Scalar Measures of Performance in Rare-Event Situations

CAREN MARZBAN

National Severe Storms Laboratory, Cooperative Institute for Mesoscale and Meteorological Studies, and Department of Physics, University of Oklahoma, Norman, Oklahoma

(Manuscript received 29 February 1996, in final form 10 December 1997)

ABSTRACT

A set of 14 scalar, nonprobabilistic measures—some old, some new—is examined in the rare-event situation. The set includes measures of accuracy, association, discrimination, bias, and skill. It is found that all measures considered herein are inequitable in that they induce under- or overforecasting. One condition under which such bias is not induced (for some of the measures) is when the underlying class-conditional distributions are Gaussian (normal) and equivariant.

1. Introduction

Forecast quality has been extensively examined by Murphy (1991, 1993). One lesson that emerges from those considerations is that forecast quality, or the performance of a forecaster or of an algorithm, is an inherently multifaceted quantity. In other words, although it is quite common to express performance in terms of a single, scalar (i.e., one-dimensional) quantity (e.g., fraction correct, the critical success index, etc.), such considerations are apt to be incomplete. A complete and faithful analysis must consider all the various components of performance quality.

As argued by Murphy and Winkler (1987), one quantity that encapsulates all the components of performance is the joint probability of observations, x, and forecasts, f. When x and f are discrete, the joint probability can be represented as a contingency table. For example, if the observations consist of the existence or the nonexistence of tornados, then the number of rows in the contingency table is 2. Additionally, if the forecasts are probabilities given in intervals of 10%, then the contingency table is 2×11, and if the forecasts are binary (yes/no), then it is 2×2. In the present article, only the 2×2 case is considered. In other words, both the observations and the forecasts are assumed to be binary.

Not withstanding the multidimensionality of performance, there exist situations in which this multidimensionality must be distilled to a single, scalar quantity. For example, in deciding the winner of a forecasting contest, this multidimensionality allows for multiple first-place winners; different first-place winners may excel one another in terms of different components of performance. As a result, even in probabilistic forecasting contests, performance is gauged in terms of some scalar quantity such as the ranked probability score (Hamill and Wilks 1995). Of course, it is possible that a unique candidate may outperform all of the other candidates in terms of all the different components of performance, or that the particular component of performance that is of interest is unambiguously self-evident. However, neither situation is guaranteed, or even likely.

For this and other reasons, scalar measures of performance are in common use. A number of these measures are derived from the contingency table itself, but at least two measures of performance are required to account for the two degrees of freedom present in the (2×2) contingency table (see next section). As mentioned above, however, frequently it is impossible to optimize both measures simultaneously. For example, it is known that the critical success index is ''inequitable'' (Gandin and Murphy 1992) in that it can induce ''hedging.'' Another way of saying this is that the critical success index and bias cannot be optimized simultaneously, that is, that the maximum of the critical success index does not correspond to unbiased (bias $= 1$) forecasts. It has also been argued (Doswell et al. 1990) that the true skill score can induce similar hedging in rare-event situations while Heidke's skill score does not. Indeed, R. L. Vislocky (1997, personal communication) has claimed that ''all'' measures are generally inequitable. In this article, 14 scalar measures based on the 2×2 contingency table will be examined in the rare-event situation. It will be shown that forecasts that optimize any single one of these measures are generally biased in a rare-event situation and can, therefore, be said to induce hedging or be inequitable. Although the concept of hedging, as put forth by Murphy and Epstein (1967), relates to probabilistic forecasts and scoring rules, these

Corresponding author address: Dr. Caren Marzban, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: marzban@gump.nssl.noaa.gov

measures do induce under- or overforecasting in a rare-event situation.

2. Measures of performance quality

The question as to what exactly is a proper measure of performance quality has been addressed extensively in the past (Brooks and Doswell 1996; Gandin and Murphy 1992; Murphy 1993, 1996; Murphy and Winkler 1992; Murphy and Winkler 1987). In this section, 14 measures of categorical forecast performance will be defined. The measures are derived from the contingency table (otherwise known as the confusion matrix), or in short the C table:

$$\text{C table} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \#\text{ of 0's predicted as 0} & \#\text{ of 0's predicted as 1} \\ \#\text{ of 1's predicted as 0} & \#\text{ of 1's predicted as 1} \end{pmatrix} = \begin{pmatrix} . & \text{false alarms} \\ \text{misses} & \text{hits} \end{pmatrix}.$$

The total number of nonevents (0's) is given by $N_0 = a + b$, that of events (1's) is $N_1 = c + d$, and the total sample size is $N = N_0 + N_1$. Note that this table has only two degrees of freedom; a general 2×2 matrix has four degrees of freedom, but with the two constraints $N_0 = a + b$ and $N_1 = c + d$, that number is reduced to 2. Two common quantities, probability of detection (POD) and false alarm ratio (FAR), are easily calculated as $\text{POD} = d/(c + d)$ and $\text{FAR} = b/(b + d)$. It is, however, convenient to write all of the measures in terms of the two error *rates*—the rate at which 0's are misclassified as 1's, $c_{01} = b/N_0$, and the rate at which 1's are misclassified as 0's, $c_{10} = c/N_1$. Therefore,

$$\text{POD} = 1 - c_{10}, \qquad \text{FAR} = \frac{c_{01}}{c_{01} + N_{10}(1 - c_{10})},$$

where N_{10} is simply the ratio of the sample sizes, $N_{10} = N_1/N_0$.

