On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables

CHARLES A. DOSWELL III, ROBERT DAVIES-JONES, AND DAVID L. KELLER[1]

NOAA/Environmental Research Laboratories, National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 22 January 1990, in final form 20 July 1990)

ABSTRACT

The so-called True Skill Statistic (TSS) and the Heidke Skill Score (S), as used in the context of the contingency table approach to forecast verification, are compared. It is shown that the TSS approaches the Probability of Detection (POD) whenever the forecasting is dominated by correct forecasts of non-occurrence, i.e., forecasting rare events like severe local storms. This means that the TSS is vulnerable to "hedging" in rare event forecasting. The S-statistic is shown to be superior to the TSS in this situation, accounting for correct forecasts of null events in a controlled fashion. It turns out that the TSS and S values are related in a subtle way, becoming identical when the expected values (due to chance in a $k \times k$ contingency table) remain unchanged when comparing the actual forecast table to that of a hypothetical perfect set of forecasts. Examples of the behavior of the TSS and S values in different situations are provided which support the recommendation that S be used in preference to TSS for rare event forecasting. A geometrical interpretation is also given for certain aspects of the 2×2 contingency table and this is generalized to the $k \times l$ case. Using this geometrical interpretation, it is shown to be possible to apply dichotomous verification techniques in polychotomous situations, thus allowing a direct comparison between dichotomous and polychotomous forecasting.

1. Introduction

In a recent paper, Doswell and Flueck (1989, hereafter referred to as DF89) described the use of the contingency table for forecasting verification. Some summary measures of verification skill were mentioned and applied to a forecasting experiment, including the so-called True Skill Statistic (TSS).[2] The TSS (under any of its myriad names) is used widely in statistics and is recommended by Murphy and Daan (1985) as a "proper formulation of a skill score." Skill scores in general measure *relative* forecasting skill, comparing the forecasts in question to some standard forecasting technique. The idea is to avoid artificial inflation (or deflation) of one's perception of the quality of the forecasts. Some examples of standard forecasts include random guessing, climatology, and persistence. Everyone in operational forecasting understands, for instance, that the percentage of correct forecasts is really not a very meaningful statistic, unless it is substantially different from what one might obtain using, say, persistence. The TSS compares the number of correct forecasts, minus those attributable to random guessing (subject to the constraint that the marginal totals of observed events in the contingency table must remain the same), to that of a hypothetical set of perfect forecasts. We will show this in more detail in what follows.

In this paper, we wish to examine several aspects of verification using contingency tables, including the TSS, in situations where one might anticipate that one of the elements of the contingency table dominates the other elements. In particular, for forecasts of rare events (like tornadoes or flash floods), one expects that correct forecasts of non-occurrence will dominate a contingency table. This creates a variety of problems, to be discussed after we have established some basic definitions associated with the standard 2×2 contingency table (see DF89 and Donaldson et al. 1975) associated with dichotomous forecasts.

The contents of this paper are necessarily rather abstract and there is a chance that some field forecasters are not familiar with these esoteric aspects of verification scores. The failure of the TSS to deal effectively with rare event forecasts led to the analysis contained in this paper. We did not set out to consider the abstract properties of verification scores, but such considerations were forced on us; we needed an alternative to the TSS. While many operational forecasters may be unaware of (and uninterested in) the subtleties of verification scores, we hope that this paper will convince at least some of them that it is in their interests to

[1] Also affiliated with the Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma.

[2] The TSS is also known as the Hanssen–Kuipers (H–K) Discriminant (see, e.g., Woodcock 1976). Murphy and Daan (1985) refer to it as Kuipers' Performance Index. This proliferation of names for the same score is relatively common and can cause considerable confusion. Confusing terminology is confounded further by the lack of standardized notation.

Corresponding author address: Dr. Charles A. Doswell III, NOAA/ERL, 1313 Halley Circle, Norman, OK 73069.

TABLE 1. Schematic actual (left) and hypothetical "perfect" (right) forecast contingency tables, using the notation of Donaldson et al. (1975). In the "perfect" table, the number of observed events in each category remains the same as in the actual table, as does $N = x + y + z + w$, the total number of events.

	Observed				Observed		
Forecast	Yes	No	Total	Forecast	Yes	No	Total
Yes	x	z	$x+z$	Yes	$x+y$	0	$x+y$
No	y	w	$y+w$	No	0	$z+w$	$z+w$
Total	$x+y$	$z+w$	N	Total	$x+y$	$z+w$	N

become concerned about these issues. As noted in DF89, if forecasts are not verified, then one is not taking those forecasts seriously. However, mere calculation of numbers via some pre-existing verification scheme is not a very thoughtful way to evaluate the quality of those forecasts. It is all too easy to misuse or misapply some statistical summary measure of forecast quality. As suggested in DF89, if one is concerned with improving forecasts, one first must distinguish good from bad forecasting, and to be able to monitor the quality of forecasting over time.

