Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score

ANTHONY G. BARNSTON

Climate Analysis Center, NMC/NWS/NOAA, Washington, D.C.

15 April 1992 and 13 July 1992

ABSTRACT

The correspondence among the following three forecast verification scores, based on forecasts and their associated observations, is described: 1) the correlation score, 2) the root-mean-square error (RMSE) score, and 3) the Heidke score (based on categorical matches between forecasts and observations). These relationships are provided to facilitate comparisons among studies of forecast skill that use these differing measures.

The Heidke score would be more informative, more "honest," and easier to interpret at face value if the severity of categorical errors (i.e., one-class errors versus two-class errors, etc.) were included in the scoring formula. Without taking categorical error severity into account the meaning of Heidke scores depends heavily on the categorical definitions (particularly the number of categories), making intercomparison between Heidke and correlation (or RMSE) scores, or even among Heidke scores, quite difficult.

When categorical error severity is taken into account in the Heidke score, its correspondence with other verification measures more closely approximates that of more sophisticated scoring systems such as the experimental LEPS score.

1. Definitions and descriptions of three verification measures

Researchers have a wide choice regarding the quantitative evaluation of forecast skill in the results of their prediction studies. The correlation coefficient, the root-mean-square error, and the Heidke score are three commonly selected measures, among many others. In this section these are defined and briefly described, and the correlation coefficient and root-mean-square error are interrelated. In later sections the focus is placed largely on the relationship between the correlation and Heidke scores. Following a discussion about modifying the Heidke score to reduce the discrepancy with the correlation, a quick examination of the correspondence between the correlation and the linear error in probability space (LEPS) score (Ward and Folland 1991) is provided. The discussion applies primarily to continuous underlying variables.

a. Correlation coefficient

One possible choice of a forecast verification measure is the correlation coefficient, which describes the strength of the linear relationship between forecasts and corresponding observations. The correlation may be computed over a period of record, over a spatial domain for a single forecast, or a combination of both. It is a continuous parameter—that is, it is sensitive to the finest details of each forecast versus observed case. If the forecasts (f) and observations (o) are standardized[1] (resulting in zero means and unit standard deviations) and denoted as z_f and z_o, the coefficient of correlation between f and o, r_{fo}, is defined as

$$r_{fo} = \sum_{i=1}^{N} (z_{f_i} z_{o_i})/N \qquad (1)$$

where N is the number of time elements if the correlation is temporal, space elements (e.g., grid points or stations) if it is spatial, or a combination over both dimensions. The i denotes the element number. The correlation would not be different if computed without first standardizing f and o; however, the complete correlation formula would be required in which standardization is accomplished using the means and standard deviations of f and o.[2] If there were an exact linear

Corresponding author address: Anthony G. Barnston, Climate Analysis Center, NMC/NWS/NOAA, 5200 Auth Road, Camp Springs, MD 20746.

[1] As described in most statistical references, a set of forecasts or observations is standardized by first computing its mean and its standard deviation. The standard deviation equals the square root of the mean of the squared differences between each member of the set and the mean. Each member is then standardized by subtracting the mean, and then dividing by the standard deviation. For samples of 100 or fewer that do not contain extreme outlier values, standardized values typically fall within the −3 to +3 range.

[2] As an example of an exception, standardization is not accomplished exactly when means and standard deviations of o, or of both f and o, are set to those of a relatively longer period of record than that used for the correlation calculation; that is, the sample means

functional relationship between forecasts and observations (implying perfect forecast skill), the correlation would be at its maximum possible value of 1.0; if there were no linear predictability whatsoever, it would be zero. In the real world of forecasting, the correlation is usually found to be at an intermediate value that describes, within a linear framework, the quality or relative accuracy of the forecasts in the sample. In the case of a nonlinear relationship, the correlation can underestimate forecast skill significantly, as when the forecasts and observations contain wavelike patterns that are partly out of phase (sine and cosine, for example). On the other hand, while less commonly found in practice, the correlation also can overrepresent forecast accuracy. For example, the correlation is not affected by the amplitude of the original (unstandardized) forecasts, which makes high correlations possible even when the set of forecasts rather uniformly tend to be too weak or too strong.

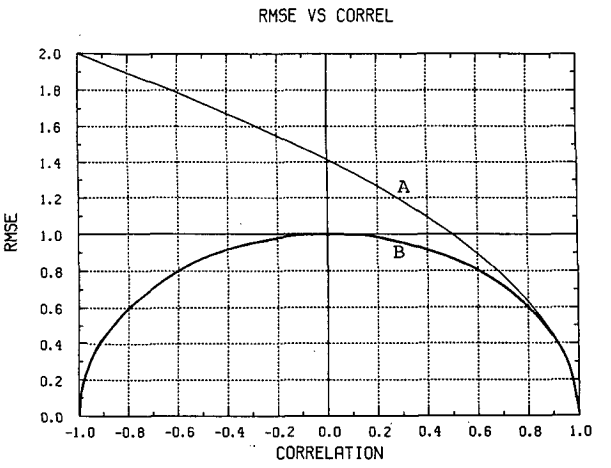

FIG. 1. Root-mean-square error (RMSE) as a function of correlation for standardized sets of forecasts and observations (curve A), and for same except that the forecasts have been damped and possibly sign reversed by multiplying by r_{fo}—i.e., the correlation between forecasts and observations (curve B).

b. *Root-mean-square error*

Another verification measure is the root-mean-square error (RMSE)[3], which is defined in the context of, as before, standardized variables as the square root of the mean of the squared differences between corresponding elements of the forecasts and observations:

