Equitable Skill Scores for Categorical Forecasts

LEV S. GANDIN

UCAR Scientist, National Meteorological Center, National Weather Service, NOAA, Washington, D.C.

ALLAN H. MURPHY

UCAR Visiting Scientist Program, National Meteorological Center, National Weather Service, NOAA, Washington, D.C.

(Manuscript received 21 March 1991, in final form 21 June 1991)

ABSTRACT

Many skill scores used to evaluate categorical forecasts of discrete variables are inequitable, in the sense that constant forecasts of some events lead to better scores than constant forecasts of other events. Inequitable skill scores may encourage forecasters to favor some events at the expense of other events, thereby producing forecasts that exhibit systematic biases or other undesirable characteristics.

This paper describes a method of formulating *equitable skill scores* for categorical forecasts of nominal and ordinal variables. Equitable skill scores are based on scoring matrices, which assign scores to the various combinations of forecast and observed events. The basic tenets of equitability require that (i) all constant forecasts—and random forecasts—receive the same expected score, and (ii) the elements of scoring matrices do not depend on the elements of performance matrices. Scoring matrices are assumed here to be symmetric and to possess other reasonable properties related to the nature of the underlying variable. To scale the elements of scoring matrices, the expected scores for constant and random forecasts are set equal to zero and the expected score for perfect forecasts is set equal to one. Taken together, these conditions are necessary but generally not sufficient to determine uniquely the elements of a scoring matrix. To obtain a unique scoring matrix, additional conditions must be imposed or some scores must be specified a priori.

Equitable skill scores are illustrated here by considering specific situations as well as numerical examples. These skill scores possess several desirable properties: (i) The score assigned to a correct forecast of an event increases as the climatological probability of the event decreases and (ii) scoring matrices in $n + 1$-event and n-event situations may be made consistent, in the sense that the former approaches the latter as the climatological probability of one of the events approaches zero. Several possible extensions and applications of this method are discussed.

1. Introduction

Skill scores are measures of the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by a reference procedure such as chance, climatology, or persistence (Murphy and Daan 1985). Many different skill scores have been formulated over the last 100 years (e.g., see Daan 1984; Murphy and Daan 1985; Stanski et al. 1989; Woodcock 1976). For example, skill scores have been defined for different types of variables (i.e., continuous, discrete) and/or different types of forecasts (i.e., categorical, probabilistic). In addition, skill scores with particular properties have been designed for various applications, such as situations involving rare events or situations in which the forecasts are produced with specific users in mind.

Although it is not widely recognized, skill scores for forecasts of variables defined in terms of categories (or events) are based on scoring matrices. A scoring matrix is a square array of numbers that assigns a score (or weight) to each possible combination of forecast and observed events. For example, the Heidke skill score (Heidke 1926) measures the accuracy of forecasts relative to the accuracy of random (or chance) forecasts, and the measure of accuracy employed in conjunction with this skill score is the frequency (or relative frequency) of correct forecasts. In applying this measure, all correct forecasts (complete correspondence between forecast and observed events) are assigned a score of 1 and all incorrect forecasts (lack of complete correspondence between forecast and observed events) are assigned a score of 0. Thus, all correct forecasts are weighted equally regardless of the relative frequencies of occurrence of the respective events, and all incorrect forecasts are weighted equally regardless of their respective degrees of incorrectness.

All scoring matrices are not equally suitable or appropriate, even from a purely meteorological point of view. In a two-event situation, for example, it may be reasonable to assign correct forecasts of the events identical scores when these events are approximately equally likely (in a climatological sense), but identical

Corresponding author address and permanent affiliation: Dr. Allan H. Murphy, Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331-2209.

scores do not appear to be appropriate when one event is much more likely than the other event. In these "unbalanced" situations, the use of a skill score based on an inappropriate scoring matrix may encourage forecasters to exhibit a preference for some events at the expense of other events. (This statement should *not* be interpreted to mean that forecasters are exhibiting inappropriate behavior. It is incumbent on the designers of verification systems to develop measures of performance that do not reward forecasters for making forecasts that differ from their best judgments.) Skill scores based on such scoring matrices can be said to be *inequitable.* Inequitable skill scores can lead to forecasts that possess various undesirable performance characteristics. (Examples of such skill scores are presented in section 3.)

This paper describes a method of formulating *equitable skill scores* for forecasts of variables defined in terms of two or more events. These skill scores discourage forecasters from exhibiting inappropriate preferences for some events at the expense of other events. In particular, constant forecasts of any particular event—as well as forecasts in which events are chosen at random—achieve the same expected score (which can be set equal to zero). Moreover, the scoring matrices associated with equitable skill scores can be shown to possess other desirable properties: (i) The scores assigned to correct forecasts of an event increase as the climatological probability of the event decreases and (ii) the scoring matrices associated with situations involving $n + 1$ events and n events may be made consistent, in the sense that the scoring matrix in the $n + 1$-event situation approaches the scoring matrix in the n-event situation as the climatological probability of one of the events approaches zero.

Basic definitions associated with performance and scoring matrices are introduced in section 2. Some examples of inequitable skill scores are presented and discussed in section 3, with particular attention given to a skill score used to evaluate long-range forecasts in the USSR. The basic concepts associated with the equitable skill scores introduced in this paper are described in section 4. Sections 5, 6, and 7 investigate and illustrate equitable skill scores in two-event, three-event, and n-event ($n > 3$) situations, respectively. Section 8 consists of a discussion and some concluding remarks.

2. Performance and scoring matrices: Basic definitions

We are concerned here with the evaluation of categorical (i.e., nonprobabilistic) forecasts of variables defined in terms of n (≥ 2) mutually exclusive and collectively exhaustive events (or categories). The underlying variables may be either nominal or ordinal. Ordinal variables consist of values (or events) that possess a natural ordering. For any two values of an ordinal variable (any two nonoverlapping events consisting of sets of such values), one value (event) is necessarily either larger or smaller than the other value (event). No such ordinal relationship exists in the case of events derived from nominal variables.