We will show that the TSS has some disadvantages in situations involving forecasts of rare events. Further, the extension of the TSS to the $k \times k$ contingency table is rather complex and involves some complicated calculation. We will show that the Heidke skill score (S) discussed in Panofsky and Brier (1958; hereafter referred to as PB58) avoids these difficulties and seems better suited as a summary measure of skill in forecasting than the TSS when evaluating forecasts of rare events. The relationship between the TSS and the S-statistic will be shown in some detail.

A geometrical interpretation of some aspects of the verification of a 2×2 contingency table will be given, and the results generalized to the $k \times l$ case. This will allow a direct comparison between dichotomous and polychotomous verification results. We will use some of the data from DF89 to illustrate the evaluation and interpretation of S in comparison with the TSS, as well as to demonstrate the applicability of the geometrical interpretation of the verification statistics. Moreover, we also will include some illustrative examples from a forthcoming paper on verification of severe thunderstorm and tornado watches that will serve to show the performance of S in rare event forecasting.

2. Some basic definitions

a. Dichotomous forecasts

Table 1 defines the elements of the basic 2×2 contingency table associated with dichotomous forecasts. Donaldson et al. (1975) employed three of the four elements of this table to define various forecast verification measures, leading to what they called the Critical Success Index [or CSI $\equiv x/(x + y + z)$]. Note that the CSI has been employed elsewhere (e.g., Bermowitz and Zurndorfer 1979) under the name of the Threat Score, and has been used widely in operational forecast verification. A problem with this score is that no account is taken of the contents of the table associated with correct forecasts of null events. Clearly, in forecasting rare events, this term (w) will be the dominant one and the CSI ignores the potential problems associated with a very large w by not employing it at all (see Mason 1989). On the one hand, many of the correct forecasts of null events are trivial in character and seemingly of little interest in the verification. Many forecasters realize that at times, however, it takes a great deal of effort to conclude correctly that nothing will happen in a given situation and it seems unfair that such an effort can have no positive influence on the forecast evaluation. Schaefer (1990) has attempted to modify the CSI to deal with this problem.

Before we turn to some alternatives for incorporating the information contained in the w-element of the 2×2 table, let us consider the ways in which the data in the table can be combined. As shown in Table 2, there are eight ways in which ratios can be formed involving one of the elements with its associated marginal sums. Donaldson et al. (1975) defined only two of these, the Probability of Detection (POD) and the False Alarm Ratio (FAR).[3] Flueck (1987) has noted another, the Probability of False Detection (POFD). For the sake of completeness, we wish to give names to all eight of the possible combinations, and these are presented in Table 2 also. We note the following relationships which suggest we need only four[4] of the eight (e.g., POD, FAR, DFR, and POFD) to describe all the combinations.

$$1 - \mathrm{FAR} = \frac{x+z}{x+z} - \frac{z}{x+z} = \mathrm{FOH}, \qquad (1a)$$

$$1 - \mathrm{POFD} = \frac{z+w}{z+w} - \frac{z}{z+w} = \mathrm{PON}, \qquad (1b)$$

$$1 - \mathrm{POD} = \frac{x+y}{x+y} - \frac{x}{x+y} = \mathrm{FOM}, \qquad (1c)$$

[3] The POD is also referred to as *prefigurance*, while 1 − FAR is sometimes called the *post agreement* in some references (see, e.g., Brier and Allen 1951).

[4] Strictly, one needs only three of the combinations in the 2×2 case, because the fourth can be derived from the other three. However, for the general $k \times l$ situation, all four are necessary.

TABLE 2. Definitions for ratios of Table 1 elements with their associated marginal sums.

$\frac{x}{x+y}$ = POD	$\frac{x}{x+z}$ = FOH	$\frac{z}{x+z}$ = FAR	$\frac{z}{z+w}$ = POFD
$\frac{y}{x+y}$ = FOM	$\frac{y}{y+w}$ = DFR	$\frac{w}{z+w}$ = PON	$\frac{w}{y+w}$ = FOCN

POD: Probability of Detection
FAR: False Alarm Ratio
FOM: Frequency of Misses
PON: Probability of a Null event
FOH: Frequency of Hits
POFD: Probability of False Detection
DFR: Detection Failure Ratio
FOCN: Frequency of Correct Null forecasts

$$1 - \text{DFR} = \frac{y+w}{y+w} - \frac{y}{y+w} = \text{FOCN}. \quad (1d)$$

For this 2×2 case, the TSS can be shown to be (see Appendix A)

$$\text{TSS} = \text{POD} - \text{POFD} = \frac{(xw - yz)}{(x+y)(z+w)}, \quad (2)$$

while the Heidke skill score[5] is

$$\text{S} = \frac{C - E}{N - E}$$

$$= \frac{2(xw - yz)}{y^2 + z^2 + 2xw + (y+z)(x+w)}, \quad (3)$$

where C is the number of correct forecasts $(x + w)$, N is the total number of forecasts $(x + y + z + w)$ and E is the *expected* number of correct forecasts due purely to chance. The latter is derived here from the marginal sums to be

$$E = \frac{(x+z)(x+y) + (z+w)(y+w)}{x+y+z+w}. \quad (4)$$

Some limiting cases relevant to S are noted in Table 3, and it can be seen that S shares certain desirable properties with the TSS; e.g., S falls within a $(-1, +1)$ range so it looks rather like a measure of correlation. It can be seen in (2) and (3) that the numerators (in the 2×2 case) of the TSS and S are very similar, differing only by a factor of two. However, the denominators are quite different. We will show the relationship between these two scores in what follows. For the moment, we observe that whenever $y = z$, the TSS and S are equal.