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2/N]^{1/2}. \quad (2)$$

There is an exact one-to-one relationship between the correlation coefficient and the RMSE parameter when the latter is formed using standardized forecasts and observations. This nonlinear relationship, shown graphically by curve A in Fig. 1, is given by

$$\mathrm{RMSE}_{fo} = [2(1 - r_{fo})]^{1/2}. \quad (3)$$

When $r = 0$, which occurs when forecasts are perfectly random with respect to the observations, the RMSE score is $\sqrt{2} = 1.41$—a less favorable outcome than when uniform climatology forecasts (i.e., $z_{f_i} = 0$ for all i) are issued, in which case r_{fo} is not computable (because the standard deviation of f is zero, and z_f is undefined) but RMSE_{fo} computed using (2) would equal 1.00. A commonly used reference of minimum usable forecast skill requires that $\mathrm{RMSE}_{fo} \leq 1.00$ (Hollingsworth et al. 1980), which corresponds to $r_{fo} \geq 0.50$ (Roads 1986) as noted in curve A in Fig. 1. For nonstandardized f and/or o, (2) is still used to compute RMSE, but the relationship to the correlation r_{fo} becomes

$$\mathrm{RMSE}_{fo} = [s_f^2 + s_o^2 - 2s_f s_o r_{fo} + b^2]^{1/2}, \quad (4)$$

where s_f and s_o are the standard deviations of the sets of forecasts and observations, respectively, and b is the forecast bias, defined as the mean of f minus the mean of o. The contribution of this overall bias and of more subtle biases [conditional, or differential, biases, which would weaken r_{fo} in (4)] to the RMSE score is examined in detail in Murphy and Epstein (1989). When f and o are standardized, b becomes zero, s_f and s_o become unity, and (4) reduces to (3). In the unstandardized case, differences in the relative amplitude of the set of forecasts versus that of the observations increase the RMSE score (i.e., show up as poorer skill) but do not affect the correlation score, which is sensitive only to the temporal (and/or spatial) *phasing* of the forecasts with the observations.

The RMSE scores corresponding to $r_{fo} < 1.00$ (almost all of curve A in Fig. 1) can be reduced by uniformly damping the amplitude of the set of imperfect forecasts (i.e., reducing s_f) by multiplying each by r_{fo}:

$$d_i = f_i r_{fo}, \quad i = 1 \text{ to } N, \quad (5)$$

where d denotes a damped forecast. (We discuss how r_{fo} is known *before* the forecasts are verified shortly.) Note that if $-1 \leq r_{fo} < 0$, (5) not only damps the forecasts but also reverses their sign, eliminating the tendency for forecasts and observations to have opposite sign. Following the damping/sign correction process, RMSE must be computed using (4) because

and standard deviations are not used because the longer-term statistics are considered to be better estimates. This has been done for certain versions of the spatial anomaly correlation (Miyakoda et al. 1972; Saha and Van den Dool 1988). While reasons for choosing this version of calculation of the correlation are well grounded, the resulting coefficient is a "partial" (not to be confused with a true partial correlation) rather than total or conventional correlation coefficient, and thus may not obey some of the relationships highlighted here.

[3] The Brier score (Brier and Allen 1951) is essentially the square of the RMSE—that is, the MSE. However, it was intended for application to probability forecasts rather than point value forecasts.

d is not standardized. For the sake of simplicity, we assume that s_o remains unity and b remains zero:

$$\mathrm{RMSE}_{do} = [s_d^2 + 1 - 2s_d r_{do}]^{1/2}.$$

Because s_d equals r_{fo} and r_{do} equals $\pm r_{fo}$, this expression can be rewritten as

$$\mathrm{RMSE}_{do} = [1 - r_{fo}^2]^{1/2}. \quad (6)$$

The relationship described by (6) is shown as curve B in Fig. 1. All sets of forecasts having nonzero RMSE scores benefit from damping and/or sign reversal, after which the maximum RMSE of 1.00 occurs at $r_{fo} = 0.00$, equaling the RMSE of uniformly issued climatology forecasts (i.e., $z_{f_i} = 0$ for all i) because the damping process reduces the original random forecasts to exactly that [i.e., $d_i = 0$ in (5); the forecast amplitude becomes zero]. By contrast, the correlation coefficient is not changed by the uniform damping (provided that the forecasts are not damped completely to zero amplitude), but a sign change would beneficially change an originally negative correlation.

It may seem peculiar to assume knowledge of r_{fo} (the damping factor) before forecast verification, when r_{fo} is determined in the verification. When a regression model is used to describe the f versus o relationship in a hindcast mode (i.e., all f_i and their corresponding o_i are known), r_{fo} is computed and then automatically used to damp the forecasts that then may be used for the computation of RMSE_{fo}. When forecasts are issued for independent cases (future times or any times for which o is not, or cannot be, used to develop the f versus o relationship), r_{fo} is estimated using only the sample over which both f_i and the corresponding o_i are used (called the *development* sample). Because r_{fo} is only an approximation of what it would be if the independent case(s) were included, the damping factor is not exactly optimal, and the success of the forecasts applied to the independent cases is expected to be somewhat less than that described by r_{fo} based on the development sample (Davis 1976). In fact, when the development sample size is small (e.g., <20) and the unknown population value of r_{fo} is high (>0.75), there is a nonnegligible chance that the damping factor will be misrepresented (underestimated) to the extent that the RMSE is actually increased (due to overdamping) from its value for undamped forecasts.

When procedures other than regression are used to produce forecasts for independent cases (e.g., in numerical weather prediction), damping may not be done automatically and application of (5) can be done "manually" based on a best estimate of r_{fo}. On the other hand, there are times when damped forecasts and the accompanying minimization of RMSE are *not* desired—perhaps because forecasts would be too conservative. In such cases forecast restandardization can be carried out, and a correlation score or Heidke score can be used as the skill measure.