Specifically, the measures analyzed are[1]

1) product of POD and (1−FAR),

$$\text{PRD} = \text{POD} \times (1 - \text{FAR}) = \frac{(1 - c_{10})^2}{1 - c_{10} + N_{01}c_{01}};$$

2) average of POD and (1−FAR),

$$\text{AVG} = [\text{POD} + (1 - \text{FAR})]/2 = \frac{1}{2}(1 - c_{10})\left(1 + \frac{1}{1 - c_{10} + N_{01}c_{01}}\right);$$

3) fraction correct,

$$\text{FRC} = \frac{a + d}{a + b + c + d} = \frac{1 - c_{01}}{1 + N_{10}} + \frac{1 - c_{10}}{1 + N_{01}};$$

4) efficiency,

$$\text{EFF} = \frac{a}{a + b} \times \frac{d}{c + d} = (1 - c_{01})(1 - c_{10});$$

5) critical success index,

$$\text{CSI} = \frac{d}{b + c + d} = \frac{1 - c_{10}}{1 + N_{01}c_{01}};$$

6) true skill score,

$$\text{TSS} = \frac{\det}{N_0}\frac{C}{N_1} = 1 - c_{01} - c_{10};$$

7) Heidke's skill score,

$$\text{HSS} = \frac{2\det C}{N_0(b + d) + N_1(a + c)} = \frac{2(1 - c_{01} - c_{10})}{2 - (1 - N_{01})c_{01} - (1 - N_{10})c_{10}};$$

8) Gilbert's skill score,

$$\text{GSS} = \frac{\det C}{\det C + N(b + c)} = \frac{1 - c_{01} - c_{10}}{1 + N_{01}c_{01} + N_{10}c_{10}};$$

9) Clayton's skill score,

$$\text{CSS} = \frac{\det C}{(a + c)(b + d)} = \frac{1 - c_{01} - c_{10}}{(1 - c_{01} + N_{10}c_{10})(1 - c_{10} - N_{01}c_{01})};$$

10) Doolittle's skill score,

$$\text{DSS} = \frac{(\det C)^2}{N_0N_1(a + c)(b + d)} = \frac{(1 - c_{01} - c_{10})^2}{(1 - c_{01} - N_{10}c_{10})(1 - c_{10} + N_{01}c_{01})};$$

11) discrimination measure,

[1] None of the measures considered here allows for assigning specific costs of misclassification; for that purpose one must construct a scoring matrix reflecting the desired costs of misclassification.

$$\mathrm{DIS} = \left(\frac{1}{1 + N_{10}}\right)^2 \times \left[1 + \frac{2(1 - c_{10})}{1 + N_{01}}\left(1 + \frac{N_{10}(1 - C_{10})}{c_{01}}\right)\right] + \left(\frac{1}{1 + N_{01}}\right)^2 \times \left[1 + \frac{2(1 - c_{01})}{1 + N_{10}}\left(1 + \frac{N_{01}(1 - c_{01})}{c_{10}}\right)\right],$$

$$\mathrm{DIS} = \left(\frac{1}{1 + N_{10}}\right)^2 \times \left[1 + \frac{2c_{10}}{1 + N_{01}}\left(1 + \frac{c_{10}}{N_{01}(1 - c_{01})}\right)\right] + \left(\frac{1}{1 + N_{01}}\right)^2 \times \left[1 + \frac{2c_{01}}{1 + N_{10}}\left(1 + \frac{c_{01}}{N_{10}(1 - c_{10})}\right)\right],$$

for $ad - bc \sim (1 - c_{01} - c_{10}) \geq 0$, and $ad - bc < 0$, respectively. We also define two new measures—a pair of angles θ and ϕ:

12)

$$\theta = \frac{1}{2}\left|\tan^{-1}\frac{2(ab + cd)}{d^2 + b^2 - a^2 - c^2}\right| = \frac{1}{2}\left|\tan^{-1}\frac{2[(1 - c_{10})c_{10} + N_{01}^2 c_{01}(1 - c_{01})]}{(1 - 2c_{10}) - N_{01}^2(1 - 2c_{01})}\right|,$$

13)

$$\phi = \frac{1}{2}\left|\tan^{-1}\frac{2(ac + bd)}{d^2 + c^2 - a^2 - b^2}\right| = \frac{1}{2}\left|\tan^{-1}\frac{2(c_{01} + c_{10})}{N_{10}(1 - 2c_{10} + 2c_{10}^2) - N_{01}(1 - 2c_{01} + 2C_{01}^2)}\right|.$$

Finally, the bias of the forecasts will be gauged with 14) bias,

$$\mathrm{Bias} = \frac{b + d}{c + d} = 1 - c_{10} + N_{01}c_{01}.$$

In the above equations N_{01} stands for N_0/N_1. Unlike the other measures, θ and ϕ are measures of ''error'' in that lower values correspond to better performance. Although they, too, can be transformed into measures of ''success,'' as shown below, that would obfuscate their geometrical interpretation.

The quantities POD $\times$ (1 $-$ FAR) and [POD + (1 $-$ FAR)]/2 are natural choices to maximize, since optimal performance would correspond to the maximum of both POD and (1 $-$ FAR); both measures have been considered by Donaldson et al. (1975). The measure FRC is equal to Finley's measure (Murphy 1993, 1996). It is the measure of accuracy on which all of the skill scores are based, and it is the weighted average of the two group-specific fractions correct, a/N_0, and d/N_1. Efficiency is simply the product of the two group-specific fractions correct. This is a commonly used quantity in high energy detector physics. The unweighted average of the two is related to TSS: $(a/N_0 + d/N_1)/2 = (\mathrm{TSS} + 1)/2$. CSI (Donaldson et al. 1975) is an example of a measure with a long history and one that has been rediscovered many times (Murphy 1996). TSS and HSS are both derived from considerations of the marginal probabilities, and they both take into account the non-skill-related contributions (e.g., chance, bias, etc.) to the C table. The technical difference between the two is in the way they are normalized (Doswell et al. 1990): $\mathrm{TSS} = \mathrm{Tr}(C - E)/\mathrm{Tr}(C^* - E^*)$, while $HSS = \mathrm{Tr}(C - E)/\mathrm{Tr}(C^* - E)$, where E is the (biased) expected matrix based on C:

$$E = \frac{1}{N}\begin{bmatrix}(a + b)(a + c) & (a + b)(b + d)\\ (c + d)(a + c) & (c + d)(b + d)\end{bmatrix}.$$