We wish to consider what happens to the TSS in the case where the w-element becomes large, relative to the other elements in the contingency table. By taking the limiting case of (2), it is easy to show that

$$\lim_{z/w \to 0} \text{TSS} = \frac{x}{x+y} = \text{POD}. \quad (5)$$

To be absolutely rigorous, (5) is valid only if yz/xw tends to zero as well. If z/w tends to zero, then yz/xw will fail to tend toward zero only if x tends to zero as well, in which case the POD is likely to decrease. When w gets very large in comparison to the other elements in the table (particularly, the z-element), one can maximize the TSS simply by maximizing the POD.

On the other hand, when taking the same limits on (3), we find the limiting case for S to be

$$\lim_{z/w \to 0} \text{S} = \frac{2x}{2x+y+z} = 2\left(\frac{1}{\text{CSI}} + 1\right)^{-1}. \quad (6)$$

Therefore, in the limiting case, S tends to a simple function of the CSI (see Fig. 1).

As an ancillary issue, Donaldson et al. (1975) introduced a factor of κ in their analysis of the CSI to account for situations in which either the POD or the FAR is most important to the forecast. In some situations, one might be willing to accept a large FAR for some forecasts. Of course, a large FAR can lead to the "cry wolf" syndrome, which may be undesirable. The technique involves replacing z in the forecast contingency table (as in Table 1) with z/κ. Some characteristics of this modification to the CSI are noted in Table 3.

b. *Polychotomous forecasts*

In the situation where the forecasts and observations fall into more than two categories, the TSS must be generalized. It is convenient to define new notation for a $k \times k$ contingency table as shown in Table 4. Using this notation, the expected value for the ijth table element is given by $E_{ij} = (n_{i.})(n_{j})/n_{..}$ and this value can be subtracted from the observed ijth element to give that part of the observed elements which is due to skill

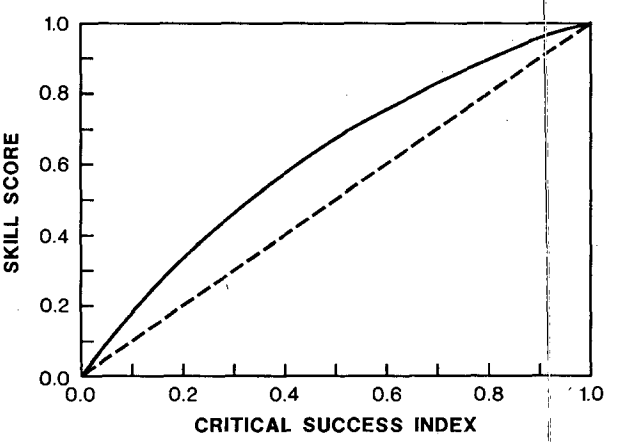

FIG. 1. Plot of the Heidke Skill Score (S) as a function of the Critical Success Index (CSI—solid line), in the limiting case [Eq. (6)] of the ratio z/w tending to zero. The dashed line is included for reference and show S ≡ CSI.

[5] Apparently, the skill score defined by Panofsky and Brier (1958) is due to Heidke (1926), as noted in Brier and Allen (1951). Hereafter, this will be referred to as the Heidke skill score.

TABLE 3. Some limiting cases.

A. Heidke Skill Score
1. $x = 0, y = 0$ (no observed events): $S \equiv 0$
2. $x = 0, z = 0$ (no events forecast): $S \equiv 0$
3. $x = 0, w = 0$ (no correct forecasts): $S = -2zy/(y^2 + z^2)$
   S attains its minimum when $y = z$ (at $S = -1$)
4. $y = 0, z = 0$ (no incorrect forecasts): $S \equiv 1$
5. $w \to \infty$ (rare events): $S = 2x/2x + y + z)$
   $= \{[(CSI)^{-1} + 1]/2\}^{-1}$