c. *Heidke skill score*

Forecasts are expected to be only roughly accurate for difficult parameters such as precipitation amount in 2+-day forecasts, or a time-mean forecast at long projection times. In such cases, a set of forecasts and their associated observations sometimes are converted to categories (such as below, near, or above normal for temperature) and scored as the number of categorical matches (i.e., correct forecasts) compared with that expected by chance alone. The rationale behind this conversion is, perhaps, that because only rough forecasting capability is expected, the precision of a continuous measurement tool such as the correlation coefficient or RMSE is unnecessary. (We leave aside whether this is a wise course of action.) One commonly used categorical verification score is the Heidke score (Heidke 1926), defined as:

$$\mathrm{Heidke} = (H - E)/(N - E), \quad (7)$$

where H is the number of categorically correct forecasts ("hits"), N is the total number of forecasts issued (over time, space, or both), and E is the number of categorically correct forecasts expected by chance in the absence of any forecasting skill. This score can be computed for any number of categories. It is defined such that a perfect set of forecasts (i.e., all categorical hits) would be scored as 1.00, a set of random forecasts would have an expected score of zero, and sets of forecasts having fewer hits than would be expected by chance would have negative scores. A number of measures similar to the Heidke score have also been used. For example, one of the scores used in Van den Dool and Toth (1991) is identical except that E does not appear in the denominator. Highly related scores are the performance index (Daan 1985; Hanssen and Kuipers 1965) and the Gringorten skill score (Gringorten 1965). Although the Heidke score behavior is most easily described using equally probable categories (such as three defined equally likely temperature categories—cold, near normal, and warm), the score can be applied to any categorical configuration, including highly asymmetric categories (as in Klein and Charney 1992), as long as E is determined properly. An examination of categorical skill measures used for asymmetric, inherently discrete categories such as those associated with the occurrence of rare weather events is found in Doswell et al. (1990). As one might expect, there is much in common between categorical skill measurement of discrete variables and continuous variables that are forced into categories.

As defined in (7), the Heidke score is insensitive to the severity of errors (in terms of the number of categories involved in a "miss") when more than two categories are used. For example, for three categories, two-class errors (e.g., forecasting cold when warm is observed) are not differentiated from one-class errors. This characteristic turns out to play an important role

TABLE 1. Some characteristics of the Heidke score for 2, 3, 4, and 5 equally likely categories. Only hits are credited; all misses count equally.

Number of equally likely categories	Minimum score (no hits; all misses count equally)	Randomly expected probability (E)[1] Hit	Miss	Equitable scoring system?	Standardized class limit cutoffs
2	−1.000	.500	.500	yes	.0000
3	−.500	.333	.667	yes	±.4307
4	−.333	.250	.750	yes	.0000, ±.6745
5	−.250	.200	.800	yes	±.2533, ±.8418

[1] The "E" used in Eq. (7) is the product of the value given here and the sample size of the forecasts, N.

in the comparison with correlation scores, as illustrated in the following section.

2. Characteristics of Heidke score for equally probable categories, and expected correspondence to correlation score

Table 1 presents some characteristics of Heidke scores for the cases of 2, 3, 4, and 5 equally probable categories. While zero is the expected score in the absence of any forecast skill for all cases, the minimum score (obtained when there are no hits) is $-1/(k-1)$ where k is the number of categories. Table 1 also presents probabilities, given random (no skill) forecasts, for a categorical hit and a miss. The value of E in (7) is based on the expected probability of a hit for a single random forecast trial (column 3 in Table 1). It is worth noting that for equally probable categories, the Heidke score is an *equitable* skill score (Gandin and Murphy 1992) in that the probability of random success is constant regardless of which category is forecast. This desirable feature permits forecasters to choose categories on more purely meteorological grounds, as opposed to "playing the odds" of a system with built-in biases. Finally, Table 1 lists the boundaries between the equally likely categories for standardized, assumed Gaussian forecasts and observations. The information in Table 1 is easily computable for higher numbers of equally likely categories and for unequally likely categories, which are mostly not examined in this study.

Figure 2 illustrates a case of 16 forecasts (ordinate) and their associated observations (abscissa) in both a correlational and a 3-class Heidke score context. The cutoffs for standardized f and o allowing for three equally likely categories in an overall sense are determined using the area under the Gaussian (normal) distribution; they are ±0.431, as shown in Table 1. (Another option would be to form these limits on the basis only of rank within the distribution—i.e., use tercile cutoffs.) The correlational aspect in the Fig. 2 illustration is reflected in the shape of the cloud of points in the scatterplot, which produces a 0.38 correlation in this case. The Heidke score is determined by the number of points (denoted by crosses) falling into the lower left, the center, or the upper right squares, representing categorical matches. In this case the number of hits is 8, compared with 5.33 (i.e., 16/3) hits expected by chance, producing a Heidke score of 0.25.

A certain degree of correspondence between correlation and Heidke scores is to be expected, because high correlations would require very linear relationships in which a high proportion of the points would be positioned in one of the three categorically matching sectors. However, there is some uncertainty in the relationship between the two scores. For example, with only small locational changes, points near the categorical boundaries can be moved into nearby parts of adjacent sectors without affecting the shape of the scatterplot (which determines the correlation) very much but changing the number of matches (which determines the Heidke score) by a greater proportion. Also, changes from 1-class to 2-class errors on the parts of several forecasts might occur with large positional changes in some of the points in the scatterplot, which would be reflected in a noticeably lower correlation score but an unchanged Heidke score. Lack of exact

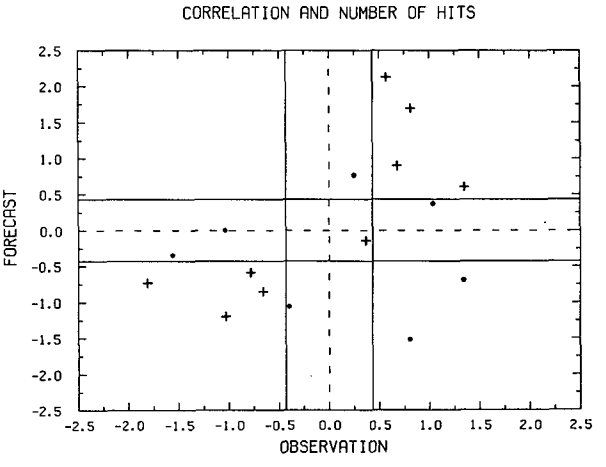

FIG. 2. Observation versus forecast scatterplot for $N = 16$, illustrating the correspondence of the correlation coefficient of 0.38 and the Heidke score (based on number of categorical matches in a three-equally-likely categorical system; note the crosses) of 0.25.

correspondence between Heidke and correlation scores is caused by 1) the discrete character of the Heidke versus the continuous character of the correlation score, 2) the crudeness of Heidke error scoring (scoring all misses equally, regardless of severity) versus the linear precision of correlation error scoring, and 3) the linear keying of errors by the Heidke score (whether error severity is acknowledged or not) versus the quadratic keying of errors in a correlation model (i.e., error outliers greatly affect the regression fit and the resulting correlation score).