This matrix is the C table that one would obtain in the absence of any skill, that is, with random guessing; the proof can be found in many statistics texts. Here, E^* is the (unbiased) expected matrix based on a hypothetical diagonal C-table, C^*, representing perfect accuracy:

$$C^* = \begin{pmatrix}a + b & 0\\ 0 & c + d\end{pmatrix},$$

$$E^* = \frac{1}{N}\begin{pmatrix}(a + b)^2 & (a + b)(c + d)\\ (a + b)(c + d) & (c + d)^2\end{pmatrix}.$$

The three measures GSS, CSS, and DSS complete the list of measures compiled by Murphy (1996). Note that many of these measures are in fact related; for example, $\mathrm{DSS} = \mathrm{TSS} \times \mathrm{CSS}$.

Murphy et al. (1989) define a measure of discrimination, DIS, derived from the conditional probability $p(f|x)$, that is, the posterior probability of a forecast f given an observation x. Specializing their formula for DIS to $f = 0, 1$ results in the expressions for DIS, given above.

The quantities θ and ϕ are measures that to our knowledge have not been considered elsewhere. Their origins are as follows: If the matrix $\mathbf{C}$ is symmetric, then it can be diagonalized by a rotation (similarity transformation) of the basis axes:

$$\mathbf{\Lambda} = \mathbf{T}_\theta \mathbf{C} \mathbf{T}_\theta^{-1},$$

where Λ would be the diagonal matrix of the eigenvalues and $\mathbf{T}$ is an orthogonal matrix written in terms of a single rotation parameter θ,

$$\mathbf{T} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Clearly, a diagonal C table would represent perfect performance, and as a result the angle of rotation could serve as a measure of performance. However, for a nonsymmetric matrix (as is the case with C tables if bias $\neq 1$) it is not possible to diagonalize with a single rotation, but one can show that a transformation of the type

$$\mathbf{\Lambda} = \mathbf{T}_\theta \mathbf{C} \mathbf{T}_\phi^{-1}$$

can render $\mathbf{\Lambda}$ diagonal, where $\mathbf{T}_\theta$ and $\mathbf{T}_\phi$ are rotation matrices but with different angles θ and ϕ.[2] In the nonsymmetric case, therefore, it requires a pair of quantities to provide a measure of performance, namely θ and ϕ. This is again a consequence of the multidimensionality of forecast quality (or the C table). It is interesting that in an $(M \times M)$ C table the number of rotation angles necessary for diagonalization [i.e., $2 \times M(M-1)/2$, the factor of 2 reflecting the nonsymmetric nature of the C table] is exactly equal to the number of independent degrees of freedom after the M ''climatological constraints'' (e.g., $N_0 = a + b$, $N_1 = c + d$, for $M = 2$) have been taken into account, that is, $M^2 - M$. However, it must be noted that these rotations cannot produce a diagonal matrix with the proper climatological frequency.

Finally, as for bias (Wilks 1995), if bias $= 1$, then the forecasts are unbiased. If bias < 1, then events are being underforecasted, otherwise overforecasting is occurring.[3] Note that bias $= 1$ implies that the C table is symmetric, that is, $b = c$. Also note that if $b = c$, then $\theta = \phi$. In other words, the difference between the two measures θ and ϕ is also a measure of bias.

3. Limiting cases

It is evident from their defining equations that PRD, AVG, and CSI are independent of a. This a independence does not imply that these measures fail to incorporate the correct classification of nonevents. The simplest way to see this is to note that one may always substitute $b = N_0 - a$ in the defining equations for the measures. Since N_0 is a fixed number, then these measures do effectively depend on the element a. In this respect, they are perfectly well-behaved measures in the rare-event situation.[4]

It is important to properly define what is meant by a ''rare-event situation.'' In a rare-event situation, the C table may look like

$$\begin{pmatrix} 9990 & 10 \\ 40 & 60 \end{pmatrix}.$$

First, note that $a \gg b$ and $c \sim d$, that is, a is much larger than b, while c is of the same order as d. For this reason, Doswell et al. (1990) consider the rare-event situation to be characterized by the inequality $a \gg b$. Also note that $N_0 = a + b = 10\,000$ and $N_1 = c + d = 100$, and thus $N_0 \gg N_1$. This inequality is simply a reflection of nature and its preferred proportion of nonevents to events. It is easy to show that

$$(a \gg b) \quad \text{and} \quad (c \sim d) \rightarrow (N_0 \gg N_1).$$

The converse is not true and so the inequality $N_0 \gg N_1$ is ''weaker'' than $(a \gg b,\ c \sim d)$. Although both inequalities are useful definitions of a rare-event situation, only $N_0 \gg N_1$ is an attribute of the ''situation''; the other is a characteristic of the classifier itself. For example, even when $N_0 = N_1$, overforecasting alone can yield a C table with $a \gg b$. To preserve the generality of the analysis, only $N_0 \gg N_1$ will be considered in this article. The question then arises as to the effect this inequality may have on the various measures.

The examination of the measures of performance in the rare-event situation is fruitful in general because even though the extreme inequality may not be realized in a given situation, the existence of any inadequacy in such extremely rare event limits may hint at the existence of an inadequacy (albeit a weaker one) even for situations where events are not extremely rare. In other words, in order for the pathologies to be of serious consequence and concern it is not necessary to have $N_0 \gg N_1$; even $N_0 > N_1$ (i.e., a common condition) may be sufficient to raise concern.