B. Miscellaneous
1. POD = 1: CSI = 1 − FAR
2. FAR = 1: CSI = 0
3. POD = 0: CSI = 0
4. FAR = 0: CSI = POD
5. $w \to \infty$: TSS → POD
6. The effect of the "κ-factor" in Donaldson et al. (1975):
   $z_\kappa = z/\kappa$
   a. $\kappa = 1$ (POD and FAR equally important): $FAR_\kappa = FAR$
   b. $\kappa \to 0$ (FAR most important): $FAR_\kappa \to 1$ ($CSI_\kappa \to 0$)
   c. $\kappa \to \infty$ (POD most important): $FAR_\kappa \to 0$ ($CSI_\kappa \to$ POD)

over and above random guessing. In effect, we treat the contingency table like a matrix $\mathbf{n}$ with elements n_{ij}. The matrix of expected values is denoted $\mathbf{E}$ and we subtract $\mathbf{E}$ from $\mathbf{n}$. This gives a new matrix ($\mathbf{R}$) with elements $R_{ij} = n_{ij} - E_{ij}$. It is relatively easy to show that $\mathbf{R}$ is symmetric. Now the sum of the diagonal elements of the matrix $\mathbf{R}$ [i.e., the *trace* of $\mathbf{R}$, denoted tr($\mathbf{R}$)] gives the number of correct forecasts beyond those attributable to random guessing, but in order to measure skill, it is desirable to compare tr($\mathbf{R}$) with some standard. For the standard, we develop a new matrix $\mathbf{R}^*$ which is based on the assumption of *perfect* forecasts. This allows us to define the generalized version of the TSS as

$$\text{TSS} = \frac{\text{tr}(\mathbf{R})}{\text{tr}(\mathbf{R}^*)}. \tag{7}$$

Clearly, if perfect forecasts were issued, the contingency table would look like that shown in Table 5. In such a case, the old $\mathbf{n}$ matrix is changed to a new matrix $\mathbf{n}^*$ which has all zeroes in the off-diagonal elements. The expected value matrix elements become E^*_{ij}

TABLE 4. Schematic $k \times k$ contingency table using the notation in Flueck (1987). The forecast and observed categories are the C_i ($i = 1, 2, 3, \cdots k$).

	Observed					
Forecast	C_1	C_2	C_3	$\cdots$	C_k	Total
C_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1k}	$n_{1.}$
C_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2k}	$n_{2.}$
C_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3k}	$n_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
C_k	n_{k1}	n_{k2}	n_{k3}	$\cdots$	n_{kk}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$\cdots$	$n_{.k}$	$n_{..}$

TABLE 5. Hypothetical $k \times k$ contingency table for "perfect" forecasts, derived from Table 4. Note that the observed totals remain unchanged.

	Observed					
Forecast	C_1	C_2	C_3	$\cdots$	C_k	Total
C_1	$n_{.1}$	0	0	$\cdots$	0	$n_{.1} = n_{1.}$
C_2	0	$n_{.2}$	0	$\cdots$	0	$n_{.2} = n_{2.}$
C_3	0	0	$n_{.3}$	$\cdots$	0	$n_{.3} = n_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
C_k	0	0	0	$\cdots$	$n_{.k}$	$n_{.k} = n_{k.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$\cdots$	$n_{.k}$	$n_{..}$

$= (n^*_{i.})(n^*_{.j})/n^*_{..}$. Thus, $\mathbf{R}^* = \mathbf{n}^* - \mathbf{E}^*$ and in the 2×2 case, the resulting TSS is identical to that shown in (2) above (see Appendix A).

While the TSS certainly can be extended to the $k \times k$ case, it is not simple and its calculation requires two passes (one for the actual, and one for the hypothetical perfect forecasts) to find the traces of the $\mathbf{R}$-matrices. In contrast, the Heidke skill statistic is rather easily modified for polychotomous forecasts. The number of correct forecasts (C) is the simple trace of the original matrix (or contingency table); $C = \text{tr}(\mathbf{n})$. The only significant change is in the calculation of the expected number of correct forecasts due solely to random chance. In the polychotomous case,

$$E = (n_{..})^{-1} \sum_{i=1}^{k} n_{i.} n_{.i}, \tag{8}$$

which, of course, is derived from the marginal totals in a relatively simple calculation.

So what is the relationship between the TSS and the Heidke S score? Had we defined the TSS with the $\mathbf{E}$-matrix remaining the same in finding the $\mathbf{R}^*$-matrix (i.e., $\mathbf{R}^* = \mathbf{n}^* - \mathbf{E}$), then we can show (see Appendix B) that TSS ≡ S! Thus, the only difference between these two seemingly quite different measures of forecasting skill is whether or not one changes the expected values for the hypothetically perfect forecasts. This is a rather subtle and not entirely obvious result. We do not know of any logic that asserts which of these alternatives is most sensible.