3. Correspondence between correlation and Heidke skill scores

Figure 3 displays the mean and the variability of the correspondence between the correlation score and the Heidke score for 2, 3, 4, and 5 equally probable categories in parts a, b, c, and d, respectively. The relationships are derived using iterative sampling of a fixed sample size (N) of forecasts and accompanying observations (in this case, $N = 64$) from a Gaussian random number generator. The mean correlation between forecasts and observations using this data source is zero, and the variability is such that correlations of absolute value greater than 0.2 occur only about 10% of the time. Higher-magnitude correlations of either sign are produced, however, using a secondary iterative procedure for each drawn sample. This procedure consists of systematically modifying the initial correlation by moving each observation versus forecast (o, f) point closer to or farther away from the 45° $o = f$ line along its perpendicular to that line by a fixed proportion of its initial perpendicular distance from the line. In other words, the scatterplot (such as the one shown for $N = 16$ in Fig. 2) is either compressed with respect to

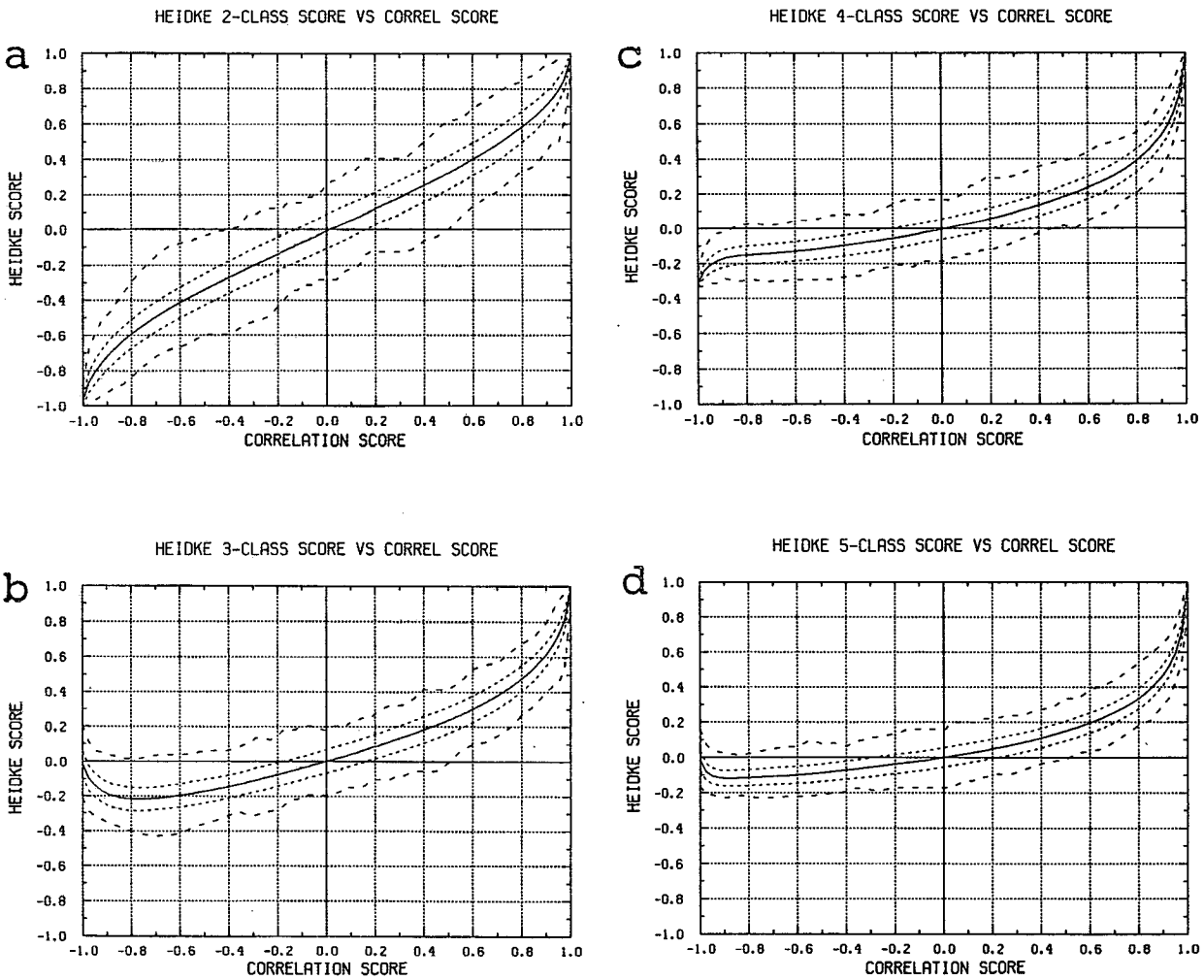

FIG. 3. Heidke score as a function of correlation score for (a) two, (b) three, (c) four, and (d) five equally likely categories, based on a large number of simulations using a random number generator. The solid curve represents mean results, the short-dashed curves the plus- and minus-one standard deviation interval, and the long-dashed curves the maximum and minimum results.

the $o = f$ line (which would increase r) or with respect to the $o = -f$ line (which would decrease r). The sets of forecasts and observations are restandardized following this modification. The procedure is employed 68 times, producing an array of approximately evenly separated correlation values from −.999 to .999 that are used to develop the smooth, sufficiently sampled relationships with Heidke scores as seen in Fig. 3. The procedure of creating 68 different correlation and corresponding Heidke values from one initial o versus f sample correlation is repeated for 500 different initial samples drawn from the Gaussian random number generator. While the sets of 68 correlation values differ among the 500 samples drawn, they and their corresponding Heidke scores are interpolated to a fixed set of 68 correlations. Thus, 34 000 o versus f pairs are used to create each plot. The five curves in each of the four plots in Fig. 3 describe the mean and the variability with respect to the 500 iterations (see caption). (Note that the extreme curves can have noticeable "wiggles" because they represent the outcome of one particular case whose degree of extremeness may differ markedly among adjacent correlation values.) Based on analogous simulations using different values of N, the distance of both sets of dashed curves from the solid (mean) curve is found to vary as $N^{-1/2}$. For example, they would bracket the mean curve at twice the distance if N were 16 instead of 64, indicating a doubling of the uncertainty in the correlation versus Heidke relationship. It is emphasized that the forecasts and observations in the simulations are standardized, thus removing distribution dissimilarities such as overall biases and scaling errors that would degrade the Heidke score but not the correlation score.