To describe the performance of a forecaster or forecasting system, it is necessary to specify the joint distribution of forecasts and observations (Murphy and Winkler 1987). In terms of a verification data sample for categorical forecasts of an n-event variable, this distribution can be represented by an $n \times n$ *performance matrix* $\mathbf{P} = (p_{ij})$ ($p_{ij} \geq 0$, $\Sigma_i \Sigma_j p_{ij} = 1$; $i, j = 1, \cdots, n$), where p_{ij} denotes the relative frequency of occasions on which the ith event is forecast and the jth event is observed. Moreover, let $\mathbf{p} = (p_j)$ represent the *climatological probability vector,* where $p_j = \Sigma_i p_{ij}$ ($j = 1, \cdots, n$) is the (sample) climatological probability of occurrence of the jth event, and let $\mathbf{q} = (q_i)$ represent the *predictive probability vector,* where $q_i = \Sigma_j p_{ij}$ ($i = 1, \cdots, n$) is the (sample) predictive probability of the ith forecast.

Let $\mathbf{S} = (s_{ij})$ represent the $n \times n$ *scoring matrix,* where s_{ij} denotes the score assigned to a forecast of event i when the jth event occurs. It will be assumed here that the elements of $\mathbf{S}$ are independent of the elements of $\mathbf{P}$. That is, it is assumed that the basic scores assigned to various combinations of forecast and observed events do not depend on the relative frequencies with which these combinations occur in the verification data sample. This assumption seems quite reasonable, in view of the desirability of separating the forecasting and scoring tasks. It should be noted that this assumption does *not* rule out the possibility that the elements of $\mathbf{S}$ may depend upon the elements of the climatological probability vector $\mathbf{p}$.

Finally, let S denote the expected score associated with performance matrix $\mathbf{P}$ and scoring matrix $\mathbf{S}$. Under the assumption that S is a linear combination of the elements of $\mathbf{S}$ we can write

$$S = \Sigma_i \Sigma_j p_{ij} s_{ij}. \tag{1}$$

Thus, S represents a weighted average of the s_{ij}, where the weights are the probabilities of the respective combinations of forecast and observed events. The expression in (1) can be used to determine the expected score associated with any performance matrix $\mathbf{P}$ including (for example) the performance matrix for constant forecasts of a particular event.

3. Inequitable skill scores: Some examples

To motivate the method presented in this paper, we first consider examples of inequitable skill scores. A good example of such a skill score is provided by the measure used to evaluate long-range forecasts of monthly precipitation totals in the USSR, which was officially approved by the Hydrometeorological Administration (Gandin 1977). The forecasts involved three events—below-normal, near-normal, and above-normal precipitation—and the events were defined in

such a way that their climatological probabilities were equal (i.e., $p_1 = p_2 = p_3 = 1/3$). A score of one was assigned to correct forecasts, a score of $1/2$ was assigned to forecasts involving one-category errors, and a score of 0 was assigned to forecasts involving two-category errors. That is, $s_{11} = s_{22} = s_{33} = 1$, $s_{12} = s_{21} = s_{23} = s_{32} = 1/2$, and $s_{13} = s_{31} = 0$.

Superficially, this scoring method seems quite logical and reasonable, and it appears to imply that a forecasting method is skillful when its expected score exceeds 0.50. However, random forecasts—that is, forecasts for which $p_{ij} (= q_i p_j) = 1/9$ $(i, j = 1, 2, 3)$—receive an expected score of $5/9$, and this result gives the impression that the zero point on the skill scale is 0.56. In reality, the situation is somewhat more unfavorable. It is easy to show that a forecaster who constantly forecasts the near-normal event (category 2) will receive an expected score of $2/3$ (or 0.67). Alternatively, a forecaster who predicts one of the anomalous events on each occasion will receive an expected score of $1/2$ (or 0.50). As a result, this scoring method encourages forecasters to predict the near-normal event more often than it occurs, at the expense of potentially successful forecasts of the other two events.

From a practical point of view, the influence of this inequitable skill score can be illustrated by considering the forecasts produced by two forecasting methods (denoted here by A and B) for monthly precipitation totals in April 1974 over 33 regions in the European territory of the USSR (see Table 1). Although the near-normal event occurred over a relatively small part of the territory, it was predicted by operational method A for many regions and by experimental method B for all regions. As a result, both methods were awarded relatively high scores; namely, 0.62 for A and 0.67 for B. More extensive statistics based on the application of these forecasting methods over a 4-yr period confirm this result. Both methods forecast category 2 about twice as often as it actually occurred (Gandin 1977).

TABLE 1. Forecast and observed monthly precipitation totals in three categories (1: below normal, 2: near normal, 3: above normal) in April 1974 for 33 regions in the European territory of the USSR (Gandin 1977). (*A*: forecast produced by method A, *B*: forecast produced by method B, *R*: observed).

Region	*A*	*B*	*R*
1	2	2	1
2	2	2	2
3	2	2	3
4	2	2	1
5	2	2	2
6	1	2	1
7	2	2	1
8	2	2	1
9	2	2	1
10	2	2	1
11	2	2	2
12	2	2	2
13	2	2	1
14	2	2	1
15	2	2	1
16	2	2	1
17	2	2	3
18	2	2	1
19	2	2	1
20	2	2	1
21	1	2	3
22	1	2	2
23	1	2	2
24	2	2	2
25	2	2	3
26	2	2	1
27	3	2	3
28	2	2	2
29	3	2	3
30	3	2	3
31	2	2	3
32	2	2	3
33	3	2	3

To provide some overall insight into the way in which such skill scores might influence forecasters, consider a forecaster who possesses a judgmental probability distribution over the categories in a three-event situation and who decides to forecast the category with the highest (subjective) expected score. Let $\mathbf{r} = (r_1, r_2, r_3)$ represent this distribution, where r_i denotes the judgmental probability of event i ($r_i \geq 0$, $\Sigma_i r_i = 1$; $i = 1, 2, 3$). Then $S_i = \Sigma_j r_j s_{ij}$ is the expected score associated with a forecast of the ith event. In this situation, the set of all possible judgmental probabilities can be represented geometrically by an equilateral triangle $R = \{(r_1, r_2, r_3): r_i \geq 0, \Sigma_i r_i = 1; i = 1, 2, 3\}$, in which the vertices represent the three possible categorical forecasts. Moreover, when the expected scores associated with the forecasts are set equal to each other in a pairwise manner (i.e., $S_1 = S_2$, $S_1 = S_3$, and $S_2 = S_3$), the triangle R is divided into three regions. Region R_i ($i = 1, 2, 3$) represents the set of all judgmental probabilities for which a categorical forecast of event i is optimal (in the sense of maximizing expected score).

