The Coefficients of Correlation and Determination as Measures of Performance in Forecast Verification

ALLAN H. MURPHY

Prediction and Evaluation Systems, Corvallis, Oregon

(Manuscript received 24 October 1994, in final form 2 May 1995)

ABSTRACT

This paper is concerned with the use of the coefficient of correlation (CoC) and the coefficient of determination (CoD) as performance measures in forecast verification. Aspects of forecasting performance that are measured—and not measured (i.e., ignored)—by these coefficients are identified. Decompositions of familiar quadratic measures of accuracy and skill are used to explore differences between these quadratic measures and the coefficients of correlation and determination. A linear regression model, in which forecasts are regressed on observations, is introduced to provide insight into the interpretations of the CoC and the CoD in this context.

Issues related to the use of these coefficients as verification measures are discussed, including the deficiencies inherent in one-dimensional measures of overall performance, the pros and cons of quadratic measures of accuracy and skill vis-à-vis the coefficients of correlation and determination, and the relative merits of the CoC and the CoD. These coefficients by themselves do not provide an adequate basis for drawing firm conclusions regarding absolute or relative forecasting performance.

1. Introduction

Forecasts of surface pressure, geopotential height, etc., produced by numerical models, as well as some forecasts of sensible weather variables provided by numerical–statistical models or weather forecasters, are often verified by calculating a coefficient of correlation between the forecasts and the corresponding observations (or analysis) (e.g., Arpe et al. 1985; Jensenius 1990; Kalnay et al. 1990; Miyakoda et al. 1972). Different forms of the correlation coefficient are used in various applications; for example, the anomaly correlation coefficient generally is computed when verifying forecasts produced by numerical models. As a well-known indicator of the strength of the statistical relationship between two variables, it is perhaps not surprising that the correlation coefficient—in one form or another—has been viewed as a natural and reasonable measure of forecasting performance.

Whatever intuitive (or other) appeal correlation coefficients may possess as measures of forecasting performance, Brier and Allen (1951, p. 845) pointed out almost 45 years ago that such measures are "insensitive to any bias or error in scale." In the years since the early 1950s, the relative merits of correlation coefficients and alternative measures of performance such as the mean square error have been discussed on several occasions (e.g., see Barnston 1992; Livezey 1994; Murphy and Epstein 1989). Although correlation coefficients may have a role to play in forecast verification as summary measures of the joint distribution of forecasts and observations (see Murphy and Winkler 1987), their use as measures of performance appears to warrant further scrutiny.

In addition to issues of a methodological nature related to the use of correlation coefficients in forecast verification, issues also arise concerning the interpretation of these coefficients as measures of forecasting performance. For example, it has become common practice in some circles (e.g., Palmer et al. 1990; Tracton 1993) to refer to the aspect of forecasting performance measured by correlation coefficients as "skill." On the other hand, Murphy and Epstein (1989) noted that correlation coefficients are more appropriately viewed (at best) as measures of potential skill rather than actual skill. Moreover, recent efforts to develop and exploit a general framework for forecast verification (see Murphy and Winkler 1987; Murphy et al. 1989) have underlined the importance of distinguishing among—and measuring separately—various basic aspects of forecast quality. Thus, issues such as the aspect of forecasting performance measured by correlation coefficients and its relationship to other aspects of forecast quality appear to require additional clarification.

The overall purpose of this paper is to investigate the proper interpretation and use of correlation coefficients as measures of forecasting performance. In section 2, the coefficients of correlation and determination are defined and the aspects of forecast quality measured—and not measured (i.e., ignored)—by these

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

coefficients are identified. Decompositions of familiar quadratic measures of accuracy and skill are used in section 3 to illustrate basic differences between the coefficients of correlation and determination as measures of linear association and these quadratic measures of performance. A linear model in which forecasts are regressed on observations is introduced in section 4 to provide additional insight into relationships between these measures and coefficients. Several issues related to the use of the coefficients of correlation and determination in forecast verification are discussed in section 5. Section 6 consists of a brief summary and some concluding remarks.

2. Coefficients of correlation and determination

a. Basic definitions

Let the variables G and Y denote the forecasts of interest and the corresponding observations (or analysis), respectively. Moreover, let $[(g_i, y_i); i = 1, \cdots, n]$ denote a sample of n matching forecasts and observations—hereafter referred to as the verification data sample (VDS). The ith pair (g_i, y_i) could represent the forecast and observation on the ith day in a VDS spanning n days (e.g., n sensible weather forecasts for a specific location). Alternatively, the pair (g_i, y_i) could represent the forecast and observation (or analyzed value) for the ith grid point in a VDS defined on n grid points (e.g., a two-dimensional forecast for n grid points on a specific day).