2. Regression and a geometric interpretation of the contingency table

When making dichotomous forecasts and verifying with dichotomous observations, the resulting 2×2 contingency table can be given a geometric interpretation. This begins by assigning numerical scores of unity to events and zero to non-events. Thus, if A and F denote variables representing actual observations and forecasts, respectively, then $(A, F) = (1, 1), (0, 0), (0, 1)$, and $(1, 0)$ represent, in order, a correctly forecast event, a correct forecast of a null event, a false alarm,

and a missed event. These ordered pairs represent the elements of the 2 × 2 contingency table. As shown in Fig. 2, therefore, all the elements of the table are represented by numbers that are plotted at the four corners of the figure. The next step is to find the *linear regression* of the forecasts upon the actual observations, denoted $\hat{F}(A)$. This is given by (Hays 1973)

$$\hat{F}(A) = b_{F\cdot A}(A - M_A) + M_F, \tag{9}$$

where the slope of the regression line is

$$b_{F\cdot A} = \frac{\mathrm{Cov}(A, F)}{s_A^2} = \frac{xw - yz}{N^2}\frac{N^2}{(x + y)(z + w)} = \mathrm{TSS},$$

in the 2 × 2 case, and where M_A and M_F are the actual observed and forecast means (respectively), while s_A^2 is the variance of the observed events and $\mathrm{Cov}(A, F)$ is the covariance between A and F. The means are given by $M_A = (x + y)/N$ and $M_F = (x + z)/N$; the covariance by $\mathrm{Cov}(A, F) = (xw - yz)/N^2$; and $s_A^2 = (x + y)(z + w)/N^2$. We also can find the regression line of A upon F, denoted $\hat{A}(F)$, which satisfies a relation similar to (9), namely

$$\hat{A}(F) = b_{A\cdot F}(F - M_F) + M_A, \tag{10}$$

where

$$b_{A\cdot F} = \frac{\mathrm{Cov}(A, F)}{s_F^2} = \frac{xw - yz}{(x + z)(y + w)},$$

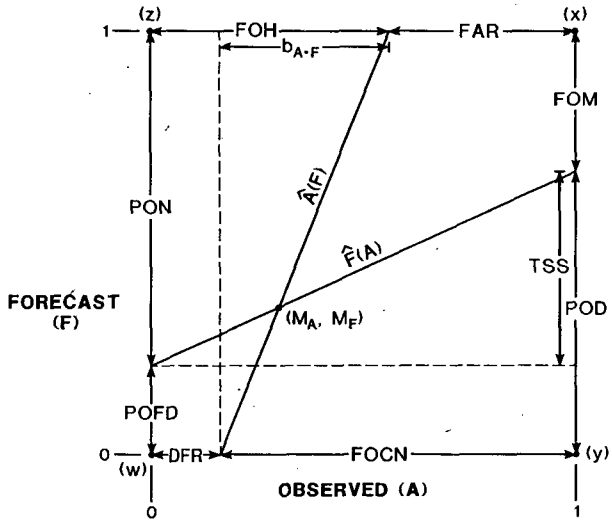

FIG. 2. Schematic example showing the regression lines of the forecast upon the observations $[\hat{F}(A)]$ and the observations upon the forecasts $[\hat{A}(F)]$. The intercepts of these with the $A = (0, 1)$ and $F = (0, 1)$ axes define the quantities listed in Table 2 (see text). The values M_A and M_F are the averages of the observations and the forecasts, respectively. The elements of the 2 × 2 table (x, y, z, w) are shown plotted at their associated corners in this figure. See the text for explanations of TSS and $b_{A\cdot F}$.

in the 2 × 2 case, and where s_F^2 is the variance of the forecasts ($s_F^2 = (x + z)(y + w)/N^2$). The two regression lines (9) and (10) are shown on Fig. 2, and it is relatively straightforward to show that the eight quantities defined in Table 2 satisfy the indicated geometric relationships, determined by the intercepts of the regression lines with the A and F lines of zero and unity. It is noteworthy that as w becomes very large, the intersection point (M_A, M_F) moves toward the origin at (0, 0), which means that the TSS tends toward the POD, as we already have shown.

Figure 2 provides an elegant picture of the contents of Table 2, as well as showing that the TSS can be interpreted as the slope of the regression line $\hat{F}(A)$, $b_{F\cdot A}$ (also noted by Woodcock 1976). Although we can show various characteristics of $b_{A\cdot F}$, such as $b_{A\cdot F}$ = FOH − DFR, we do not know of any summary statistic comparable to the TSS that is associated with $b_{A\cdot F}$. If one were using *probabilities* to forecast dichotomous events (as in National Weather Service precipitation probabilities) the data points would be scattered along the $A = 0$ and $A = 1$ axes. In the most general case with both polychotomous forecasts and observations, the data would be scattered all over the figure.

As a first test of this generalization, Fig. 3 shows the regression lines for the probability forecasts of "go/no go" days (a dichotomous observation) that were done at the same time as the dichotomous forecasts summarized in Table 5 of DF89.[6] By finding these regression lines, one can use the graphical interpretation of such quantities as FAR and POFD (as in Fig. 2) to evaluate the polychotomous forecasts in a way comparable to that done in dichotomous forecast verification. In Fig. 3, for example, the intercept of $\hat{F}$ with the $A = 1$ axis (a "POD-like" quantity) is 0.61, while the intercept of $\hat{A}$ with the $F = 1$ axis (an "FOH-like" quantity) is 0.93, giving a value to the "FAR" of 0.07. Also shown on Fig. 3 are the regression lines derived from the dichotomous forecast verification (see Table 5 in DF89). The results are similar but not identical to the dichotomous forecasts made at the same time, suggesting an inconsistency between the dichotomous and polychotomous forecasts. Given the inexperience with probability forecasting among the DOPLIGHT '87 forecasters, this inconsistency between "categorical" and "probabilistic" forecasts issued at the same time is not surprising.