The geometrical framework corresponding to the scoring matrix used to evaluate long-range forecasts in the USSR is depicted in Fig. 1. Note that region R_2 is much larger than regions R_1 and R_3. In fact, when $r_1 = r_3$, a forecast of category 2 is optimal for *all* nonzero r_2. Specifically, category 1 is the optimal forecast when $r_1 > 1/2$, category 2 is the optimal forecast when $r_1 < 1/2$ *and* $r_3 < 1/2$, and category 3 is the optimal forecast when $r_3 > 1/2$.

The fraction correct (FC), the measure of forecast accuracy underlying the Heidke skill score (and other skill scores), provides a second example of an inequitable performance measure. As noted in section 1, this measure is based on a scoring matrix equal to the identity matrix. That is, $s_{ij} = 1$ for $i = j$ and $s_{ij} = 0$ for $i \neq j$, and thus FC $= \Sigma_j p_{jj}$. Consider a two-event ($n = 2$) situation in which $p_1 = 0.1$ and $p_2 = 0.9$. In this situation, constant forecasts of categories 1 and 2 obtain expected scores of 0.1 and 0.9, respectively. Clearly, it

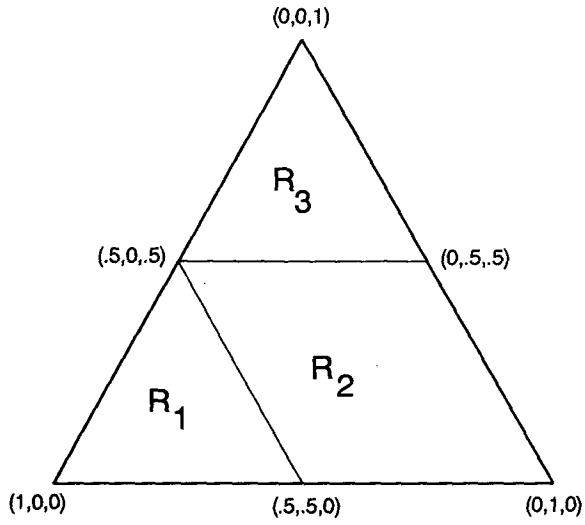

FIG. 1. Equilateral triangle $R = \{(r_1, r_2, r_3): r_i \geq 0, \Sigma_i r_i = 1; i = 1, 2, 3\}$ representing the set of all possible judgmental probabilities over three events. Region R_i denotes the set of judgmental probabilities for which a forecast of event i receives the highest expected score, based on the inequitable scoring matrix used to evaluate long-range forecasts in the USSR. See text for additional details.

is advantageous to forecast category 2 in this situation. In general, for performance measures based on the identity matrix, it is advantageous to forecast the category k for which $p_k = \max_j p_j$.

The threat score or critical success index, originally defined by Gilbert (1884) and denoted here by TS, provides a third example of an inequitable performance measure. In terms of the notation introduced in section 2, this measure can be expressed as $TS = p_{11}/(p_{11} + p_{12} + p_{21})$. Note that a constant forecast of category 1 yields an expected threat score of p_1, whereas a constant forecast of category 2 yields an expected threat score of 0. Thus, TS is inequitable and may encourage forecasters to make an undue number of category 1 forecasts. [For further discussion of the TS, see Doswell et al. (1990) and Schaefer (1990), as well as the references cited in these papers.]

These examples illustrate the principal shortcomings of inequitable skill scores. Such skill scores may not only lead to erroneous conclusions about the relative skill of different forecasters, but they also may encourage forecasters—and even the developers of forecasting methods—to produce forecasts that are biased toward one event (or several events) at the expense of other events. Thus, special conditions must be imposed on performance measures to ensure their equitability.

4. Equitable skill scores: Basic concepts

All skill scores possess a range of numerical values, with an explicit origin (zero point) and scale (unit of measurement). The choices of origin and scale are arbitrary, but some choices are more convenient than others. Here, we choose α as the origin (a numerical value for α will be specified later). Since it is assumed that the skill score S has a positive orientation (i.e., it is defined in such a way that larger scores are better), scores greater (less) than α denote positive (negative) skill.

The elements of the scoring matrix depend upon the basic relationships that define the concept of equitability (see below). Any linear transformation of these elements changes the origin and scale, but otherwise yields an equivalent scoring matrix. That is, the scoring matrices $\mathbf{S}$ and $\mathbf{S}'$ are said to be equivalent if $s'_{ij} = a + bs_{ij}$ $(i, j = 1, \cdots, n)$, where a and b are known constants. If S has a positive orientation, then S' has a positive (negative) orientation when $b > 0$ $(b < 0)$.

As noted in section 2, the underlying variable may be either nominal or ordinal. We impose certain conditions on the relative magnitudes of the elements of the scoring matrix, depending on the basic nature of the variable. For all variables (i.e., nominal *and* ordinal), it seems reasonable to require that $s_{ij} \leq s_{ii}$ and $s_{ij} \leq s_{jj}$ for all i and j $(i, j = 1, \cdots, n)$. That is, we require that the scores assigned to incorrect forecasts be less than or equal to the scores assigned to correct forecasts. In the 3×3 situation, this requirement implies that $s_{12} \leq s_{11}$, $s_{12} \leq s_{22}$, $s_{13} \leq s_{11}$, $s_{13} \leq s_{33}$, $s_{23} \leq s_{22}$, and $s_{23} \leq s_{33}$ (similar relationships are assumed to hold for the corresponding s_{ij} for which $i > j$).