The product–moment coefficient of correlation (CoC) r_{gy} for this matched sample of forecasts and observations can be defined as follows:

$$r_{gy} = \frac{s_{gy}}{s_g s_y}, \tag{1}$$

where

$$s_{gy} = (1/n) \sum_{i=1}^{n} (g_i - \bar{g})(y_i - \bar{y}) \tag{2}$$

is the sample covariance between G and Y,

$$s_g = [(1/n) \sum_i (g_i - \bar{g})^2]^{1/2} \quad (i = 1, \cdots, n) \tag{3}$$

is the sample standard deviation of G, and

$$s_y = [(1/n) \sum_i (y_i - \bar{y})^2]^{1/2} \quad (i = 1, \cdots, n) \tag{4}$$

is the sample standard deviation of Y. The quantities $\bar{g} = (1/n)\sum_i g_i$ and $\bar{y} = (1/n)\sum_i y_i$ in (2)–(4) are the sample means of G and Y, respectively.

As defined in (1), the CoC is a measure of the strength of the linear statistical relationship between G and Y (Neter et al. 1988, 609). Its values range from -1 to $+1$ inclusive, with $|r_{gy}| = 1$ indicating a perfect linear relationship and $r_{gy} = 0$ indicating the absence of a linear relationship. Further discussion of r_{gy} and its interpretation as a verification measure is postponed until section 2b.

In the context of model verification, the anomaly correlation coefficient (ACC) is frequently used to measure the "degree of correspondence" between forecasts and observations (or analysis). The ACC is defined in terms of deviations of forecast and observed (or analyzed) values from mean historical climatological values. Suppose that these deviations (or anomalies) are denoted by $G' = G - \mu$ and $Y' = Y - \mu$, respectively, where μ represents the appropriate climatological mean. Then, replacing g and y in (1)–(4) by g' and y', respectively, the expression in (1) becomes the ACC, denoted here by $r_{g'y'}$. Under the assumption that the ACC is defined in this manner, relationships and results analogous to those described in sections 3 and 4 for r_{gy} (and r_{gy}^2) can also be obtained for $r_{g'y'}$ (and $r_{g'y'}^2$).

The strength of the linear statistical relationship between G and Y can be measured by the square of the CoC as well as by r_{gy} itself. From (1), it follows that

$$r_{gy}^2 = \frac{s_{gy}^2}{s_g^2 s_y^2}, \tag{5}$$

where s_g^2 and s_y^2 are the sample variances of the forecasts and observations, respectively. Note that $0 \leq r_{gy}^2 \leq 1$. The quantity r_{gy}^2 in (5) is generally referred to as the coefficient of determination (CoD) (Neter et al. 1988, pp. 607–609). Further discussion of r_{gy}^2 and its interpretation as a verification measure is postponed until section 2b.

b. Aspects of performance measured by CoC and CoD

The aspect of the relationship between two variables measured by the coefficients of correlation and determination is usually referred to as *association* (or *linear association*) (Neter et al. 1988, pp. 172–173). In the case of the CoC, the values of $+1$ and -1 indicate perfect positive and negative association, respectively, between G and Y. On the other hand, when $r_{gy} = 0$, these variables exhibit no linear association. In the case of the CoD, its values range from zero to one, with the former indicating no linear association and the latter indicating perfect linear association.

In addition to its interpretation as a measure of linear association, the CoD represents the proportion of the (total) variability of Y, s_y^2, accounted for when a linear regression model—that is, a linear equation in G (see section 4a)—is used to predict Y. With regard to the relative magnitudes of the CoD and the CoC, it is of interest to note that $r_{gy}^2 \leq |r_{gy}|$, with equality being reached only when $r_{gy} = 0$ or 1. For further discussion of r_{gy}^2 and r_{gy} as verification measures, see section 3.

When the forecasts of interest consist of two-dimensional fields (e.g., forecasts produced by numerical models), the CoD can be interpreted as a measure of the correspondence in *phase* between the forecasts and observations (under the assumption that $r_{gy} > 0$). This correspondence is relatively strong when large and small values of the variable of interest (e.g., ridges and troughs in a geopotential height field) occur in the same regions of the forecast and observed (or analyzed) fields, and it is relatively weak otherwise. Correspondence in phase is complete only when $r^2_{gy} = 1$ (or $r^2_{g'y'} = 1$, in the case of fields defined in terms of anomalies). Thus, in the context of model verification, the quantity $1 - r^2_{gy}$ (or $1 - r^2_{g'y'}$) can be viewed as a measure of the phase error in the forecasts.

c. *Aspects of performance ignored by CoC and CoD*

It is a well-known result in elementary statistics that the variance of a random variable X is unaffected by the addition or subtraction of a constant (k say). That is, if $X^* = X + k$, then $s^2_{x^*} = s^2_x$. Moreover, multiplication of the variable X by a constant simply multiplies its variance by the square of the constant. That is, if $X^* = kX$, then $s^2_{x^*} = k^2 s^2_x$. Analogous results hold for covariances.