One aim of this study is to examine whether or not different measures of performance induce under- or overforecasting in rare-event situations. For that reason, the role played by bias is somewhat different from that of the other measures. To see how bias enters the analysis, it is sufficient to consider the way in which one arrives at a C table. Typically, the forecaster makes a decision based on some quantity, for example, dewpoint, gate-to-gate velocity difference, probability, or a regression function representing many variables, by introducing a decision threshold. If the measure of choice is inequitable (Gandin and Murphy 1992), then the fore-

[2] In performing a pair of transformations of this type the orthonormality of the axes is lost, weakening the geometrical significance of the angles of rotation. However, this is not a problem since the C table is only a table and not a true matrix, that is, it does not transform as a rank (1, 1) tensor on $V \otimes V^*$, where V is a vector space and V^* its dual.

[3] The author is indebted to R. L. Vislocky for introducing this notion of bias.

[4] In fact, since the C table has only two degrees of freedom, it is sufficient for a measure to depend on only two elements of the C table, as long as one of them is either a or b, and the other is either c or d.

TABLE 1. The values of the measures at four limiting cases: perfect prediction of both events and nonevents (I), constant forecasts of events (II) and nonevents (III), complete misclassification of both events and nonevents (IV), and classification by random guessing (V). The C-tables in these four cases are, respectively, $\begin{pmatrix} N_0 & 0 \\ 0 & N_1 \end{pmatrix}$, $\begin{pmatrix} 0 & N_0 \\ 0 & N_1 \end{pmatrix}$, $\begin{pmatrix} N_0 & 0 \\ N_1 & 0 \end{pmatrix}$, and $\begin{pmatrix} 0 & N_0 \\ N_1 & 0 \end{pmatrix}$. Here, N_0 and N_1 are the number of nonevents and events, respectively. Also, $N = N_0 + N_1$, $N_{01} = N_0/N_1$, and $N_{10} = N_1/N_0$.

	I	II	III	IV	V
PRD	1	$\frac{N_1}{N}$	0	0	$\frac{1 - c_{10} + N_{01}c_{01}}{1 + N_{01}^2}$
AVG	1	$\frac{1}{2}\left(1 + \frac{N_1}{N}\right)$	$0, \frac{1}{2}$	0	$\frac{2 - c_{10} + N_{01}c_{01}}{2(1 + N_{01})}$
FRC	1	$\frac{N_1}{N}$	$\frac{N_0}{N}$	0	$\frac{N_{01}(1 + N_{01}) + (1 - N_{01})(1 - c_{10} + N_{01}c_{01})}{(1 + N_{01})^2}$
EFF	1	0	0	0	$\frac{N_{01}(1 - c_{01} + c_{10})(1 - c_{10} + N_{01}c_{01})}{(1 + N_{01})^2}$
CSI	1	$\frac{N_1}{N}$	0	0	$\frac{1 - c_{10} + N_{01}}{1 + N_{01}(2 - c_{10} + N_{01}c_{01})}$
TSS	1	0	0	-1	0
HSS	1	0	0	$\frac{-2N_{01}}{1 + N_{01}^2}$	0
GSS	1	0	0	$\frac{-N_{01}}{1 + N_{01} + N_{01}^2}$	0
CSS	1	$-\frac{N_0}{N}, \frac{N_1}{N}$	$\frac{N_0}{N}, -\frac{N_1}{N}$	-1	0
DSS	1	0	0	1	0
DIS	∞	1	1	∞	1
θ	0	0	0	$\pi/2$	$\frac{1}{2}\tan^{-1}\left[\frac{2(1 - c_{10} + N_{01}c_{01})(c_{10} + N_{01}(1 - c_{01}))}{(1 + N_{01})(1 - N_{01} - 2c_{10} + 2N_{01}c_{01})}\right]$
ϕ	0	$\frac{1}{2}\tan^{-1}\left[\frac{2N_{01}}{1 - (N_{01})^2}\right]$	$\frac{1}{2}\tan^{-1}\left[\frac{2N_{01}}{1 - (N_{01})^2}\right]$	0	$\frac{1}{2}\tan^{-1}\left[\frac{2N_{01}}{1 - N_{01}^2}\right]$
Bias	0	$1 + \frac{N_0}{N_1}$	0	$\frac{N_0}{N_1}$	$1 - c_{10} + N_{01}c_{01}$

caster may be encouraged to lower or raise the decision threshold, in order to optimize that measure. However, there is no guarantee that the optimum of the measure corresponds to unbiased forecasts. In other words, in attempting to optimize a measure the forecaster may be unintentionally under- or overforecasting.

Table 1 lists the values of the measures in several limits.[5] The C table of case I represents perfect accuracy, while that of case IV reflects a complete lack of accuracy. At the same time, cases I and IV are equally and completely discriminatory. Cases II and III represent constant forecasts of all observations as events, or as nonevents, respectively. In other words, case II corresponds to very low decision thresholds, that is, overforecasting, and case III represents very high decision thresholds, that is, underforecasting. Another common standard of reference is the expected matrix **E** (previous section), and the values of the measures in this no-skill limit are given in column V.

Gandin and Murphy (1992) first note that CSI approaches N_1/N_0 in the limit II—a value larger than the corresponding limits in III and V—and then argue that CSI is inequitable in that a forecaster may increase his/her CSI by simply underforecasting. By the same token, they argue that any measure whose values in columns II, III, and V are unequal may encourage under- or overforecasting and is therefore inequitable.