When the underlying variable is ordinal, it seems reasonable to place an additional requirement on the scores (since the magnitude of the error is now a meaningful concept). Specifically, we require that $s_{i'j} \leq s_{ij}$ for $|i' - j| > |i - j|$ and $s_{ij'} \leq s_{ij}$ for $|i - j'| > |i - j|$. That is, the scores assigned to large errors are less than or equal to the scores assigned to small errors. In the 3×3 situation, this condition imposes the additional requirements that $s_{13} \leq s_{12}$ and $s_{13} \leq s_{23}$ (similar relationships are assumed to hold for the corresponding s_{ij} for which $i > j$).

The equitable skill scores formulated in this paper are based on the fundamental concept that constant forecasts of any event, as well as forecasts produced by a procedure in which a forecast event is chosen at random, should be accorded the same level of skill. This level of skill represents the zero point on the skill scale and is denoted here by α. Let S_i denote the expected score for a constant forecast of event i $(i = 1, \cdots, n)$. Thus, it is assumed here that

$$S_i = \Sigma_j p_j s_{ij} = \alpha \quad (i = 1, \cdots, n). \tag{2}$$

The assumption embodied in (2) implies that the expected score for random forecasts S_r is also equal to α. Specifically,

$$S_r = \Sigma_i \Sigma_j q_i p_j s_{ij}. \tag{3}$$

From (2), it follows that

$$S_r = \Sigma_i q_i S_i = \alpha. \tag{4}$$

Let S_p denote the expected score for perfect forecasts ($p_{jj} = p_j$ for all j). Specification of a value for S_p defines a scale for the expected scores. Here, we set

$$S_p = \Sigma_j p_j s_{jj} = \beta. \tag{5}$$

That is, the best possible expected score is equal to β.

When the relationship in (5) is added to the n relationships involving constant forecasts [embodied in (2)], a total of $n + 1$ relationships are available to determine the n^2 scores s_{ij} ($i, j = 1, \cdots, n$). Note that $n^2 > n + 1$ for all n ($n \geq 2$). Obviously, we must either increase the number of relationships or decrease the number of scores in order to obtain a unique solution. In this regard, no a priori reason exists to expect the relationships embodied in the concept of equitability to be sufficient by themselves to determine uniquely the elements of the scoring matrix. In general, these relationships represent a set of necessary but not sufficient conditions on the elements of **S**.

We are concerned here with the formulation of skill scores that are not related explicitly to any particular users or uses of the forecasts. With this consideration in mind, we pose the following specific question: What relationship should exist between elements s_{12} and s_{21} in the scoring matrix? For example, should the score when event 1 is forecast and event 2 is observed be greater (or less) than the score when event 2 is forecast and event 1 is observed? In view of the fact that we are concerned with performance in a purely meteorological sense, it seems reasonable to assume that $s_{12} = s_{21}$. Thus, the following general assumption is made regarding the structure of the scoring matrix **S**:

$$s_{ji} = s_{ij} \quad (i, j = 1, \cdots, n). \tag{6}$$

That is, we assume that **S** is symmetric, in the sense that the score assigned to a forecast of the ith event when the jth event occurs is the same as the score assigned to a forecast of the jth event when the ith event occurs. Obviously, such an assumption generally would *not* be appropriate in the case of user-related scores.

The assumption that **S** is symmetric reduces the number of scores that must be determined from n^2 to $n(n+1)/2$. Note that $n(n+1)/2 \geq n+1$, with equality only when $n = 2$. Thus, the $n + 1$ relationships embodied in (2) and (5) are necessary conditions for the determination of the $n(n + 1)/2$ scores, but they represent sufficient conditions only in the two-event situation.

Another feature of the results produced by the method introduced here is that the scoring matrices are consistent between situations of different dimensionality. Specifically, the scoring matrix for a particular 3×3 situation may be made to approach the scoring matrix for the corresponding 2×2 situation when the climatological probability of one of the events approaches zero. This type of consistency between scoring matrices in the context of two-event and three-event situations is discussed briefly in sections 5 and 6, respectively.

5. Equitable skill scores: Two-event situation

In two-event ($n = 2$) situations,

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \tag{7}$$

$\mathbf{p} = (p_1, p_2)$, $\mathbf{q} = (q_1, q_2)$, and

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \tag{8}$$

(the assumption of a symmetric scoring matrix implies that $s_{21} = s_{12}$). The expected-score relationships in this situation are

$$S_1 = p_1 s_{11} + p_2 s_{12} = \alpha, \tag{9}$$

$$S_2 = p_1 s_{12} + p_2 s_{22} = \alpha, \tag{10}$$

and

$$S_p = p_1 s_{11} + p_2 s_{22} = \beta \tag{11}$$

($p_1 + p_2 = 1$). It is convenient here—and throughout this paper—to set $\alpha = 0$ and $\beta = 1$; that is, we take 0 as the origin and 1 as the unit of measurement on the skill scale.

The system of three equations in (9)–(11) contains three unknowns: s_{11}, s_{12}, and s_{22}. Thus, this system possesses a unique solution, which is (with $\alpha = 0$ and $\beta = 1$)

$$s_{11} = p_2/p_1, \tag{12}$$

$$s_{12}\,(=s_{21}) = -1, \tag{13}$$

and

$$s_{22} = p_1/p_2. \tag{14}$$

It is interesting to note that the scores for correct forecasts of events 1 and 2 are equal to the climatological odds against these events; namely, p_2/p_1 and p_1/p_2, respectively. In this context in which forecasting performance is measured in a purely meteorological sense, the score for both types of incorrect forecasts is equal to -1.

As an example, consider a situation in which $p_1 = 0.05$. Thus, category 1 represents a relatively rare event and category 2 represents a relatively common event. Then $s_{11} = 0.95/0.05 = 19$ and $s_{22} = 0.05/0.95 = 1/19 = 0.053$. In this situation, the forecaster receives a score of 19 for a correct forecast of the rare event, a score of 0.053 for a correct forecast of the common event, and a score of -1 for an incorrect forecast. By definition, constant forecasts of category 1 or category 2, as well as random forecasts, receive an expected score of 0, and perfect forecasts receive an expected score of 1. Note that, although the maximum expected score is

1, individual elements in the scoring matrix can be greater than 1.