In light of these basic results, it is evident that neither r_{gy} in (1) nor r^2_{gy} in (5) are affected by changes in the origin (addition of a constant) or unit of measurement (multiplication by a constant) of the scale on which G and/or Y are measured. In the case of multiplication by a constant, the square of the constant appears in both numerator and denominator of r_{gy} and thereby cancels. To dramatize this fact, Brier and Allen (1951, 845) noted that the CoC would not be affected if temperature forecasts made using the Fahrenheit scale were verified using the Celsius scale.

As a consequence of these basic relationships, r_{gy} and r^2_{gy} are insensitive to two types of bias that may be present in forecasts: 1) unconditional bias and 2) conditional bias. Unconditional (or systematic) bias relates to the difference between the mean forecast $\bar{g}$ and mean observation $\bar{y}$. Forecasts are unconditionally unbiased if $\bar{g} = \bar{y}$. Conditional bias relates to the difference between the mean observation given a particular forecast $\bar{y}_g$ and the forecast $G = g$. Forecasts are conditionally unbiased for the forecast $G = g$ if $\bar{y}_g = g$, and they are conditionally unbiased for all forecasts if $\bar{y}_g = g$ for all g. Forecasts that are conditionally unbiased for all g are also unconditionally unbiased, but the converse is not true. To reiterate, the presence of conditional and/or unconditional bias in a VDS is *not* reflected in the value of r_{gy} or r^2_{gy}.

The fact that the CoD ignores conditional and unconditional bias, as defined above, implies that r^2_{gy} also ignores correspondence in *amplitude* in the case of forecasts expressed in the form of two-dimensional fields. In this context, correspondence in amplitude is determined by the relative magnitudes of the respective means and standard deviations of the forecasts and observations. Further discussion of this issue is postponed until section 3c.

3. Relationships between aspects of performance

a. *Quadratic measures of accuracy and skill*

Insight into the differences between (linear) association and other aspects of forecasting performance, such as accuracy and skill, can be obtained by examining the relationships between the CoC or the CoD and common measures of these other aspects of forecast quality. Accuracy is defined as the average degree of correspondence between forecasts and observations (i.e., between G and Y) over a VDS, and skill is defined as the accuracy of the forecasts of interest relative to the accuracy of forecasts based on a standard of reference such as climatology or persistence (e.g., Wilks 1995, pp. 236–237).

Here, we explore relationships between r_{gy} and the mean square error and between r^2_{gy} and a skill score based on the mean square error. The mean square error of the forecasts in the VDS is defined as

$$\mathrm{MSE}(g, y) = (1/n) \sum_i (g_i - y_i)^2 \quad (i = 1, \cdots, n). \tag{6}$$

MSE(g, y) in (6), denoted hereafter simply as MSE, is a quadratic measure of the accuracy of the forecasts. Note that MSE $\geqslant 0$, with equality indicating perfect accuracy (i.e., $g_i = y_i$ for all i).

The skill score based on the MSE is defined as

$$SS(g, r, y) = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}(r, y)}, \tag{7}$$

in which MSE(r, y) is the mean square error of the reference forecasts r_i ($i = 1, \cdots, n$). Under the assumption that the reference forecasts represent constant forecasts of the sample mean of the observations (i.e., $r_i = \bar{y}$ for all i), MSE(r, y) $= s^2_y$ [see (6)]. Thus, the standard of reference here is sample climatology. The $SS(g, r, y)$ in (7), denoted hereafter simply as SS, is a measure of skill (i.e., relative accuracy). Note that $SS \leqslant 1$, with $SS = 1$ indicating perfect skill (i.e., MSE $= 0$) and $SS = 0$ indicating no skill [i.e., MSE $=$ MSE(r, y)].

b. *Association vis-à-vis accuracy*

In the case of the MSE in (6), the following basic decomposition can be formulated

$$\mathrm{MSE} = (\bar{g} - \bar{y})^2 + s^2_g + s^2_y - 2 s_g s_y r_{gy} \tag{8}$$

(e.g., Murphy 1988, p. 2419). By adding and subtracting the quantity $r^2_{gy} s^2_y$ on the rhs of (8), the following useful expression can be obtained (after rearranging and combining terms):

$$\mathrm{MSE} = (\bar{g} - \bar{y})^2 + (s_g - r_{gy}s_y)^2 + (1 - r_{gy}^2)s_y^2. \quad (9)$$