However, this does not preclude the remaining measures from inducing biased forecasts as well. This can be seen by noting that even for a measure with vanishing

[5] To obtain the values of the measures in these limits, one must first introduce small parameters, ϵ, λ, in place of the zeros in the C table, for example, $\begin{pmatrix} N_0 - \epsilon & \epsilon \\ \lambda & N_1 - \lambda \end{pmatrix}$ in II. After the measures are calculated, then one may take the ϵ, $\lambda \to 0$ limit. However, the limits of AVG, and CSS, involve the ratio (λ/ϵ), leading to ambiguous results. Later in this article, these ambiguities will be shown to be related to the relative size of the standard deviations of the two classes.

limits in II, III, and IV, it is possible that the value of the threshold that optimizes such a measure corresponds to a C table whose bias is not equal to 1. As such, this measure is inequitable because in the process of optimizing it one will be biasing the forecasts.[6]

To examine the measures for any such inequitability, we expose the threshold dependence of the measures. That dependence is entirely contained to the quantities c_{01} and c_{10}, and so they can be written as $c_{01}(t)$ and $c_{10}(t)$, with t being the decision threshold. Then, the optima of the measures can be found by differentiating them with respect to t and setting the results equal to zero.

4. Some exact results

From column III of Table 1 it is evident that FRC approaches N_0/N, which in the rare-event situation is approximately 1. But this is the value of FRC in the perfect skill limit (column I). Therefore, by simply underforecasting one may increase FRC all the way up to its maximum value. Similarly, CSS may approach N_0/N (columns II and III) and can, therefore, suffer the same fate as FRC; the precise condition under which CSS approaches N_0/N will be given in the next section. To a lesser degree AVG has the same problem, since by simply underforecasting it approaches 0.5 (column III) suggesting nontrivial skill when in fact there is no skill at all. Both θ and ϕ have values in columns II and III that are either zero or approach 0 in the rare-event situation, but zero is also their perfect-accuracy value (column I), and so they cannot distinguish between under-, over-, or perfect forecasts. As such, AVG, FRC, CSS, θ, and ϕ are problematic measures.

As mentioned previously, the value of the decision threshold at which a given measure is optimized is an important quantity, because if the bias at that critical threshold is not equal to one, then the use of such a measure can induce under- or overforecasting. For the sake of brevity the details of the calculation will not be presented here, but it is easy (though lengthy) to show that the derivatives of the measures CSI, HSS, and GSS are equal in the rare-event situation. Therefore, they can be optimized simultaneously at a unique threshold. However, it is not easy to compute the value of bias at this threshold. To that end, an approximation must be made.

5. Gaussian approximation

One may assume that the underlying distributions of the events and nonevents are gaussian (normal) with means μ_0, μ_1 and standard deviations σ_0, σ_1, respectively (Fig. 1). Although this assumption may not be generally valid, it is often a fair approximation and it can aid in capturing some general properties of the measures. It is then straightforward to show (Marzban 1998)

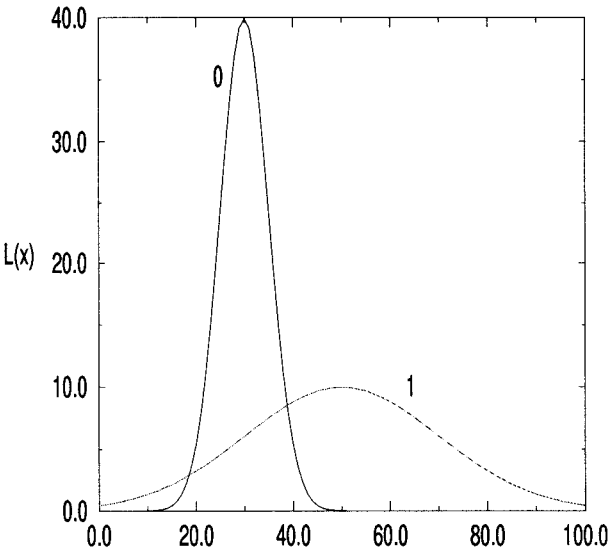

FIG. 1. Two Gaussian (normal) distributions with unequal variances.

$$c_{01} = \frac{1}{2}[1 - \text{erf}(t_0)], \quad \text{and} \quad c_{10} = \frac{1}{2}[1 + \text{erf}(t_1)],$$

where erf(x) is the Gaussian error function and t_i (i = 0, 1) are defined as

$$t_i \equiv \frac{t - \mu_i}{\sqrt{2}\,\sigma_i},$$

where t is the decision threshold.

TSS and FRC are special in that their critical thresholds can be computed, exactly, by noting

$$\frac{d}{dt}\text{erf}(t) = \frac{2}{\sqrt{\pi}}\exp^{-t^2},$$

which follows from the definition of the Gaussian error function

$$\text{erf}(t) = \frac{2}{\sqrt{\pi}}\int_0^t \exp^{-x^2}\,dx.$$

Then, one can prove that the critical threshold, t_c, maximizing FRC satisfies the quadratic equation

$$\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)t_c^2 + 2\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)t_c - \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2}\right) + 2\log\left(\frac{\sigma_0}{\sigma_1}\right) - 2\log\left(\frac{N_0}{N_1}\right) = 0.$$

It can also be shown that the relevant equation for TSS is given by the same quadratic but without the last term

[6] The author is indebted to one of the reviewers of this article for pointing out this extremely important and subtle point.