The expected score S associated with the scoring matrix defined by (12)–(14) can be written as follows:

$$S = p_{11}(p_2/p_1) + p_{12}(-1) + p_{21}(-1) + p_{22}(p_1/p_2) \quad (15)$$

[see (1)], or

$$S = (p_{11}p_{22} - p_{12}p_{21})/(p_1p_2). \quad (16)$$

It is interesting to note that S in (16) is identical to a score originally defined by Peirce (1884) and frequently identified today as Kuipers' performance index (see Murphy and Daan 1985). This expected score is also linearly related to a measure of skill S' defined by Gringorten (1967). In Gringorten's formulation, $s_{11} = 1/p_1$, $s_{12} = s_{21} = 0$, and $s_{22} = 1/p_2$, and, as a result, $S' = S + 1$. In summary, the skill score S in (16)—and monotonic transformations of this measure—are the only performance measures in the two-event situation that satisfy the basic conditions of equitability set forth in this paper.

In the two-event situation, it is relatively easy to describe the impact of **S** on the process of translating a judgmental probability distribution $\mathbf{r} = (r_1, r_2)$ into a categorical forecast. Here, the set of all judgments is the unit line segment $R = \{(r_1, r_2): r_i \geq 0, \Sigma_i r_i = 1; i = 1, 2\}$, and forecasters who want to maximize their expected scores predict event 1 (2) when r_1 (r_2) > p_1 (p_2). Thus, the critical threshold for translating judgments into categorical forecasts in the two-event situation is simply the climatological probability.

A special 2 × 2 situation arises when $p_1 = p_2 = 1/2$. In this situation, $s_{11} = s_{22} = 1$ and $s_{12} = s_{21} = -1$ [see (12)–(14)]. Thus, correct forecasts of the two events are assigned the same score when the climatological probabilities of the events are equal.

The scoring matrix in this special situation can be shown to be equivalent to the identity matrix. Specifically, if S_1 and S_2 in (9) and (10), respectively, are set equal to 1/2 (i.e., $\alpha = 1/2$ instead of $\alpha = 0$), then $s_{11} = s_{22} = 1$ and $s_{12} = s_{21} = 0$. This result implies that performance measures based on the identity matrix (see section 3) are equitable in the two-event situation only when $p_1 = p_2$.

6. Equitable skill scores: Three-event situation

In the three-event ($n = 3$) situation,

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}, \quad (17)$$

$\mathbf{p} = (p_1, p_2, p_3)$, $\mathbf{q} = (q_1, q_2, q_3)$, and

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix} \quad (18)$$

($s_{21} = s_{12}$, $s_{31} = s_{13}$, and $s_{32} = s_{23}$). In this situation, the expected-score relationships are (with $\alpha = 0$ and $\beta = 1$)

$$S_1 = p_1s_{11} + p_2s_{12} + p_3s_{13} = 0, \quad (19)$$

$$S_2 = p_1s_{12} + p_2s_{22} + p_3s_{23} = 0, \quad (20)$$

$$S_3 = p_1s_{13} + p_2s_{23} + p_3s_{33} = 0, \quad (21)$$

and

$$S_p = p_1s_{11} + p_2s_{22} + p_3s_{33} = 1 \quad (22)$$

($p_1 + p_2 + p_3 = 1$).

The system of four equations in (19)–(22) contains six unknowns: s_{11}, s_{12}, s_{13}, s_{22}, s_{23}, and s_{33}. In general, a solution for this system of equations requires either that two additional relationships be imposed or that values be specified for two of the scores. To provide insight into the solution procedure for the general 3 × 3 problem, we consider first a special and simpler situation in which the climatological probabilities of events 1 and 3 are assumed to be equal (i.e., $p_1 = p_3$). Forecasts on monthly and seasonal time scales for below-normal, near-normal, and above-normal temperatures provide examples of forecasts that possess this characteristic.

a. Special situation

When (i) $p_1 = p_3$, (ii) the underlying variable is ordinal, and (iii) the categories are defined in such a way that events 1 and 3 can be considered to be equally distant from event 2, it follows that $s_{23} = s_{12}$ and $s_{33} = s_{11}$. As a result, the number of unknown scores is reduced from six to four. However, $S_1 = S_3$ under these conditions [cf. (19) and (21)], so that the number of independent expected-score relationships is reduced from four to three; namely, the relationships embodied in (19), (20), and (22), respectively (with $s_{23} = s_{12}$ and $s_{33} = s_{11}$). Thus, one score must be specified in order to obtain a unique solution to the relevant system of equations. Specifically, we set $s_{12} = k$. The solution is then

$$s_{11}\ (=s_{33}) = (1 + 2kp_1)/(1 - p_2), \quad (23)$$

$$s_{13} = -[1 + 2k(1 - p_1)]/(1 - p_2), \quad (24)$$

and

$$s_{22} = -2kp_1/p_2. \quad (25)$$

Since the underlying variable is assumed to be ordinal in this situation, we require that $s_{12} \leq s_{22}$ and $s_{13} \leq s_{12}$. These relationships, together with the assumption that $p_1 = p_3$, imply that $-1/2 \leq k \leq 0$. (If the underlying variable is nominal instead of ordinal, then the relevant relationships are $s_{12} \leq s_{22}$ and $s_{13} \leq s_{11}$. In this situation, it follows that $-1 \leq k \leq 0$.) Moreover, since $p_2 = 1 - 2p_1$, the unknown scores s_{11}, s_{13}, and s_{22} depend only on p_1 and $k\ (=s_{12})$. The values of these scores are

shown in Fig. 2 as functions of p_1 for selected values of $k\,(=s_{12})$. Note that the ranges of values of the scores generally are quite limited, except in the cases of s_{13} and s_{22} for small and large values of p_1, respectively. The score s_{11} is restricted to a particularly narrow range of values. These restricted ranges indicate that the structure of the scoring matrix **S** is less sensitive to the choice of a value for k than might have been expected a priori.

With regard to the extreme values of k, $s_{11} = s_{33} = (1 - p_1)/2p_1$, $s_{12} = s_{13} = s_{23} = -1/2$, and $s_{22} = p_1/(1 - 2p_1)$ when $k = -1/2$. In this case, the values of s_{12}, s_{13}, and s_{23} are equal, a limiting case with respect to the structure of scoring matrices for forecasts of ordinal variables. When $k = 0$, $s_{11} = s_{33} = 1/2\,p_1$, $s_{12} = s_{22} = s_{23} = 0$, and $s_{13} = -1/2\,p_1$. In this case, s_{12}, s_{22}, and s_{23} are equal, another limiting case with respect to the structure of the relevant scoring matrices.