The decomposition in (9) defines an analytical relationship between the MSE, a measure of accuracy, and r_{gy}, a measure of association. Examination of (9) reveals that perfect linear association (i.e., $r_{gy} = 1$) is *not* equivalent to perfect accuracy (i.e., MSE = 0). When $r_{gy} = 1$, the first and second terms on the rhs of (9) remain; namely, $(\bar{g} - \bar{y})^2$ and $(s_g - s_y)^2$. Perfect accuracy is achieved only when these two terms vanish (i.e., when $\bar{g} = \bar{y}$ and $s_g = s_y$). The third term on the rhs of (9) represents the fraction of the variance of the observations (s_y^2) that is not accounted for by a simple linear model in the forecasts (see section 4).

The expression for the MSE when $r_{gy} = 1$, as well as analogous expressions for the MSE when $r_{gy} = 0$ and $r_{gy} = s_g/s_y$, are given in Table 1 (since $r_{gy} \geqslant 0$ in most verification problems of interest in the real world, negative values of the CoC are not considered). The value of the CoC when $r_{gy} = s_g/s_y$ is of interest because it defines the value of s_g ($= r_{gy}s_y$) that minimizes the MSE under the condition that the other statistics in (9) are fixed (see section 4b).

c. *Association vis-à-vis skill*

In the case of the *SS* in (7), the following decomposition can be formulated:

$$SS = r_{gy}^2 - [r_{gy} - (s_g/s_y)]^2 - [(\bar{g} - \bar{y})/s_y]^2 \quad (10)$$

(Murphy 1988, p. 2419). The three terms on the rhs of (10) are measures of association, conditional bias, and unconditional bias, respectively. In this context, unconditional bias vanishes if $\bar{g} = \bar{y}$ and conditional bias vanishes if $r_{gy} = s_g/s_y$ *and* $\bar{g} = \bar{y}$. As noted by Murphy and Epstein (1989), this decomposition indicates that the coefficient of determination r_{gy}^2 can be viewed as a measure of potential skill, in the sense that $SS = r_{gy}^2$ when both conditional and unconditional bias vanish.

Expressions for the *SS* in (10) for specific values of r_{gy} are given in Table 2. Note that $SS \leqslant 0$ when $r_{gy} = 0$, with equality being reached only if $\bar{g} = \bar{y}$ *and* $s_g = 0$. (The forecast g in this case is constant and equals $\bar{y}$; see section 4a.) That is, skill is negative when G and Y are uncorrelated, unless the forecasts are conditionally and unconditionally unbiased. Moreover, $SS \leqslant 1$ when $r_{gy} = 1$, with equality only if $\bar{g} = \bar{y}$ *and* $s_g = s_y$. Thus, perfect skill is attained only if the forecasts are conditionally and unconditionally unbiased as well as being perfectly correlated with the observations.

As noted in section 2b, it is possible to interpret r_{gy}^2 (or $r_{g'y'}^2$) as a measure of correspondence in phase when the forecasts and observations (or analysis) are expressed in the form of two-dimensional fields. Correspondence in amplitude for such forecasts is measured by the quantities $[r_{gy} - (s_g/s_y)]^2$ and $[(\bar{g} - \bar{y})/s_y]^2$ [see (10)]. This latter correspondence is complete only when $r_{gy} = s_g/s_y$ and $\bar{g} = \bar{y}$.

TABLE 1. Expressions for the MSE for three specific values of the coefficient of correlation r_{gy}.

r_{gy}	MSE
0	$(\bar{g} - \bar{y})^2 + (s_y^2 + s_g^2)$
s_g/s_y	$(\bar{g} - \bar{y})^2 + (s_y^2 - s_g^2)$
1	$(\bar{g} - \bar{y})^2 + (s_y - s_g)^2$

TABLE 2. Expressions for the *SS* for three specific values of the coefficient of correlation r_{gy}.

r_{gy}	*SS*
0	$-(s_g/s_y)^2 - [(\bar{g} - \bar{y})/s_y]^2$
s_g/s_y	$(s_g/s_y)^2 - [(\bar{g} - \bar{y})/s_y]^2$
1	$(s_g/s_y)[2 - (s_g/s_y)] - [(\bar{g} - \bar{y})/s_y]^2$