Several features of the correspondence are noteworthy. In all four cases, the slope of the curves at and near the origin is less than 1, such that for low and moderate positive correlation scores the Heidke scores are always somewhat lower. This tendency increases with the number of categories. For example, for a correlation score of 0.40, the mean corresponding Heidke score is 0.26 for two equally probable categories, 0.18 for three, 0.14 for four, and 0.11 for five categories. These findings are consistent with the outcome in Van den Dool and Toth (1991), where the Heidke score for a fixed set of fairly unskillful simulated forecasts increased when the width of the middle category in a three-category system was diminished, making the scoring more like that of a two-category system. Thus, intercomparisons among Heidke scores are meaningful at face value only for cases having equal numbers of identically defined categories. The same conclusion is implied in Daan (1985) for the similarly behaved performance index; see his Fig. 3. The variation of slope with the number of categories occurs because, given the same RMSE and corresponding imperfect correlation, the probability of a miss increases with decreasing class width. An accompanying feature in cases with larger numbers of categories is a rapid increase in the Heidke score as the correlation closely approaches 1.00; this is evident in Fig. 3. Along with the lower slope values for the greater number of categories there also appears a somewhat lower standard deviation of the Heidke score associated with a given level of correlation.

The plots for the odd category numbers (3, 5) show improvements in the Heidke score as the correlation becomes highly negative. This occurs as a result of increases in the number of hits in the middle category, as forecasts in either half of the middle category (i.e., upper versus lower half) result in observations on the opposite half of that category. With an even number of categories there is no middle category for this to occur. While negative forecast skill scores can occur when actual skill is close to zero and sampling variability contributes negatively, correlation scores of less than −0.50 are uncommon, assuming a reasonable number of degrees of freedom (N-2) in the sample (i.e., >15). Hence, any behavior in the highly negative portions of the plots in Fig. 3 should have little bearing on our choice of a verification measure.

A plot similar to that of Fig. 3b except for the unequal category definition of 0.30, 0.40, 0.30 (i.e., that used for the Climate Analysis Center's long-range categorical forecasts; Gilman 1986) is not shown here but is virtually indistinguishable from Fig. 3b for correlation scores greater than −0.30. This suggests that, for a given number of categories, the Heidke score is not very sensitive to small departures from equality of likelihood of occurrence among the categories.

4. Penalizing for error severity in the Heidke score

Section 3 and Fig. 3 describe several inconvenient irregularities in the relationship between conventional Heidke and correlation scores. While the correlation score is not necessarily regarded as the standard of accuracy in the comparisons (recall that it has the weakness of being tuned only to linearity), the major undesirable features of the correspondence are attributable to the rough and incomplete nature of the Heidke score. We now ask whether any simple modifications can be applied to the Heidke score that would cure some of these problems without inducing any serious side effects.

The correspondence between correlation and Heidke scores becomes closer and less irregular when the classes of categorical error are taken into account in the Heidke scoring. One simple way to do this is to score hits as +1 (as before), one-class errors as 0, two-class errors as −1, etc. Acknowledging the severity of the error in terms of the number of categories of the failure is a crude distance measure that makes the Heidke scoring model and the correlation scoring model more similar.

Table 2 provides basic probabilistic information for Heidke scoring when the severity of categorical error is linearly taken into account. As in Table 1, it presents

TABLE 2. Some characteristics of a modified Heidke score for 2, 3, 4, and 5 equally likely categories, where error severity (the number of classes separating the forecast from the observation) is linearly taken into account. The two-category system is unaffected by the modification. See text for further explanation.

Number of equally likely categories	Expected probabilities: Hit	Miss	Error severity: 1-class error	2-class error	3-class error	4-class error	Randomly expected probability (E)[1]	Equitable scoring system? (E value by category) [Heidke score by category]
2	.500	.500	.500	—	—	—	.500	yes (.50, .50) [0, 0]
3	.333	.667	.444	.222	—	—	.111	no (0, .33, 0) [−.125, .25, −.125]
4	.250	.750	.375	.250	.125	—	−.250	no (−.50, 0, 0, −.50) [−.20, .20, .20, −.20]
5	.200	.800	.320	.240	.160	.080	−.600	no (−1.00, −.40, −.20, −.40, −1.00) [−.25, .125, .25, .125, −.25]

[1] When the class of the error is acknowledged, one point is added for a "hit," nothing is added for one-class errors, one point is subtracted for two class errors, etc. Equation (7) would still be used, but the "E" term would be modified to the product of the value given here and the sample size of the forecasts, N.

probabilities, given random forecasts, for a categorical hit or a miss. In addition, Table 2 gives probabilities for various possible class errors from 1 up to the $k-1$ class error for a k-category system. The value of E in (7) is now based on a composite of the expected probability of a hit along with that of any of the possible classes of misses for any single forecast trial (second-to-last column in Table 2).

Figure 4 illustrates the effects of acknowledging categorical error classes for the 3, 4, and 5 equal-probability category cases shown in Fig. 3b,c,d. In this comparison it is noted that accounting for error classes makes possible a more complete, balanced Heidke skill evaluation, as evidenced by the better comparability between correlation and Heidke scores in terms of greater uniformity and closeness to unity of the slope of the curves over the range of positive as well as negative correlation. In addition, the inclusion of error category severity also nearly eliminates the difference in slope of the mean curve from one number of categories to another, rendering Heidke scores more intercomparable across these different conditions. For example, the mean Heidke scores corresponding to a correlation score of 0.40 for two, three, four, and five equally likely categories become 0.26, 0.25, 0.25, and 0.24, respectively.