Another test of this idea for generalizing the statistical evaluation of the dichotomous contingency table to polychotomous situations can be derived from our own Tables 6 and 7, to be discussed below in more detail.

[6] The probability forecasts were categorized in Tables 8 and 9 of DF89 and then plotted in the reliability diagram (Fig. 2) of that paper, although the uncategorized forecasts were used in DF89 to find the average probability forecast. For the present paper, the uncategorized probability values were used exclusively.

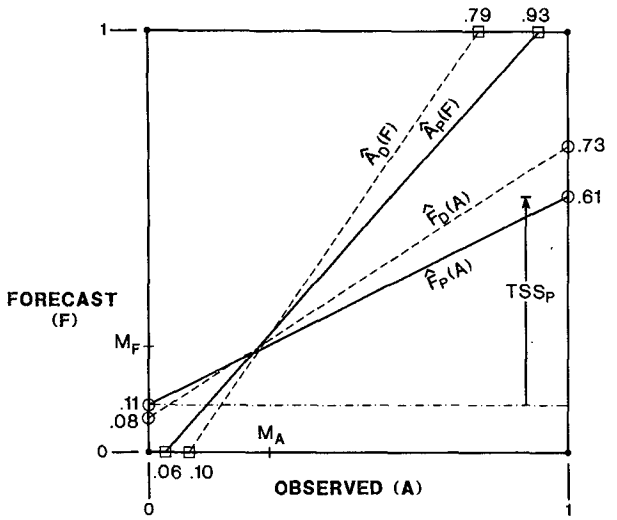

FIG. 3. Regression lines associated with the probability (solid lines) and the categorical (dashed lines) forecasts, and the "go/no go" days. Lines labelled $\hat{F}_P(A)$ and $\hat{A}_P(F)$ are the polychotomous forecasts upon the observations and the observations upon the forecasts, respectively. Similar quantities with "D" subscripts are for the dichotomous forecasts. The values of the intercepts are shown for the regression lines, associated with the POD, POFD, DFR, and FOH values (see Fig. 2 and Table 2), and the M_A and M_F values are for the polychotomous forecasts located on the axes. Heavy lines indicate the POD-, POFD-, DFR-, and FAR-like quantities associated with the set of polychotomous forecasts and dichotomous observations (see text for discussion). Note that all the data points fall on the $A = 0$ and $A = 1$ axes.

Table 6 is polychotomous in both the forecasts and the observations, although neither is quantitative. Rather, the forecast and observed quantities describe the qualitative character of the events (null, nontornadic severe, tornadic). If one assigns a numerical value of zero to a non-event (either forecast or observed), what value is given to the other events? In effect, Table 7 assigns each of the event types the same value, namely unity, thereby being reduced to a dichotomous situation. Table 8 shows the effect on the statistics of assigning two different values to severe thunderstorm events, relative to a tornado event, which is assumed to have a unit value. The effect of reducing the relative weight on severe thunderstorms vis-a-vis tornadoes is to increase the POD, FAR, and TSS while reducing the CSI, due to the large number of severe thunderstorm versus tornado events, and to the relative unimportance of the FAR in the TSS when the table is dominated by correct forecasts of non-events. For nonquantitative forecasts and observations like this, the relative weights assigned to the different events constitute a sort of extra "degree of freedom" in the statistical analysis of polychotomous, non-quantitative event contingency tables. For quantitative, polychotomous forecasts (e.g., probability) and/or observations, one is not free to make arbitrary weighting assignments.

In graphical terms, then, the slope and intercept of the regression lines are driven by the contents of the contingency table in a basically simple way. For perfect forecasts (purely diagonal tables) the regression lines lie along the 45°, $F = A$ line. As forecasts depart from perfection, the regression lines turn away from this line. The point (M_A, M_F) represents the center of mass of the points on the diagram. The geometry of the regression lines is clearly the determining factor for quantities like the POD and the DFR and, hence, is directly related the verification statistics. It appears to be useful to generalize the techniques of Donaldson et al. (1975) to allow one to compute statistics comparable to those of the 2 × 2 dichotomous situation in polychotomous cases.

TABLE 6. Contingency table for verification of severe thunderstorm ("blue") and tornado ("red") watches for 1984. For this verification the entire country has been covered with grid boxes (roughly 40 km on a side) and each hour is considered separately. The numbers represent units of "grid-box hours" and for purposes of reference, the average watch in 1984 was valid over a space-time period of about 320 grid-box h units. Multiple events within a grid-box h do not affect the totals and watch cancellations have been accounted for, as have overlapping watches.