4. Further insight: A linear regression model

a. *A linear regression model*

In view of the fact that the coefficients of correlation and determination describe the strength of the linear statistical relationship between G and Y, additional insight into the interpretation of r_{gy} and r_{gy}^2 as measures of performance can be obtained by considering a simple linear regression model involving the forecasts and observations. When the forecasts are regressed on the observations, this model takes the following form:

$$E(Y|g) = \alpha + \beta g, \quad (11)$$

where $E(Y|g)$ is the expected (or mean) value of Y given the forecast $G = g$, and α and β are (unknown) regression coefficients representing the intercept and slope, respectively, of the linear model. The least-squares estimates of these coefficients are denoted here by a and b, respectively, where

$$a = \bar{y} - b\bar{g}, \quad (12)$$

and

$$b = (s_y/s_g)r_{gy}. \quad (13)$$

Substituting the expressions for a and b into (11) yields a linear regression equation that provides a least-squares estimate $\hat{Y}_g$ of $E(Y|g)$:

$$\hat{Y}_g = \bar{y} + [(s_y/s_g)r_{gy}](g - \bar{g}). \quad (14)$$

It should be noted that both the intercept and slope of the regression equation in (14) are influenced by the value of the correlation coefficient r_{gy} [see (12) and (13)]. Moreover, the slope coefficient $b = (s_y/s_g)r_{gy}$ consists of two factors, r_{gy} and s_y/s_g.

Expressions for the slope coefficient b and the conditional mean $\hat{Y}_g$ for $r_{gy} = 0$, $r_{gy} = s_g/s_y$, and $r_{gy} = 1$ are shown in Table 3. When $r_{gy} = 0$, $b = 0$ and $\hat{Y}_g = \bar{y}$.

TABLE 3. Expressions for the slope b and the conditional mean $\hat{Y}_g$ when the forecasts (G) are regressed on the observations (Y), for three specific values of the coefficient of correlation r_{gy}.

r_{gy}	b	$\hat{Y}_g$
0	0	$\bar{y}$
s_g/s_y	1	$\bar{y} + (g - \bar{g})$
1	s_y/s_g	$\bar{y} + (s_y/s_g)(g - \bar{g})$

TABLE 4. Expressions for the terms in the decomposition of the MSE based on the linear regression model for three specific values of the coefficient of correlation r_{gy}.

r_{gy}	MSE_1	MSE_2
0	$(\bar{g} - \bar{y})^2 + s_g^2$	s_y^2
s_g/s_y	$(\bar{g} - \bar{y})^2$	$s_y^2 - s_g^2$
1	$(\bar{g} - \bar{y})^2 + (s_y - s_g)^2$	0

In this case, the forecasts G and observations Y are uncorrelated, and the best estimate of Y (in a least squares sense) is the unconditional sample mean $\bar{y}$. When $r_{gy} = s_g/s_y$, $b = 1$ and $\hat{Y}_g = \bar{y} + (g - \bar{g})$. In this case, the best estimate of the observation Y is the forecast $G = g$ adjusted by the difference between the respective sample means (i.e., $\bar{y} - \bar{g}$). Finally, when $r_{gy} = 1$, $b = s_y/s_g$ and $\hat{Y}_g = \bar{y} + (s_y/s_g)(g - \bar{g})$. In this case, the variables G and Y are perfectly correlated, and the best estimate of Y requires an adjustment of $G = g$ involving the sample means *and* the sample standard deviations.

b. Decomposition of MSE

The MSE in (6) can be decomposed into two terms taking into account the conditional mean $\hat{Y}_g$; specifically,

$$\text{MSE} = \text{MSE}_1 + \text{MSE}_2, \tag{15}$$

where

$$\text{MSE}_1 = \text{MSE}(g, \hat{Y}_g) = (\bar{g} - \bar{y})^2 + (s_g - r_{gy}s_y)^2 \tag{16}$$

and

$$\text{MSE}_2 = \text{MSE}(\hat{Y}_g, y) = (1 - r_{gy}^2)s_y^2. \tag{17}$$

The term MSE_1 in (16) represents the contribution to MSE due to differences between the forecasts and the conditional means of the observations, whereas the term MSE_2 in (17) represents the contribution to MSE due to differences between the conditional means of the observations and the individual observations. It is this latter term (i.e., MSE_2) that is minimized in a least-squares sense by choosing the values of the regression coefficients equal to a and b in (12) and (13), respectively.