When the categorical error class is squared in the Heidke score, the correspondence between correlation and Heidke scores increases still further (not shown). Because such a squared-error-based Heidke score is just a discretized version of the correlation score, this outcome is hardly surprising.

An undesirable side effect associated with the suggested class-error recognition system is that this scoring system is not equitable (Gandin and Murphy 1992), as it tends to produce higher scores for random forecasts of the middle category(s) than the extreme categories. The numbers in parentheses beneath the "no" entries in the last column of Table 2 are the E values associated with category 1, 2, . . . , k for the k-category system for $k > 2$, and the numbers in brackets below the E values are the corresponding skill scores computed using the overall E value in the second to last column. In the absence of forecasting skill, the extreme categories are more dangerous to forecast because of the more extreme class errors that are possible.

The inequitability of the modified Heidke scoring system is eliminated by adjusting the credits and penalties such that there are no expected differences in E from one forecast category to another. If the E values specific to each forecast category were subtracted from the scores assigned to each such category, equitability would be restored. (In the three-category system, for example, 0.333 would be subtracted from the credit assigned for forecasts of the middle category and nothing would be subtracted for forecasts of the extreme categories.) Following this adjustment a final recalibration would be necessary to retain unity as the overall expected score for hits and zero as the overall score for random forecasts.

The four versions of Heidke scoring described, starting with the original system and ending with the recalibrated equitable version that accommodates error classes, are shown in Table 3 for three equally probable categories. Each element in any of the matrices is the score assigned for a particular combination of forecast and observed categories. The equitability of the final two versions (parts c and d) is confirmed by the fact

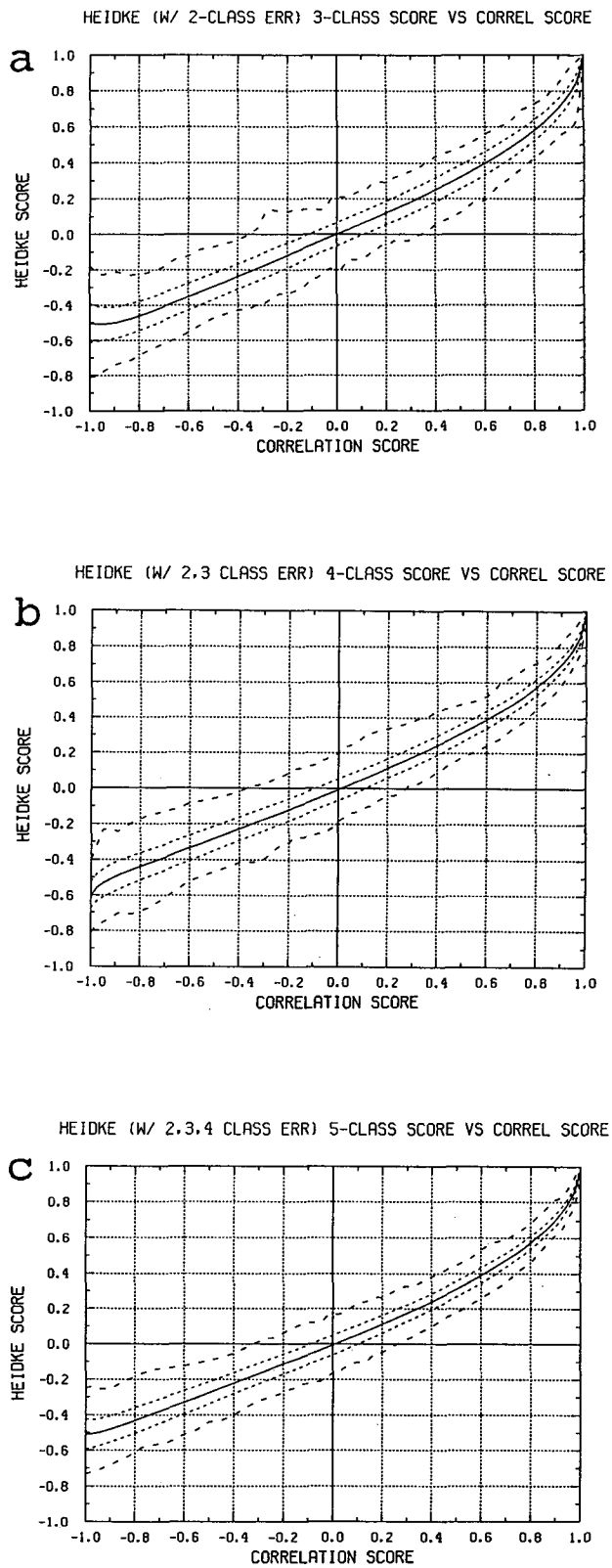

FIG. 4. As in Fig. 3 except for the case in which error severity (number of categories of error) is linearly accounted for in the Heidke scoring for (a) three, (b) four, and (c) five equally likely categories.

that the means of the scores for all rows is the same. The final matrix (part d) has the additional feature of a mean diagonal value (representing hits) of 1.00 and a matrix mean of 0.00. If the second row of a table like Table 2 (i.e., for the three-category system) were constructed for this matrix, all columns except for the last two would have the same entries as those in Table 2; the "E" column would contain zero, and the last column would contain "yes" and zeros in both the parentheses and the brackets. Note that the credit or penalty in matrix d varies not only with the class of error but also with the extremeness of the forecast and observed categories. A related characteristic is the higher scores given for correct forecasts of extreme categories (even though all categories have equal likelihood) than middle categories. While repetitive forecasts of any one of the categories have equal skill expectations, the same characteristic does not apply to the observation categories (columns). However, the latter is not a requirement for equitability.

Table 4 presents scoring matrices for two through five equiprobable categories for the skill modified to accommodate error classes, adjusted to be equitable and recalibrated to 0 and 1 for overall mean random and hit scores, respectively. The correspondences between these scores and correlation scores (not shown) are virtually indistinguishable from those of Fig. 4, which use inequitable Heidke scores modified to account for error classes (the three-category version of whose scoring matrix is shown in Table 3b). A minor difference is that the variabilities of the equitable version do not converge to zero for very high correlations due to the variation in credit given for hits over the forecast categories.