	Observed			
Forecast	Red	Blue	None	Total
Red	360	1235	64 043	65 638
Blue	38	464	40 181	40 683
None	471	3328	39 707 774	39 711 573
Total	869	5027	39 811 998	39 817 894

TSS = 0.246, S = 0.026

TABLE 7. As in Table 6, except that tornadoes and severe thunderstorms have been combined into one category of "severe" events.

	Observed		
Forecast	Severe	None	Total
Severe	2097	104 224	106 321
None	3799	39 707 774	39 711 573
Total	5896	39 811 998	39 817 894

POD = 0.356, FAR = 0.980, CSI = 0.019, TSS = 0.353, S = 0.037

TABLE 8. Statistics generated from assigning numerical values to the forecasts of Table 6, as indicated, and using the generalized definitions of POD, FAR, etc. for polychotomous situations. The values assigned are given as ordered arrays (T, ST, N), where T is the value assigned to tornado events, ST the value assigned to severe thunderstorm events, and N the value assigned to non-events.

Values	POD	FAR	CSI	TSS	
(1, 1, 0)	0.356	0.980	0.019	0.353	(Table 7)
(1, .75, 0)	0.426	0.982	0.017	0.423	
(1, .5, 0)	0.522	0.985	0.014	0.520	

4. Evaluation of DOPLIGHT '87 results using S

As discussed in DF89, a forecasting experiment (called DOPLIGHT '87) was conducted during the spring of 1987 by the National Severe Storms Laboratory in collaboration with the Norman, Oklahoma National Weather Service Forecast Office. Interested readers can consult DF89 for details. Here, our primary purpose is to present S-values for selected results in DF89 in order to see if S offers any additional insight.

For the dichotomous forecasts shown in DF89's Tables 3, 4, 5, and 7, we have the results shown in Table 9. Although these tables are dominated by the w-element, the dominance is not so strong that there is a great deal of difference between the CSI, TSS, and S scores. In effect, for these data, the distinctions among the three different summary statistics are more or less negligible. As also shown in Table 9, the data for mesocyclone forecasting (DF89's Table 9) suggest that the CSI, TSS, and S statistics are rather different. It was pointed out in DF89 that there were too few mesocyclone events in the data set to put much faith in the statistics; here, we note that this particular table is dominated by the w-element to a greater extent than the others.

Finally, Table 9 shows the computations for the data in DF89's Table 10. These were the so-called convective mode forecasts which served to illustrate the calculations for polychotomous forecasts. Note that the TSS and S statistics are nearly identical. Again, there is one dominant element in the table, but its dominance is not exceedingly great.

Although there appear to be some minor differences between the TSS and S statistics for the DOPLIGHT '87 data, these do not seem to be very important. This result confirms our abstract analysis that suggests little difference between the two when the contingency table is not characterized by overwhelming dominance of the w-element in the table.

TABLE 9. Additional calculations from DOPLIGHT '87 data presented in DF89. See the indicated tables in DF89 for the raw contingency tables and the other summary statistics.

Forecast product	Expected # correct	Observed # correct	Heidke skill score	TSS	DF89 table #
Advance outlook	55.6	79	.66	.63	3
Morning update	54.7	75	.56	.55	4
Noon outlook	54.7	79	.67	.65	5
SELS day one	53.0	77	.63	.65	7
Mesocyclones	76.2	83	.49	.61	9
Convective mode	24.4	47	.35	.36	10

5. Comparison of TSS and S in rare event forecasting

In order to give the Heidke skill score a real test against the TSS, we need a data set which truly is overwhelmed by correct forecasts of null events. We currently are working on such a data set, which will be the subject of a forthcoming paper (a preliminary version of which is given in Doswell et al. 1990). Here, we will give a sample of the sort of data we will be evaluating to show the value of the S statistic relative to the TSS and CSI.

The forecasts being verified are severe thunderstorm and tornado watches issued by the Severe Local Storms forecasting unit of the National Severe Storms Forecast Center (NSSFC). We have used a grid to break the contiguous 48 states into small grid boxes and have used the NSSFC verification data sets to determine the number of hours each grid box is under a tornado or severe thunderstorm watch. The validating data of severe thunderstorm and tornado reports (also supplied by NSSFC) are then used to construct contingency tables for each grid box. Results for the year of 1984, for all the boxes in the 48 contiguous states are shown in Tables 6 and 7, where Table 6 is for the 3×3 version of the contingency table while Table 7 is the condensed, 2×2 form.

These tables quite clearly are dominated by the correct forecasts of non-events. It may be seen that the TSS and S scores differ substantially (by about an order of magnitude) in both tables. It is also obvious that when watches are verified in this fashion, one finds that a significant fraction of the watch area is not affected by any severe weather, giving a distinct impression of overforecasting. This article is not the place for a detailed treatment of this subject, but we should observe that one expects not to see every part of a severe thunderstorm or tornado watch contain an event for the entire duration of the watch. This expectation is due, at least in part, to the difference between point and area forecasts, the watch being one of the latter. Apart from this question of how to interpret these numbers, we should point out that these figures are preliminary in character, and are subject to slight changes by the time the project is finished. They do serve to illustrate our point about the TSS and S statistics, however.