The expressions for these terms in the cases of the three selected values of r_{gy} are given in Table 4. This table helps to clarify the interpretations given in conjunction with Table 1 (see section 3b). For example, it can be seen that $\text{MSE}_2 = 0$, but $\text{MSE}_1 \geqslant 0$ when $r_{gy} = 1$. That is, the contribution due to MSE_2 does indeed vanish when $r_{gy} = 1$, but the contribution due to MSE_1 remains and vanishes only for conditionally and unconditionally unbiased forecasts. It also should be noted that MSE_1 is minimized when $r_{gy} = s_g/s_y$, whereas MSE_2 is minimized when $r_{gy} = 1$.

c. Decomposition of SS

In the case of the *SS*, it follows from (7) and (15)–(17) that

$$SS = SS_0 - SS_1 - SS_2, \tag{18}$$

where $SS_0 = 1$,

$$SS_1 = [(\bar{g} - \bar{y})/s_y]^2 + [r_{gy} - (s_g/s_y)]^2, \tag{19}$$

and

$$SS_2 = 1 - r_{gy}^2. \tag{20}$$

The term SS_1 in (19) represents the loss in skill due to unconditional and conditional bias, whereas the term SS_2 in (20) represents the loss in skill due to the lack of perfect linear association between the forecasts and observations. Alternatively, SS_2 and SS_1 can be interpreted as the loss in skill due to phase and amplitude errors in model forecasts, respectively.

Expressions for the terms in the *SS* in (18) in the cases of the three specific values of r_{gy} are given in Table 5. This table helps to clarify the interpretations given in conjunction with Table 2 (see section 3c). For example, it can be seen that $SS_2 = 0$, but $SS_1 \geqslant 0$ when $r_{gy} = 1$. That is, the contribution due to SS_2, which measures association or correspondence in phase, does indeed vanish when $r_{gy} = 1$. However, the contribution due to SS_1 remains and vanishes only when the forecasts are either conditionally and unconditionally unbiased or exhibit complete correspondence in amplitude. Finally, it should be noted that SS_1 is minimized when $r_{gy} = s_g/s_y$, whereas SS_2 is minimized when $r_{gy} = 1$.

TABLE 5. Expressions for the terms in the decomposition of the SS based on the linear regression model for three specific values of the coefficient of correlation r_{gy}.

r_{gy}	SS_0	SS_1	SS_2
0	1	$[(\bar{g} - \bar{y})/s_y]^2 + (s_g/s_y)^2$	1
s_g/s_y	1	$[(\bar{g} - \bar{y})/s_y]^2$	$1 - (s_g/s_y)^2$
1	1	$[(\bar{g} - \bar{y})/s_y]^2 + [1 - (s_g/s_y)]^2$	0

TABLE 6. Decomposition of the skill score SS in (10) for minimum temperature forecasts for Portland, Oregon, in the cool season during the period of October 1980 through March 1987 (o = objective forecasts, s = subjective forecasts).

Lead time (hours)	Type of forecast	Sample size n	Skill score SS	Linear association r^2_{gy}	Conditional bias $[r_{gy} - (s_g/s_y)]^2$	Unconditional bias $[(\bar{g} - \bar{y})/s_y]^2$
24	o	977	0.678	0.731	0.002	0.052
	s	977	0.742	0.761	0.000	0.018
36	o	957	0.581	0.650	0.000	0.069
	s	957	0.668	0.684	0.000	0.016
48	o	967	0.554	0.606	0.002	0.051
	s	967	0.624	0.647	0.000	0.024
60	o	956	0.448	0.507	0.002	0.058
	s	956	0.559	0.580	0.000	0.021

5. Discussion

First and foremost, it should be clear from the definitions and interpretations given in sections 2–4 that r_{gy} and r^2_{gy} are measures of the linear association between the forecasts and observations. That is, they measure the correspondence between a linear function of the forecasts, $\alpha + \beta G$ (where α and β are chosen to minimize the sum of the squared deviations of the values of this function from the values of Y) and the observations. In particular, the CoC and CoD do *not* measure the correspondence between the values of G and Y and thus do *not* in general represent measures of accuracy or skill. They "approach" measures of these latter aspects of performance only when the forecasts are both conditionally and unconditionally unbiased, a situation that seldom if ever holds exactly in the real world.

In the case of forecasts and observations presented in the form of two-dimensional fields, the CoC and the CoD can be viewed as measures of the correspondence in phase between G and Y (see section 2b). Interpretation of r_{gy} and r^2_{gy} as measures of correspondence in phase represents an extension (to two dimensions) of their one-dimensional interpretation as measures of linear association. Since the extent to which the phase of forecasts is in agreement with the phase of observations (or the analysis) is of considerable interest in model verification, the use of such measures is not entirely inappropriate in this context. What must be kept in mind, however, is that the CoC and the CoD are measures of only one of many potentially relevant aspects of model performance.