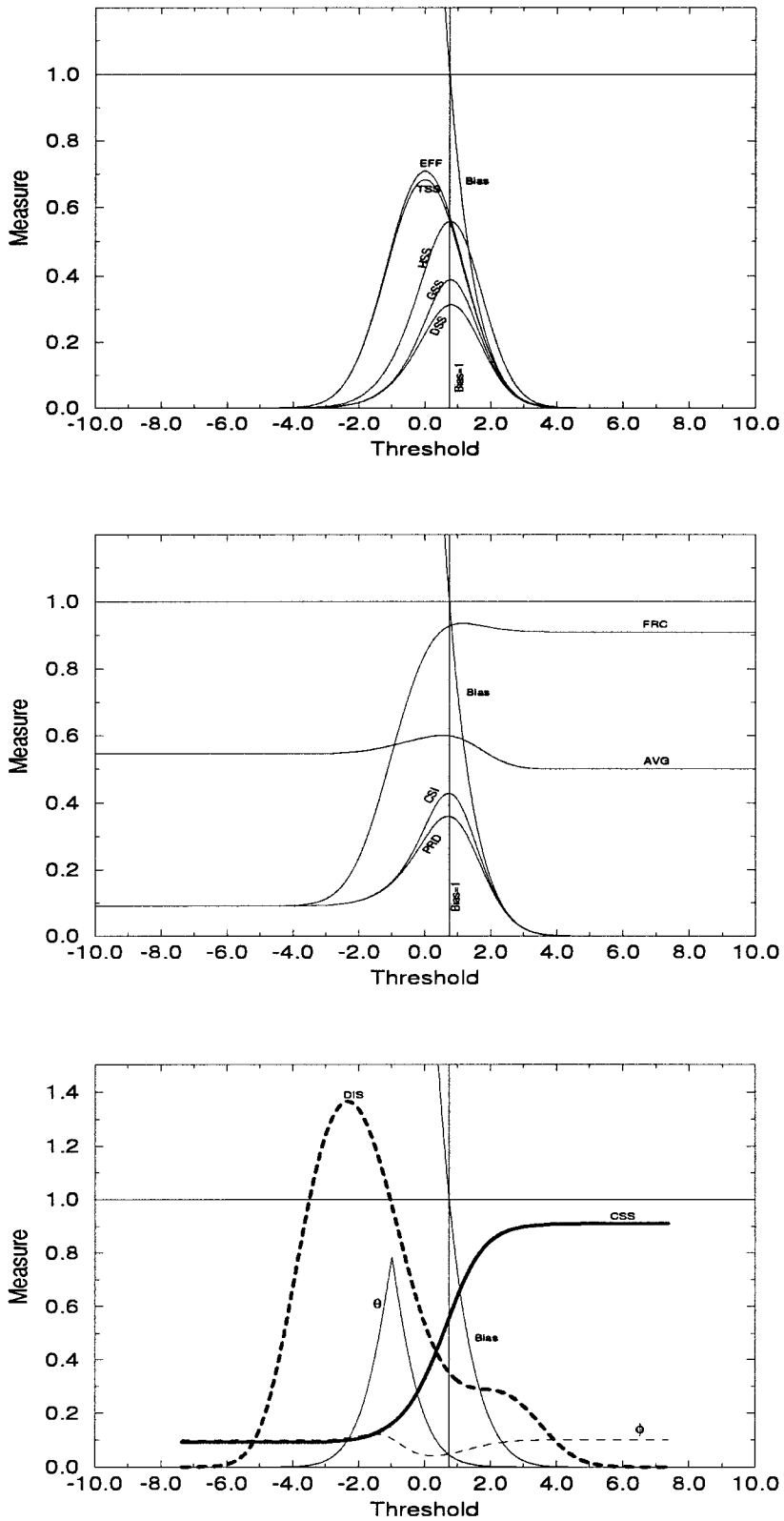

FIG. 2. The measures as a function of the decision threshold in a Gaussian (normal), equivariant ($\sigma_0 = \sigma_1$) approximation. The measure DIS is plotted on a log scale to allow for complete presentation. The vertical line represents the threshold at which bias is equal to 1.