Possibilities exist for equitable error class-sensitive scoring other than the Heidke-derived system developed above. For example, the categorical version of LEPS (linear error in probability space) studied at the British Meteorological Office (Ward and Folland 1991) offers an equitable scoring system that recognizes error severity. For either continuous or discrete variables, the LEPS score is determined by the distance between the forecast and observed values, in terms of the variable's cumulative probability distribution, compared with the chance score for the same forecast–observation pair (see Appendix in Ward and Folland 1991). Table 5 displays the categorical LEPS credit/penalty matrix for two through five equally probable categories, with entries calibrated such that perfect forecasts have an expected score of unity and random forecasts an expected score (E) of zero. Equitability is confirmed by the equality of the means of scores in any row in any one matrix (zero in this case, allowing for roundoff errors). Here the credit or penalty varies with the extremeness of the forecast and observed categories to a greater extent than that of the modified Heidke (Table 4). An added, perhaps desirable "bonus," not required for equitability, is that the columns also have equal

TABLE 3. Four versions of Heidke credit/penalty scoring systems for three equally likely categories, progressing from (A) the original system in which all misses count equally, (B) the system in which penalties depend only on the class of error (no longer equitable), (C) system (B) adjusted for E value by forecast category to restore equitability, and (D) system (C) recalibrated for expected hit mean of 1 and grand mean of 0.

(A) The original Heidke scoring system.

Forecast category	Observed category 1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

(B) Heidke system modified to account for error classes.

Forecast category	Observed category 1	2	3
1	1	0	−1
2	0	1	0
3	−1	0	1

(C) Matrix (B) made equitable by adjusting for E by forecast category.

Forecast category	Observed category 1	2	3
1	1	0	−1
2	−0.333	0.667	−0.333
3	−1	0	1

(D) Matrix C recalibrated for (0, 1) random vs perfect levels.

Forecast category	Observed category 1	2	3
1	1.125	0.000	−1.125
2	−0.375	0.750	−0.375
3	−1.125	0.000	1.125

means, giving the matrix two-dimensional symmetry. The categorical version of the Folland–Painting scoring system used for assessment of montly forecast at the British Meteorological Office (Folland et al. 1986) is also two-dimensionally equitable.

Figure 5 shows the correspondence between LEPS skill scores (using the scoring matrices of Table 5) and the correlation scores for three and five equally likely categories, using the same simulation procedure as for the previous comparisons. The correspondences are quite similar to those shown in the nonequitable modified Heidke counterparts (Fig. 4a,c). However, close examination reveals that the slope of the correspondence curves using LEPS is slightly greater (i.e., closer

TABLE 4. A Heidke credit/penalty scoring system for (A) two, (B) three, (C) four, and (D) five equally likely categories that accounts for the class of error and also retains equitability with respect to forecast classes. Calibrated for random forecasts having an expected score (E) of 0.00 and perfect forecasts 1.00.

(A) Two equiprobable categories

Forecast category	Observed category 1	2
1	1.00	−1.00
2	−1.00	1.00

(B) Three equiprobable categories

Forecast category	Observed category 1	2	3
1	1.125	0.000	−1.125
2	−0.375	0.750	−0.375
3	−1.125	0.000	1.125

(C) Four equiprobable categories

Forecast category	Observed category 1	2	3	4
1	1.20	0.40	−0.40	−1.20
2	0.00	0.80	0.00	−0.80
3	−0.80	0.00	0.80	0.00
4	−1.20	−0.40	0.40	1.20

(D) Five equiprobable categories

Forecast category	Observed category 1	2	3	4	5
1	1.250	0.625	0.000	−0.625	−1.250
2	0.250	0.875	0.250	−0.375	−1.000
3	−0.500	0.125	0.750	0.125	−0.500
4	−1.000	−0.375	0.250	0.875	0.250
5	−1.250	−0.625	0.000	0.625	1.250

TABLE 5. The equitable credit/penalty scoring system based on LEPS (Ward and Folland 1991) for (A) two, (B) three, (C) four, and (D) five equally likely categories, calibrated such that random forecasts have an expected score (E) of zero and perfect forecasts unity.

(A) Two equiprobable categories

FORECAST CATEGORY	Observed category 1	2
1	1.00	−1.00
2	−1.00	1.00

(B) Three equiprobable categories

FORECAST CATEGORY	Observed categeory 1	2	3
1	1.35	−0.15	−1.20
2	−0.15	0.29	−0.15
3	−1.20	−0.15	1.35

(C) Four equiprobable categories

FORECAST CATEGORY	Observed category 1	2	3	4
1	1.51	0.37	−0.66	−1.22
2	0.37	0.49	−0.19	−0.66
3	−0.66	−0.19	0.49	0.37
4	−1.22	−0.66	0.37	1.51

(D) Five equiprobable categories

FORECAST CATEGORY	Observed category 1	2	3	4	5
1	1.60	0.68	−0.22	−0.85	−1.22
2	0.68	0.71	0.03	−0.57	−0.85
3	−0.22	0.03	0.37	0.03	−0.22
4	−0.85	−0.57	0.03	0.71	0.68
5	−1.22	−0.85	−0.22	0.68	1.60

to unity, indicating greater comparability between LEPS and correlation scores), and more identical between the three- and five-category cases. Specifically, for a correlation of .40 the LEPS skill score is 0.28 for both cases, versus 0.25 and 0.24, respectively, for the modified Heidke scores. (The four-category LEPS versus correlation case, not shown, exhibits behavior nearly identical to that of the three- and five-category cases.) The LEPS variability about the mean correspondence differs slightly from that of the modified Heidke, depending on the number of categories and the sector of the plot; however, in general the two methods yield similar dispersions. The behavior for strongly negative correlations differs somewhat, but this need not concern us, for reasons already mentioned in the Heidke versus correlation score comparison.

