$) (Case I)—and the terms in its decomposition—for the objective PoP and T_{MIN} forecasts for Portland, Oregon, in the warm season for the period 1980–85.

Lead time (h)	Sample size n	Skill score SS($f, \bar{x}, x$)	Term IA r^2_{fx}	Term IB $[r_{fx} - (s_f/s_x)]^2$	Term IC $[(\bar{f} - \bar{x})/s_x]^2$
			PoP forecasts		
12–24	870	0.347	0.354	0.005	0.002
24–36	868	0.247	0.258	0.002	0.010
36–48	850	0.276	0.281	0.004	0.001
			T_{MIN} forecasts		
24	804	0.672	0.801	0.001	0.128
36	783	0.659	0.768	0.003	0.107
48	785	0.596	0.716	0.013	0.107
60	775	0.573	0.682	0.018	0.091

TABLE 2. Numerical values of the four skill scores and the square of the correlation coefficient for the objective PoP and T_{MIN} forecasts for Portland, Oregon, in the warm season for the period 1980–85.

Lead time (h)	Case I SS($f, \bar{x}, x$)	Case II SS(f, x^*, x)	Case III SS($f, \bar{\mu}, x$)	Case IV SS(f, μ, x)	Term IA r^2_{fx}
			PoP forecasts		
12–24	0.347	0.310	0.347	0.314	0.354
24–36	0.247	0.209	0.247	0.231	0.258
36–48	0.276	0.239	0.276	0.242	0.281
			T_{MIN} forecasts		
24	0.672	0.295	0.739	0.552	0.801
36	0.659	0.260	0.725	0.520	0.768
48	0.596	0.146	0.676	0.451	0.716
60	0.573	0.080	0.653	0.398	0.682

TABLE 3. Numerical values of the terms in decompositions of SS(f, x^*, x) (Case II), SS($f, \bar{\mu}, x$) (Case III), and SS(f, μ, x) (Case IV) for the objective PoP and T_{MIN} forecasts for Portland, Oregon, in the warm season for the period 1980–85.

Lead time (h)	SS(f, x^*, x) Term IIA	SS(f, x^*, x) Term IIB	SS($f, \bar{\mu}, x$) Term IIIA	SS(f, μ, x) Term IVA	SS(f, μ, x) Term IVB	SS(f, μ, x) Term IVC
			PoP forecasts			
12–24	0.054	0.000	0.001	0.049	0.000	0.001
24–36	0.047	0.000	0.000	0.026	0.005	0.000
36–48	0.048	0.000	0.000	0.045	0.000	0.000
			T_{MIN} forecasts			
24	0.535	0.000	0.254	0.530	0.007	0.254
36	0.540	0.000	0.239	0.534	0.006	0.239
48	0.527	0.000	0.248	0.520	0.008	0.248
60	0.536	0.000	0.231	0.530	0.008	0.231

Key:

Term IIA $= r^2_{x^*x}$ Term IVA $= r^2_{\mu x}$

Term IIB $= [r_{x^*x} - (s_{x^*}/s_x)]^2$ Term IVB $= [r_{\mu x} - (s_\mu/s_x)]^2$

Term IIIA $= [(\bar{\mu} - \bar{x})/s_x]^2$ Term IVC $= [(\bar{\mu} - \bar{x})/s_x]^2$

The results in Table 2 are consistent with these relationships. In brief, it can be seen that the differences between the skill scores and r_{fx}^2 in Cases I and III (single-valued climatology) are about 1% or less in the case of the PoP forecasts and roughly 3–13% in the case of the T_{MIN} forecasts (the unconditional biases are considerably larger for the T_{MIN} forecasts than for the PoP forecasts). The differences between the skill scores and r_{fx}^2 in Cases II and IV (multiple-valued climatology) are much larger. In these cases, the climatological standards of reference are based on monthly (as opposed to six-monthly) averages. These standards obviously produce more accurate reference forecasts, with the result that the forecasts of interest appear less skillful. Not surprisingly, this effect is strongest in Case II in which the reference forecasts are based on the sample of observations from the experimental period. In any case, it is evident that, depending on the choice of the standard of reference, the use of the square of the correlation coefficient as a measure of skill overestimates actual skill by amounts that range from a few percent to as much as 50% or more.

The numerical values of the non-common terms arising in Cases II, III, and IV are reported in Table 3. They are included here to "explain" the differences among the skill scores in Table 2. It can be seen that the differences between the mean external and internal climatologies (i.e., $\bar{\mu}$ and $\bar{x}$—see Terms IIIA and IVC) are negligible for the PoP forecasts, but are certainly not insignificant for the T_{MIN} forecasts. The terms that relate to the conditional biases in the reference forecasts (i.e., Terms IIB and IVB) are zero in the case of internal climatology and quite small in the case of external climatology, for both types of forecasts. Finally, the terms representing the square of the coefficient of correlation between the reference forecasts and the observations (i.e., Terms IIA and IVA) are modest (0.025–0.055) for the PoP forecasts and large (0.520–0.540) for the T_{MIN} forecasts. The differences in these values account for the fact that the difference between the respective skill scores and r_{fx}^2 is rather modest for the PoP forecasts but quite large for the T_{MIN} forecasts (see Table 2). It is interesting to note that, as expected, $r_{x^*x}^2$ (Term IIA) $> r_{\mu x}^2$ (Term IVA) for all combinations of variable and lead time. However, the difference between these quantities is less than 0.01 for all but one combination (PoP forecasts for the 24–36 hour lead time).