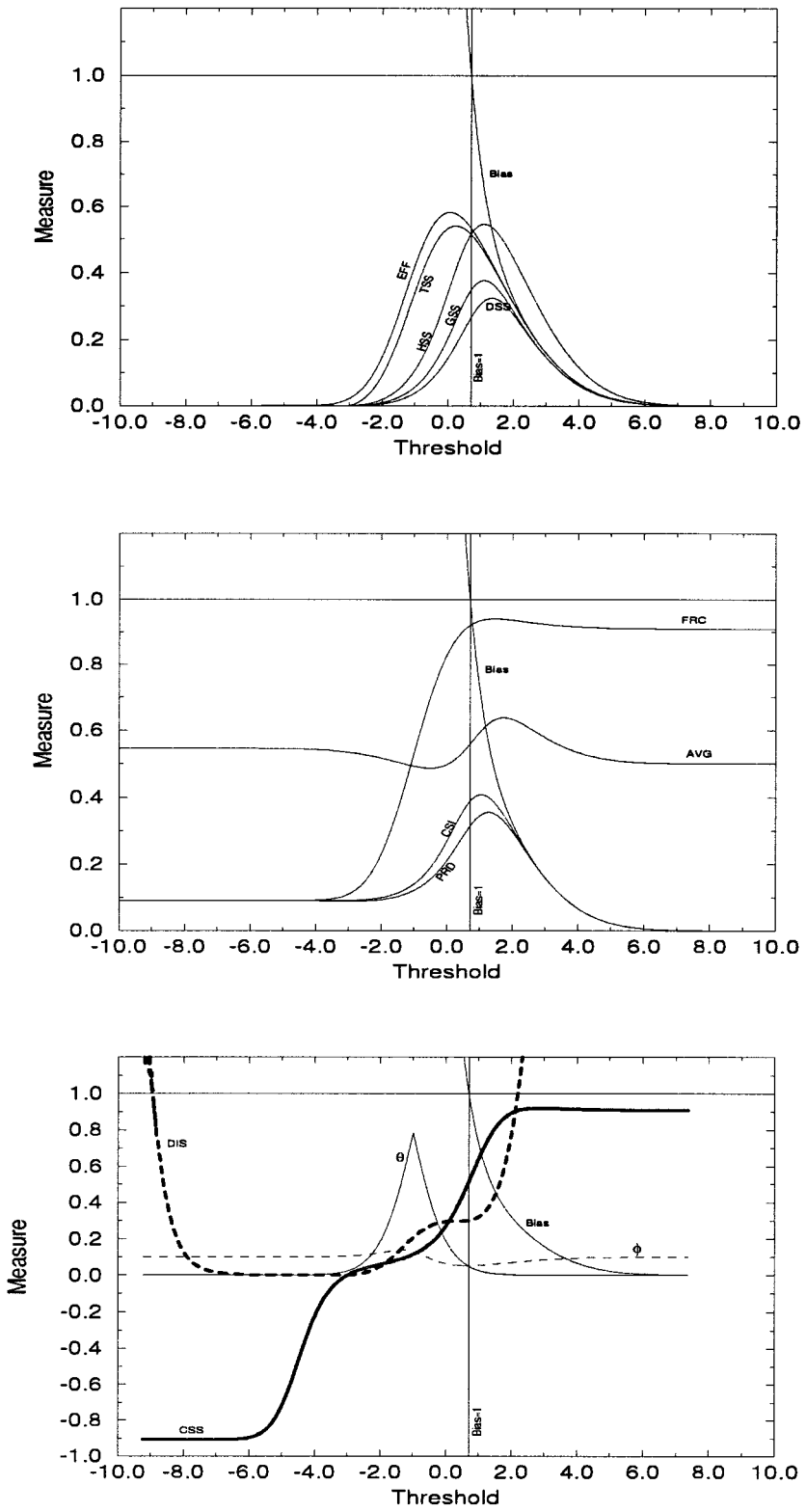

FIG. 3. The measures as a function of the decision threshold in a Gaussian (normal) approximation with $\sigma_0 < \sigma_1$. DIS is plotted on a log scale.

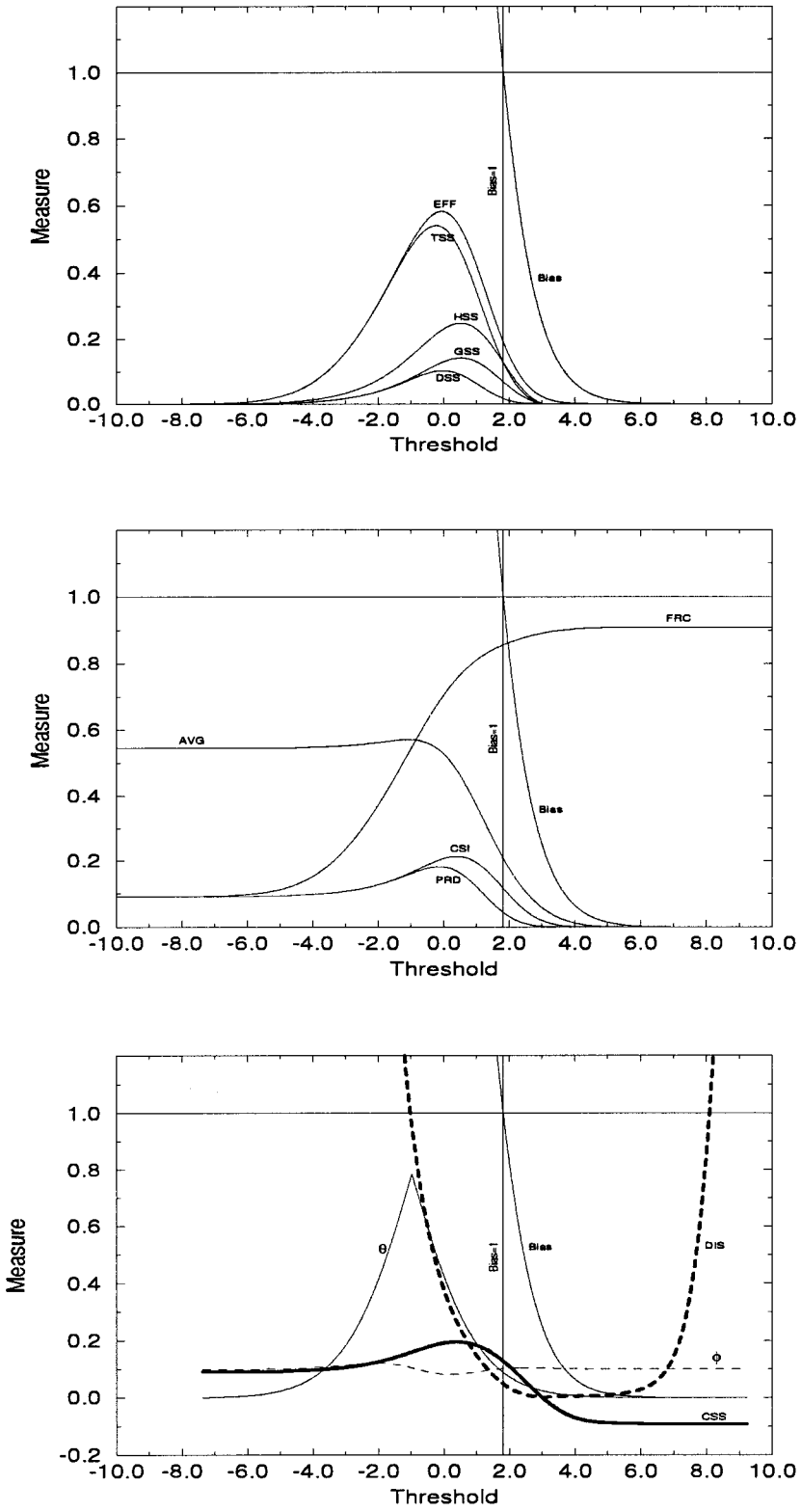

FIG. 4. The measures as a function of the decision threshold in a Gaussian (normal) approximation with $\sigma_0 > \sigma_1$. DIS is plotted on a log scale.