The foregoing outcomes demonstrate that more than one equitable scoring system that accounts for error classes is practicable, with Heidke-based and LEPS systems being just two of probably many possibilities. While the LEPS scoring scheme appears desirably consistent with respect to its correspondence with the correlation score, it has other minor problems (see section 5 in Ward and Folland 1991) and should not be regarded as a perfect or final scoring system. In fact, some

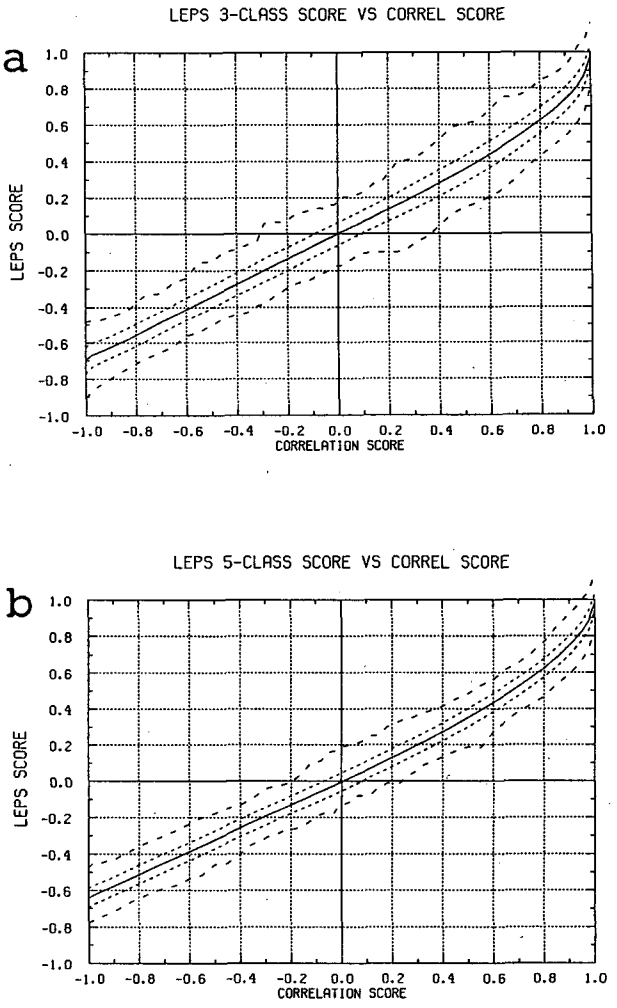

FIG. 5. As in Figs. 3 and 4 except for the LEPS categorical scoring system (see text and Table 5) as a function of correlation score for (a) three and (b) five equally likely categories.

of the finer details of LEPS are still undergoing development (C. K. Folland, personal communication, 1992).

It should be recognized that equitability is not the only consideration in choosing a discrete scoring system, just as comparability to the correlation score should not be an exclusive criterion for desirability. Every system has a set of attributes that may or may not be suitable for the problem at hand. The comparisons described here treat a small proportion of the issues relevant to evaluating skill in research or operational settings.

5. Conclusions

Correlation, RMSE, and Heidke verification scores have been compared under experimental conditions. In many of the experiments, initial standardization of forecasts and observations has removed the complicating effects of overall biases and scaling errors. The correlation and RMSE scores have an exact correspondence when the forecasts and observations are standardized, and a different exact correspondence when the forecasts are multiplied by the forecast versus observation correlation to minimize the RMSE (i.e., damped and sign corrected if needed).

The relationship between correlation (and corresponding RMSE) and Heidke scores has some inherent uncertainty that varies inversely as the square root of the number of points in the sample. More important, the mean correspondence of results varies significantly with the number of equally likely Heidke categories. Clearly, Heidke scores cannot be intercompared unless they are produced using comparable categorical definitions. If they are not, they can be compared following conversion to a common reference such as the correlation or RMSE score.

Many of the complications with Heidke scoring are largely overcome by accounting for the number of classes of error for the cases of three or more classes. This greatly increases the intercomparability among Heidke scores and makes possible more meaningful correspondence to the other verification measures.

There are many possible methods of accounting for categorical error severity. The simple method used here to modify the Heidke score initially causes a loss of scoring equitability (i.e., the randomly expected score no longer remains constant as a function of which category is forecast), but subsequent readjustment restores equitability. Another scoring technique not derived from the Heidke (LEPS, developed in the United Kingdom) that combines the benefits of equitability and accounting for the classes of error is found to relate to the correlation score at least as consistently as (if not more so than) the modified Heidke score.

Acknowledgments. I am grateful for the helpful discussions with Robert E. Livezey and Huug M. van den Dool.

REFERENCES

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

Daan, H., 1985: Sensitivity of verification scores to the classification of the predictand. *Mon. Wea. Rev.,* **113,** 1384–1392.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.,* **6,** 249–266.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

Folland, C. K., A. Woodcock, and L. D. Varah, 1986: Experimental monthly long-range forecasts for the United Kingdom. Part III. Skill of the monthly forecasts. *Meteor. Mag.,* **115,** 377–395.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.,* **120,** 361–370.

Gilman, D. L., 1986: Expressing uncertainty in long-range forecasts, *Namias Symposium,* John O. Roads, Ed., Scripps Institute of Oceanography Reference Series 86-17, University of California, San Diego, La Jolla, CA.

Gringorten, I. I., 1965: A measure of skill in forecasting a continuous variable. *J. Appl. Meteor.,* **4,** 47–53.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.,* **81,** 3–15.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler,* **8,** 301–349 (In German).

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and H. Savijärvi, 1980: The performance of a medium-range forecast model in winter—Impact of physical parameterizations. *Mon. Wea. Rev.,* **108,** 1736–1773.

Klein, W. H., and J. J. Charney, 1992: Verification of monthly outlooks for temperature, dewpoint and wind speed in the contiguous United States. *Proc. of the 16th Annual Climate Diagnostics Workshop,* Lake Arrowhead, California, United States Department of Commerce, 441–445.

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

Roads, J. O., 1986: Forecasts of time averages with a numerical weather prediction model. *J. Atmos. Sci.,* **43,** 871–892.

Saha, S., and H. M. van den Dool, 1988: A measure of the practical limit of predictability. *Mon. Wea. Rev.,* **116,** 2522–2526.

van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. Forecasting,* **6,** 76–85.

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