6. Conclusion

In this paper we have formulated skill scores based on the familiar mean-square-error measure of accuracy and four alternative climatological standards of reference. The latter relate to whether the observations on which the reference forecasts are based are taken from the experimental period or from an historical period (internal versus external climatology) and whether these observations are used to produce a single constant forecast valid on all occasions in the experimental period or multiple forecasts each of which is valid on all occasions in a subperiod of the experimental period (single-valued versus multiple-valued climatology). Decompositions of these skill scores were derived, and they were shown to yield analytical relationships between the respective skill scores and the coefficient of correlation between the forecasts and observations.

Specifically, the decompositions contain three common terms as well as other terms relating to the association between the reference forecasts and the observations. Interpretations of these terms were obtained by appealing to a linear regression model in which the observations are regressed on the forecasts. The common terms involve summary measures of the joint and marginal distributions of forecasts and observations and possess the following interpretations: (i) the square of the coefficient of correlation between the forecasts and observations; (ii) a measure of the nonsystematic (i.e., conditional) bias in the forecasts; and (iii) a measure of the systematic (i.e., unconditional) bias in the forecasts. All three terms are nonnegative and their signs are such that the skill score is generally less than or equal to the square of the correlation coefficient. As a result, it is reasonable to view the square of the correlation coefficient as a measure of potential skill (i.e., the level of skill attainable when the biases are eliminated) and the skill score as a measure of actual skill.

Samples of operational NWS precipitation probability and minimum temperature forecasts for Portland, Oregon, were used to investigate the typical magnitudes of the terms in these decompositions. Both bias terms were found to be quite small in the case of the precipitation probability forecasts but somewhat larger in the case of the minimum temperature forecasts. As a result, the difference between potential skill and actual skill, when the standard of reference is single-valued internal climatology, is very small (less than 1%) in the case of the precipitation probability forecasts and modest (3–13%) in the case of the minimum temperature forecasts. Use of multiple-valued climatology (whether internal or external) substantially increases this difference between potential and actual skill. It is evident from these numerical results that the square of the correlation coefficient, as a measure of potential skill, may substantially overestimate actual skill in some situations. In this regard it would be of some interest to investigate the magnitudes of the bias terms—and the differences between potential and actual skill—for forecasts of other variables for which forecasting methodology is less well-developed or forecasting experience is more limited.

Before discussing the implications of these results for the practice of forecast verification, it may be appropriate to point out that the skill scores described in this paper are *general* skill scores in the sense that they are applicable to all types of forecasts. That is, these

scores can be used as measures of skill whether the forecasts are expressed in categorical or probabilistic format. In this regard, $SS(f, \bar{x}, x)$ is identical to the so-called sample skill score based on the Brier score (Brier 1950), a measure frequently used to determine the skill of probabilistic forecasts (e.g., see Murphy 1985). As the results in Table 1 demonstrate, this measure can also be used to assess the skill of categorical temperature forecasts.

What are the implications of the results presented in this paper for current practices in forecast verification? First and foremost, it is evident that great care should be exercised in using the correlation coefficient to evaluate the performance of forecasts. Since the decompositions presented here demonstrate that this measure ignores two important types of biases in forecasts, use of the correlation coefficient (or its square) may lead to substantial overestimation of forecasting performance. In this regard, it is more appropriate to interpret the square of the correlation coefficient as a measure of potential skill than as a measure of actual skill. Second, if the correlation coefficient is of interest in a verification study, then the other terms in the decomposition of the relevant skill score—as well as the skill score itself—also should be computed. These other terms will provide quantitative information regarding the difference between potential and actual skill, as well as additional insight into basic characteristics of forecasting performance. This insight may prove useful in subsequent efforts to enhance the skill of the forecasts.

Acknowledgments. Harald Daan, Edward S. Epstein, Harry R. Glahn, and Ian B. Mason provided valuable comments on earlier versions of this paper. Joong Bae Ahn made the numerical calculations involving the precipitation probability and minimum temperature forecasts. This research was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8714108.

REFERENCES

Arpe, K., A. Hollingsworth, M. S. Tracton, A. C. Lorenc, S. Uppala and P. Kallberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verifications and implications for predictability. *Quart. J. Roy. Meteor. Soc.*, **111**, 67–101.

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

——, and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology*, T. F. Malone, Ed., Amer. Meteor. Soc., 841–848.

Dagostaro, V. J., G. M. Carter and J. P. Dallavalle, 1988: AFOS-era verification of guidance and local aviation/public weather forecasts—No. 8 (April 1987–September 1987). Silver Spring, NOAA, Natl. Wea. Serv., TDL Office Note 88-1, 43 pp.

Glahn, H. R., 1985: Statistical weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 289–335.

Miyakoda, K., G. D. Hembree, R. F. Strickler and I. Shulman, 1972: Cumulative results of extended forecast experiments: I. Model performance for winter cases. *Mon. Wea. Rev.*, **100**, 836–855.

Murphy, A. H., 1974: A sample skill score for probability forecasts. *Mon. Wea. Rev.*, **102**, 48–55.

——, 1985: Probabilistic weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 337–377.

——, 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.

Sanders, F., 1987: Skill of NMC operational dynamical models in prediction of explosive cyclogenesis. *Wea. Forecasting*, **2**, 322–336.

Yates, J. F., 1982: External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, **30**, 132–156.