As a specific example, consider the case in which $k = -1/4$ (the midpoint of its range of values). In this case, $s_{11} = s_{33} = (2 - p_1)/4p_1$, $s_{12} = s_{23} = -1/4$, $s_{22} = p_1/2p_2$, and $s_{13} = -(1 + p_1)/4p_1$. Suppose that $p_1\,(=p_3) = p_2 = 1/3$ (equally likely events). Then

$$\mathbf{S} = \left(\frac{1}{24}\right)\begin{pmatrix} 30 & -6 & -24 \\ -6 & 12 & -6 \\ -24 & -6 & 30 \end{pmatrix}. \qquad (26)$$

Correct forecasts of events 1 and 3 are assigned a score of $30/24$ $(=5/4)$, whereas a correct forecast of event 2 is assigned a score of $12/24$ $(=1/2)$. As in the 2×2 situation (see section 5), individual scores can be greater than 1 (or less than -1; see below). Incorrect forecasts involving one-category and two-category errors receive scores of $-6/24\,(=-1/4)$ and $-24/24\,(=-1)$, respectively.

Alternatively, suppose that $p_1\,(=p_3) = 0.3$ and $p_2 = 0.4$, a format consistent with the National Weather Service's monthly and seasonal forecasts (e.g., Epstein 1988). Then

$$\mathbf{S} = \left(\frac{1}{24}\right)\begin{pmatrix} 34 & -6 & -26 \\ -6 & 9 & -6 \\ -26 & -6 & 34 \end{pmatrix}. \qquad (27)$$

Comparison of matrices in (26) and (27) reveals that the scores for correct forecasts of the anomalous events increased (from $15/12$ to $17/12$) and that for the near-normal event decreased (from $4/8$ to $3/8$). The score for forecasts involving two-category errors decreased slightly (from -1 to $-13/12$).

The influence of the equitable skill score associated with the scoring matrix in (27) on a forecaster's decisions regarding optimal categorical forecasts is described geometrically in Fig. 3. In this case, the region R_2 corresponding to a categorical forecast of event 2 is much smaller than that associated with the scoring matrix used to evaluate long-range forecasts in the USSR (cf. Fig. 1). Specifically, event 1 is forecast when $r_1 > r_3$ *and* $r_1 > (1/12)(4 - r_2)$, event 2 is forecast when $r_2 > 4(1 - 3r_1)$ *and* $r_2 > 4(1 - 3r_3)$, and event 3 is forecast when $r_3 > r_1$ *and* $r_3 > (1/12)(4 - r_2)$.

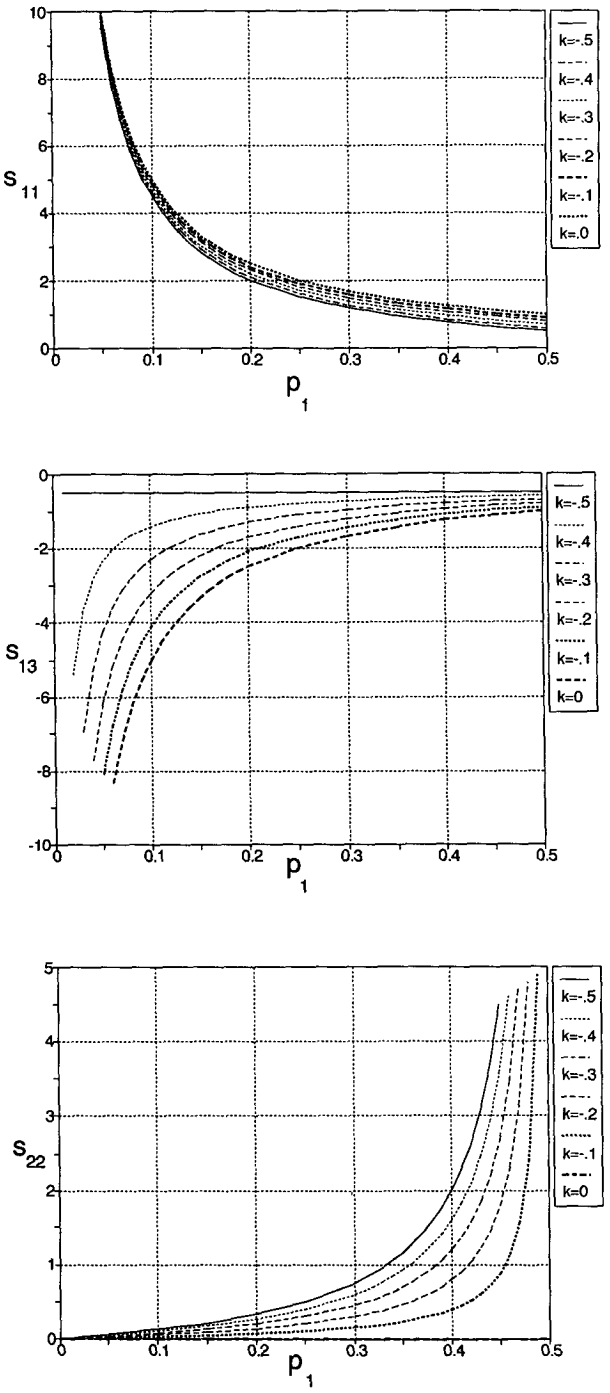

FIG. 2. Numerical values of the elements in the scoring matrix in the special 3×3 situation as a function of the climatological probability p_1 for selected values of the specified element $k\,(=s_{12})$: (top) s_{11}, (middle) s_{13}, (bottom) s_{22}.