With regard to the relative merits of r_{gy} and r^2_{gy} as measures of linear association (or correspondence in phase), the CoD would appear to be superior to the CoC. First, r^2_{gy} possesses a meaningful operational interpretation as a measure of the proportion of the variability in the observations that is accounted for by a linear model in the forecasts. No equivalent clear-cut interpretation can be given to r_{gy}. Second, since $|r_{gy}| \geqslant r^2_{gy}$ (with equality being reached only when $r_{gy} = 0$ or 1), the CoC gives the impression of a closer relationship between G and Y than is reflected by the CoD. In this regard, it should be noted that when $r_{gy} = 0.6$ (for example), the forecasts are not 60% of the way toward perfect association (or correspondence in phase) with the observations but are instead only 36% ($= r^2_{gy}$) of this "distance." For these reasons, measuring and reporting values of r^2_{gy} instead of values of r_{gy} seems more appropriate in the context of forecast verification, including model verification.

Comparative verification in which the CoD or the CoC (including the ACC) serves as the sole measure of forecasting performance is a practice fraught with considerable danger. To illustrate some of the problems that can arise when relative performance is judged solely on the basis of the CoD (i.e., r^2_{gy}), the terms in the decomposition of the skill score SS in (10) for samples of minimum temperature (TMIN) and probability of precipitation (PoP) forecasts for Portland, Oregon, are presented in Tables 6 and 7, respectively. These tables contain numerical values of the various terms in this decomposition for both objective (i.e., numerical–statistical) forecasts and subjective forecasts for several lead times. The values of r^2_{gy} and SS indicate that association and skill, respectively, for both variables, and all lead times are greater for the subjective forecasts than for the objective forecasts (recall that the latter are provided to forecasters as guidance in the preparation of the former). Also, conditional bias is generally quite small (especially for the TMIN forecasts) and unconditional bias is considerably larger for the objective forecasts than for the subjective forecasts.

Comparison of the two types of forecasts based on r^2_{gy} alone suggests that differences in performance (i.e., association) are generally quite modest, especially in the case of the 24-h PoP forecasts. However, such a comparison ignores both conditional and unconditional bias, and these biases are larger—in the case of unconditional bias considerably larger—for the objective forecasts than for the subjective forecasts. Thus, a comparison of the two types of forecasts based on the SS, which incorporates these two aspects of bias, indicates appreciably larger differences in performance

TABLE 7. Decomposition of the skill score *SS* in (10) for probability of precipitation forecasts (0000 cycle time) for Portland, Oregon, in the cool season during the period of October 1980 through March 1987 (o = objective forecasts, s = subjective forecasts).

Lead time (hours)	Type of forecast	Sample size n	Skill score SS	Linear association r_{gy}^2	Conditional bias $[r_{gy} - (s_g/s_y)]^2$	Unconditional bias $[(\bar{g} - \bar{y})/s_y]^2$
24	o	1015	0.470	0.488	0.007	0.011
	s	1015	0.493	0.494	0.000	0.001
36	o	1014	0.352	0.376	0.005	0.019
	s	1014	0.397	0.402	0.000	0.005
48	o	1000	0.270	0.285	0.000	0.015
	s	1000	0.309	0.316	0.002	0.005

(i.e., skill). In fact, although these datasets do not contain any such results, it is possible for r_{gy}^2 to indicate that forecasting method A is superior to forecasting method B, whereas the *SS* indicates that method B is superior to method A. Although such a result may be relatively unlikely in the case of the two types of forecasts considered here, it is certainly not impossible in this or other contexts.

To be even-handed, it should be noted that although measures such as the *SS* or the MSE take unconditional and conditional bias into account, they do so in a way that is not entirely satisfactory. In essence, these measures represent composite, one-dimensional measures of several aspects of quality [e.g., see (8) and (10)]. Thus, the relative magnitudes of contributions to the *SS* or the MSE—contributions associated with specific basic characteristics of the forecasts, the observations, and their relationship—are not evaluated and compared separately when comparative verification is based solely on the respective values of these overall measures of accuracy and skill.

The solution to this apparent dilemma is not to choose between r_{gy} and the MSE, or between r_{gy}^2 and the *SS*, but rather to compute separate measures of these different aspects of quality. The relative performance of A's and B's forecasts can then be compared in a framework that presumably more closely approximates the multidimensional structure of the underlying verification problem (see Murphy 1991). Of course, comparisons in such a framework may lead to mixed results, in the sense that A's forecasts are superior to B's forecasts on one dimension and B's forecasts are superior to A's forecasts on another dimension. This result should not be viewed as uninformative or unsatisfactory, but should instead be recognized for what it is—a more realistic appraisal of relative performance in terms of *basic* dimensions of quality. Moreover, to the extent that these aspects of quality can be related to specific features of forecasting methods or models, they may be helpful in identifying ways of improving the forecasts produced by these methods.