(involving N_0, N_1). Note that for the general case of unequal variances, there are in fact two thresholds at which FRC and TSS are maximized, although one of them occurs at very large values of the threshold. This is a consequence of having two crossing points between the two distributions (Fig. 1). The special case of equivariant distributions, $\sigma_0 = \sigma_1 = \sigma$, yields the intuitive results

$$t_c = \left(\frac{\mu_1 + \mu_0}{2}\right) + \frac{\sigma^2}{\mu_1 - \mu_0}\log\left(\frac{N_0}{N_1}\right), \quad \text{for FRC,}$$

$$t_c = \left(\frac{\mu_1 + \mu_0}{2}\right) \quad \text{for TSS.}$$

In a rare-event situation the second term in the t_c of FRC dominates the first term thereby tending to increase, or decrease, t_c without bound depending on the relative size of μ_1 and μ_0. Therefore, FRC induces underforecasting if $\mu_1 > \mu_0$, and overforecasting otherwise. Evaluating the bias at $t_c = (\mu_0 + \mu_1)/2$ yields a positive quantity (if $N_0 > N_1$), and therefore, TSS always induces overforecasting in a rare-event situation.

The remaining measures are difficult to address analytically, but they can be handled graphically. Figures 2, 3, and 4 display all of the measures when $\sigma_0 = \sigma_1$, $\sigma_0 < \sigma_1$, and $\sigma_0 > \sigma_1$, respectively. Without loss of generality the means have been set at $\mu_0 = -1$ and $\mu = 1$, and the sample size ratio has been set at $N_0/N_1 = 10$. For more extreme rare-event situations, for example, $N_0/N_1 \sim 50, 100, \ldots$, the behavior of the curves is mostly unchanged, and what change that does occur can be anticipated from the limiting values in Table 1. For example, FRC has a ''slight'' peak in Figs. 1 and 2; these peaks disappear as N_0/N_1 increases because according to Table 1 the value of FRC for large values of the threshold (i.e., extreme right-hand side of the graphs) approaches 1.

If the variances are equal (Fig. 2), then it can be seen that AVG, PRD, CSI, HSS, GSS, and DSS reach their maxima at the threshold for which bias = 1. Therefore, these measures are equitable in the equivariant case. By contrast, the optima of the remaining measures occur far from the bias = 1 line; EFF, TSS, DIS, and ϕ induce overforecasting (bias > 1), FRC and CSS induce underforecasting (bias < 1), while θ is capable of inducing either.

For $\sigma_0 \neq \sigma_1$, all measures are inequitable. If $\sigma_0 < \sigma_1$ (Fig. 3), then EFF, TSS, and ϕ, induce overforecasting, while PRD, AVG, FRC, CSI, HSS, GSS, DSS, and CSS all induce underforecasting. DIS and θ can induce either. If $\sigma_0 > \sigma_1$ (Fig. 4), all measures induce overforecasting, except for FRC, which induces underforecasting, and DIS and θ, which can induce either. Note that the results of the previous sections can be seen in these figures. For example, the values of the measures in columns II and III of Table 1 correspond to the values of the measures in the extreme left- and extreme right-hand side of the figures. Additionally, CSI, HSS, and GSS all have the same critical threshold, as anticipated. Also, one of the crossing points at which $\theta = \phi$ coincides with the bias = 1 line. This is a consequence of the comment made at the end of section 2.

It is worth emphasizing that the equality or the inequality of the variances are statistical statements. In other words, in a practical situation if the two variances are statistically equivalent (to some level of significance), then it behooves one to assume equivariance of the distributions. In that case, as shown above, PRD, AVG, CSI, HSS, GSS, and DSS are all equitable measures in a statistical sense.

6. Conclusions

A number of scalar measures of performance quality are examined in the rare-event situation. It is shown that AVG, FRC, CSS, θ, and ϕ are ill behaved in that their perfect-performance value coincides with their constant forecast value. Additionally, it is found that CSI, HSS, and GSS are optimized simultaneously at the same value of the decision threshold. It is further shown that in a Gaussian (normal) approximation if the variances of the distributions are statistically distinct, then all of the measures considered herein are inequitable in that they induce under- or overforecasting in rare-event situations. If the Gaussian distributions are statistically equivariant, then such bias is precluded for some of the measures; these measures are PRD, AVG, CSI, HSS, GSS, and DSS.

Acknowledgments. The author is grateful to H. Brooks, B. Davies-Jones, C. Doswell, J. Kuehler, and A. Murphy for many useful discussions, and thanks Mike Eilts and Arthur Witt for a careful reading of the original version of this manuscript. He is indebted to the editor, Michael Fritsch, for his almost unreasonable cooperation, and to all of the reviewers without whose input this work would simply not have existed. Special acknowledgement is in order to Robert L. Vislocky who first pointed out to the author that most, if not all, measures are generally inequitable; as a result of his numerous contributions his name could have justifiably appeared on this article. Partial support was provided by the FAA and the NWS/OSF.

REFERENCES

Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting,* **11,** 288–303.

Donaldson, R. J., R. M. Dyer, and M. J. Krauss, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conf. on Severe Local Stroms,* Norman, OK, Amer. Meteor. Soc., 321–326.

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

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

Hamill, T. M., and D. S. Wilks, 1995: A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting,* **10,** 620–631.

Marzban, C., 1998: Bayesian probability and scalar performance measures in Gaussian models. *J. Appl. Meteor.,* **37,** 72–82.

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

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

——, and E. S. Epstein, 1967: A note on probabilistic forecasts and ''hedging.'' *J. Appl. Meteor.,* **6,** 1002–1004.

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

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

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

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.