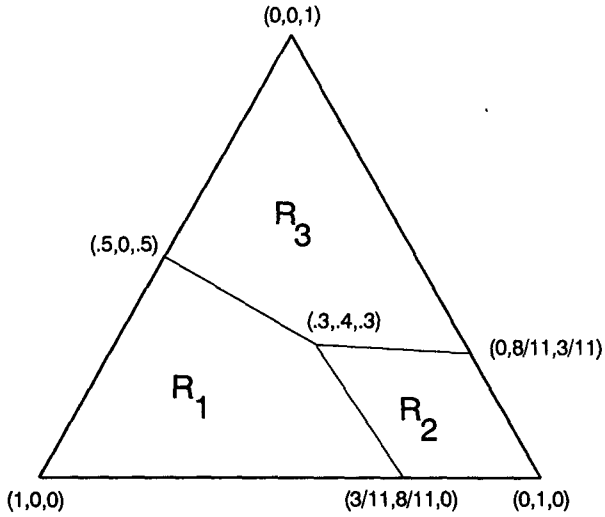

FIG. 3. Same as Fig. 1, except based on the equitable scoring matrix defined by (27).

It is interesting to note that the special case of the 3 × 3 situation considered here can be made to approach the special case of the 2 × 2 situation when p_2 approaches 0. Under these circumstances (i.e., p_2 approaching 0) p_1 approaches 1/2. In order to achieve this consistency, a relationship must exist between p_2 and k such that k approaches 0 as p_2 approaches 0. Note that when $k = 0$, $s_{11} = s_{33} = 1/2p_1$, $s_{12} = s_{23} = s_{22} = 0$, and $s_{13} = -1/2p_1$. Since $p_1 = p_3 = 1/2$ in this case, it follows that $s_{11} = s_{33} = 1$, $s_{13} = s_{31} = -1$, and all other scores are equal to 0. Thus, in order to make the scoring matrices in the 3 × 3 and 2 × 2 situations consistent, it is necessary to consider $k\ (=s_{12})$ as a function of p_1. An analogous situation exists in the general 3 × 3 problem (see section 6b).

b. General situation

As noted previously, it is necessary to specify two scores in the general 3 × 3 situation. Here, we set $s_{12} = k_1$ and $s_{23} = k_2$. The solution is then

$$s_{11} = [p_3 + p_1(p_3 - p_2)k_1 + p_3(p_2 + p_3)k_2]/[p_1(p_1 + p_3)], \quad (28)$$

$$s_{13} = -[1 + (p_1 + p_2)k_1 + (p_2 + p_3)k_2]/(p_1 + p_3), \quad (29)$$

$$s_{22} = -(p_1k_1 + p_3k_2)/p_2, \quad (30)$$

and

$$s_{33} = [p_1 + p_1(p_1 + p_2)k_1 + p_3(p_1 - p_2)k_2]/[p_3(p_1 + p_3)]. \quad (31)$$

Once again, the nature of the underlying variable (nominal or ordinal), and its implications with respect to the relative magnitudes of the elements of the scoring matrix **S** lead to restrictions on the ranges of values of $k_1\ (=s_{12})$ and $k_2\ (=s_{23})$. These restrictions, as well as the relationship between k_1 and k_2, are illustrated in Fig. 4 for the case in which $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. The square in this figure defined by the points (0, 0), (0, −1/2), (−1/2, −1/2), and (−1/2, 0) defines the basic set of independent pairs of values for these two scores. The quadrilateral with oblique angles at (1/4, −1) and (−1, 1) identifies the set of permissible pairs of values of k_1 and k_2 when the underlying variable is ordinal. It also should be noted that, when the climatological probabilities (p_1, p_2, and p_3) are known, specification of values for k_1 and k_2 determines the values of the other four scores (i.e., s_{11}, s_{13}, s_{22}, and s_{33}).

As a numerical example, we consider the situation in which $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. In this situation, $s_{11} = (4 - k_1 + 2k_2)/7$, $s_{13} = -(10 + 8k_1 + 5k_2)/7$, $s_{22} = -(5k_1 + 2k_2)/3$, and $s_{33} = (25 + 20k_1 + 2k_2)/7$. Specifically, when $k_1 = -1/2$ and $k_2 = -1/4$,

$$\mathbf{S} = \left(\frac{1}{28}\right)\begin{pmatrix} 16 & -14 & -19 \\ -14 & 28 & -7 \\ -19 & -7 & 58 \end{pmatrix}. \quad (32)$$

Note that a correct forecast of event 3 is assigned a score more than twice that of a correct forecast of event 2 and more than 3.5 times greater than that of a correct forecast of event 1 (recall that the climatological probabilities of these events are 0.5, 0.3, and 0.2, respectively).

As in the case of the special situations (see section 6a), consistency should exist between the general 3

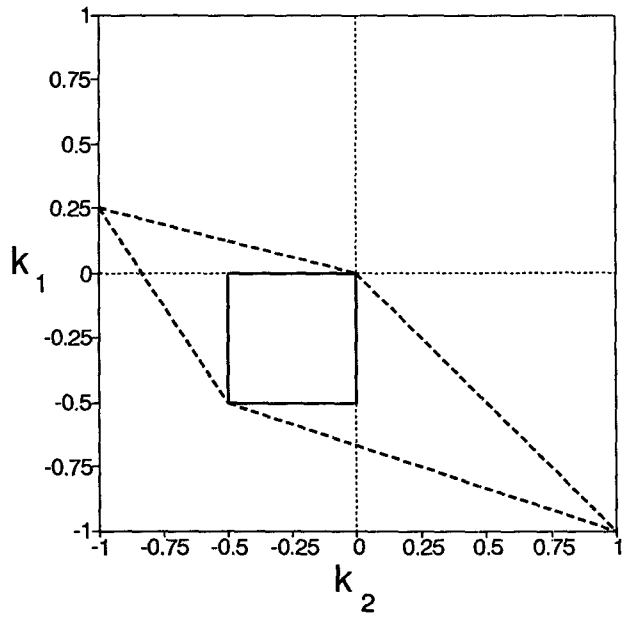

FIG. 4. Acceptable domains for numerical values of the specified elements $k_1\ (=s_{12})$ and $k_2\ (=s_{23})$ in the general 3 × 3 situation. See text for additional details.