6. Conclusions

This paper has demonstrated that the coefficients of correlation and determination, as measures of forecasting performance, describe the degree of linear association—or correspondence in phase—between forecasts and observations. As a result, these coefficients ignore both unconditional bias and conditional bias (or correspondence in amplitude), basic aspects of forecasting performance. It has also been shown here that these coefficients are *not* measures of accuracy and skill, according to long-standing and widely accepted definitions of these familiar aspects of forecast quality.

The measures of aspects of forecasting performance described and compared here are considered to be of potential interest within what might be identified as a "squared-error approach" to verification problems. Verification measures associated with alternative approaches or frameworks could also be applied to these problems. Two such frameworks are the "linear-distance approach" pioneered by Mielke and colleagues (e.g., Mielke 1991; Gray et al. 1993) and the "linear-error-in-probability-space approach" developed by Folland and colleagues (e.g., Potts et al. 1995; Ward and Folland 1991). Comparison of these approaches—and/or their respective measures of performance—is beyond the scope of the current paper.

In view of the multidimensional and multifaceted nature of verification problems, it is important to recognize that *all* one-dimensional measures of forecasting performance either (a) focus on a specific aspect of quality, such as linear association or correspondence in phase in the case of the coefficients of correlation and determination, or (b) represent a particular composite of several aspects of quality, such as linear association, conditional bias, and unconditional bias in the case of the skill score based on the mean square error. To the extent that evaluation and/or comparison of forecasting methods based on a single one-dimensional verification measure is a meaningful exercise, it would appear that measures that account for several aspects of performance (such as the mean square error or the skill score, within the context of the framework adopted here) are to be preferred to measures that focus on a single aspect of quality (such as the coefficients of correlation and determination).

Recent efforts to develop a conceptually and methodologically sound approach to verification problems suggest that each identifiable aspect of forecast quality should be measured separately, and then absolute or relative performance should be judged in the multidimensional space defined by these aspects of quality. The fact that this approach promises to be more insightful and useful than the traditional approach to verification problems can be illustrated by recognizing that it is the errors in forecasts that occur consistently over a sample (unconditional bias) or over an identifiable subsample (conditional bias) that are most likely to be amenable to reduction or elimination. Thus, the need to identify and measure these basic aspects of quality individually should be self-evident. This need can be met by adopting a multifaceted approach and by (inter alia) calculating and reporting all of the terms in the decompositions of the mean square error and the associated skill score. For recent applications of such multidimensional, aspects-oriented approaches to verification problems, see Livezey et al. (1995), Murphy et al. (1989), and Murphy and Winkler (1992).

Acknowledgments. This work was undertaken during a period in which the author was a freelance scientist at the Swedish Meteorological and Hydrological Institute. The facilities made available to the author during this period are gratefully acknowledged. Constructive comments on earlier versions of the paper were provided by three anonymous reviewers.

REFERENCES

Arpe, K., A. Hollingsworth, M. S. Tracton, A. C. Lorenc, S. Uppala, and P. Kållberg, 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.

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Wea. Forecasting,* **7,** 699–709.

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

Gray, W. M., C. W. Landsea, P. W. Mielke, and K. J. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting,* **8,** 73–86.

Jensenius, J. S., 1990: A statistical comparison of the forecasts produced by the NGM and LFM for the 1987/88 cool season. *Wea. Forecasting,* **5,** 116–127.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at NMC. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428.

Livezey, R. E., 1994: The evaluation of forecasts. *Analysis of Climate Variability: Applications of Statistical Techniques,* H. von Storch, and A. Navarra, Eds., Springer-Verlag, 177–196.

——, J. D. Hoopingarner, and J. Huang, 1995: Verification of official monthly mean 700 hpa height forecasts: An update. *Wea. Forecasting,* **10,** 512–527.

Mielke, P. W., 1991: The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Sci. Rev.,* **31,** 55–71.

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

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

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

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

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

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

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

Neter, J., W. Wasserman, and G. A. Whitmore, 1988: *Applied Statistics.* Allyn and Bacon, Inc., 1006 pp.

Palmer, T. N., C. Brankovic, F. Molteni, S. Tibaldi, L. Ferranti, A. Hollingsworth, U. Cubasch, and E. Klinker, 1990: The European Centre for Medium-Range Weather Forecasts (ECMWF) program on extended-range prediction. *Bull. Amer. Meteor. Soc.,* **71,** 1317–1330.

Potts, J. M., C. K. Folland, I. T. Jolliffe, and D. Sexton, 1995: Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *J. Climate,* **8,** in press.

Tracton, M. S., 1993: On the skill and utility of NMC's medium-range central guidance. *Wea. Forecasting,* **8,** 147–153.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.,* **11,** 711–743.

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