$\times$ 3 and 2 $\times$ 2 situations. In particular, when the probability p_1 (or p_2 or p_3) approaches 0, the scoring matrix in the 3 $\times$ 3 situation should approach the corresponding scoring matrix in the 2 $\times$ 2 situation. Realization of these limiting conditions requires that linear combinations of the remaining probabilities (e.g., p_2 and p_3 in the case of p_1 approaching zero) and the specified scores s_{12} ($=k_1$) and s_{23} ($=k_2$) satisfy certain constraints.

7. Equitable skill scores: *n*-event situation

In this section we briefly describe the procedure that would be used to determine the elements of the scoring matrix **S** in the general n-event situation. In this situation, **S** is a symmetric matrix with $n(n + 1)/2$ distinct scores. The set of $n + 1$ relationships consists of n expected-score relationships for constant forecasts, $S_i = \Sigma_j p_j s_{ij} = 0$ ($i = 1, \cdots, n$), and the expected-score relationship for perfect forecasts, $S_p = \Sigma_j p_j s_{jj} = 1$ (assuming $\alpha = 0$ and $\beta = 1$). Since $n(n + 1)/2 > n + 1$ ($n > 2$), it follows that $(n + 1)(n - 2)/2$ scores must be specified. After these scores have been specified (which generally will require some analysis of the permissible ranges of numerical values of the scores), the system of $n + 1$ equations can be solved for the values of the remaining $n + 1$ scores in terms of the p_j ($j = 1, \cdots, n$).

For example, 15 scores must be determined in the situation in which $n = 5$. Since six expected-score relationships exist in such a situation, nine scores must be specified. After the values of these nine scores have been specified, the six equations can be solved for the remaining six scores in terms of the p_j ($j = 1, \cdots, 5$).

In some cases, special conditions can be invoked to reduce the number of scores that must be determined. For example, in situations involving equally spaced categories for ordinal variables and symmetric climatological distributions, it follows that $s_{ij} = s_{ji}$ ($i, j = 1, \cdots, n; j > i$) and $s_{11} = s_{nn}$, $s_{22} = s_{n-1,n-1}$, etc. As a result, the number of scores is $(3n - 1)/2$ if n is odd and $(3n - 2)/2$ if n is even. Since only $n - 1$ independent expected-score relationships exist under these conditions, the number of scores to be specified equals $(n + 1)/2$ if n is odd and $n/2$ if n is even.

8. Discussion and conclusion

This paper has described a method of formulating skill scores that is based explicitly on the concept of a scoring matrix. A scoring matrix assigns scores to the various combinations of forecast and observed events. These skill scores are equitable in the sense that they assign all constant forecasts, as well as forecasts based on a random choice of the forecast event, the same score. By appropriate choices of origin and scale, this method produces skill scores that yield expected scores of zero for constant and random forecasts and an expected score of one for perfect forecasts. In contrast to many existing skill scores (and other performance measures), equitable skill scores do *not* encourage forecasters to favor forecasts of one (or more) events at the expense of other events.

The method described here can be applied to categorical forecasts of both nominal and ordinal variables and yields scoring matrices—and skill scores—that possess considerable intuitive appeal. Specifically, the score assigned to correct forecasts of events increases as the relative frequency of the events decreases and the score assigned to incorrect forecasts decreases as the error in the forecasts increases (in the case of forecasts of ordinal variables). The fact that the scores are sensitive to the climatological probabilities of the events suggests that equitable skill scores may be useful in rare-event situations in which it is important to encourage (or not to discourage) forecasts of these events and to reward correct forecasts of such events appropriately.

Some practical problems may be encountered in implementing the equitable skill score methodology introduced in this paper, especially in situations involving a moderate or large number of events. In particular, the assumptions on which such skill scores are based represent necessary but generally not sufficient conditions for the unique determination of the elements of the relevant scoring matrices. As a result, some arbitrariness (or degrees of freedom) exists in assigning numerical values to specific scores. However, an initial analysis of this problem (see section 6) reveals that the scores are not as sensitive to these assignments as might have been expected a priori. In any case, this ambiguity can be resolved by choosing representative values of the relevant scores after conducting some exploratory analyses of their respective ranges of permissible values. Moreover, these degrees of freedom provide the evaluator with an opportunity to tailor the scoring matrix—and the associated skill score—to the specific situation at hand.

Several directions for future work related to equitable skill scores can be readily identified. For example, it would be desirable to explore alternative methods of reducing the arbitrariness associated with scoring matrices in multiple-event situations (e.g., by introducing additional relationships or conditions on the elements of the scoring matrices, by making use of consistency relationships between scoring matrices associated with situations of different dimensionality). In a different vein, by making the assumption that the scoring matrices of interest are symmetric, attention has been focused here on equitable skill scores as measures of forecasting performance in a purely meteorological sense. If this assumption is relaxed, it might be possible to formulate a class of equitable skill scores that could serve as measures of forecasting performance in a general user-related sense. Of course, relaxation of this assumption would lead to an increase in the number of

scores that must be specified a priori. It would also be interesting to investigate the use of equitable scoring matrices as a means of transforming probabilistic forecasts into categorical forecasts in those contexts in which such transformations are required. Finally, the possible extension of the concept of equitable skill scores to forecasts of continuous variables also warrants exploration in the future.

Acknowledgments. This work was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8714108. H. Daan, E. S. Epstein, I. B. Mason, and an anonymous reviewer provided helpful comments on earlier versions of this paper. W. G. Collins is due special thanks for kindly and expertly producing the final line diagrams.

REFERENCES

Daan, H., 1984: Scoring rules in forecast verification. Geneva, Switzerland, World Meteorological Organization, Report, 60 pp.

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.

Epstein, E. S., 1988: Long-range weather prediction: Limits of predictability and beyond. *Wea. Forecasting,* **3,** 69–75.

Gandin, L. S., 1977: On methodologically unbiased estimates of the successfulness of three-category forecasts. *Trudy, Main Geophys. Observ.,* No. 397, 130–136. (In Russian)

Gilbert, G. F., 1884: Finley's tornado predictions. *Amer. Meteor. J.,* **1,** 166–172.

Gringorten, I. I., 1967: Verification to determine and measure forecasting skill. *J. Appl. Meteor.,* **6,** 742–747.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geogr. Ann.,* **8,** 301–349. (In German)

Murphy, A. H., 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.

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

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

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

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.