Table 7 shows that the S-score indeed is higher than the CSI, as it should be if correct forecasts of null events are accounted for in the scoring. If we note that there are about 30 times as many false alarms (the z-element in Table 8) as there are detection failures (the y-element) then we can re-calculate the CSI using a "κ-factor" of 30 (which means that the resulting table has its z-value roughly equal to its y-value). In such a case, we find that $CSI_\kappa = 0.224$. Similarly, one can re-calculate the Heidke skill score to find $S_\kappa = 0.366$, again revealing credit given for correctly forecasting non-events. The TSS also can be re-computed as $TSS_\kappa = 0.356$, which represents only a very small change and the TSS_κ has become virtually identical to the POD. Also, the TSS_κ and S_κ are nearly equal, which is the

result of the use of the κ-factor to make z very nearly the same as y.

6. Summary and discussion

It is evident from what we have shown that the TSS has a problem in dealing with cases involving forecasts of rare events. We have shown that the TSS becomes close to the POD when correct forecasts of nothing happening dominate the forecast contingency table. Therefore, the TSS is neither "strictly proper" nor "proper," because a forecaster can maximize the score by overforecasting in those situations where there is even a remote chance of the event occurring (provided that the condition $w \gg z$ remains valid). This approach only works if there is a great preponderance of situations where there is virtually no chance of the event (i.e., rare events). Thus, use of this skill score encourages rare event forecast "hedging" of the sort described by Murphy and Epstein (1967)—i.e., deviating from the forecaster's true beliefs in order to increase the verification score. We conclude that the TSS is an improper scoring rule for rare event forecasting.

We have offered the Heidke skill score (S) as a statistic which avoids this problem and still retains most of the desirable features of the TSS noted in Flueck (1987) and DF89. It incorporates information from the w-element of the 2×2 contingency table in a controlled way, such that credit is given for correct forecasts of non-events, but the effect of false alarms is considered, even in the limiting case where the ratio of false alarms to correct null forecasts goes to zero. Moreover, S is easier to calculate than the TSS, especially in tables of higher order than the 2×2 case. Since S $\geqslant$ CSI throughout its range, this implies that the S-score is, in effect, giving credit for correct null forecasts (i.e., the contents of the w-element in the contingency table) but in a reasonably controlled way, unlike the TSS. Its advantage over the CSI itself is that the w-element is being factored into the summary measure of skill.

We have shown how the TSS and S are related. The key element in this derivation is the equality of the marginal sums $n_{.i}$ and $n_{i.}$ (for $i = 1, 2$) in the perfect forecast matrix $\mathbf{n}^*$, used in showing (A3) of Appendix A. This has the effect of changing the expected value matrix in the case of perfect forecasts to account for the change in those marginal sums resulting from different forecasts than in the actual matrix. The bogus situation in which the expected value matrix remains the same gives the unexpected result that TSS = S. We observe that the perfect forecast table is an *implicit* part of the Heidke skill score, because N can be interpreted as the number of correct forecasts when the forecasts are perfect. S is also equal to the ratio of the traces of **R**-matrices (i.e., a form of the TSS) when the E-matrix is held fixed.

We also have shown that several aspects of the verification associated with a contingency table have a geometric interpretation. This has allowed us to compare dichotomous and polychotomous forecast verification scores directly, simply by looking at the intercepts of the regression lines with the forecast and actual observed coordinate axes. Thus, we have generalized the notions of Donaldson et al. (1975) to both partially and fully polychotomous situations, including those where the polychotomous categories are not quantitative. Because we have shown that the dichotomous techniques can be applied to polychotomous situations, this underscores further the point made in DF89 that the distinction between dichotomous (i.e., "categorical") and polychotomous (e.g., "probabilistic") situations is illusory.

We have illustrated differences among the three summary measures, TSS, CSI, and S, in forecast situations where the contingency table is not overwhelmingly dominated by correct null forecasts, and in a case where the table is so dominated. It is clear that the Heidke skill score offers the distinct advantages of being usable in both situations and incorporating information about correct null forecasts in a controlled way.

Use of the TSS in rare event forecasting is comparable to letting κ become very large in the Donaldson et al. modification of the CSI, because in both cases the summary statistic tends toward the POD. In fact, we have shown (e.g., Table 8) the TSS, CSI, and S scores to be nearly equal when a κ-factor is employed to reduce the impact of false alarms by making the y-element and the z-element in the contingency table of similar size.

Much of this paper has been concerned with the relative merits of the Heidke S score vs. the TSS. We believe, at least in cases involving rare event forecasting, the Heidke score is superior to the TSS. However, we observe that both the TSS and S scores are unchanged if we interchange the diagonal elements, since the trace remains the same. If we consider interchanging the diagonal elements in Table 8, for example, we would have a quite different picture of forecasting success but the S score would not be changed. The CSI, in contrast, would be significantly different. Clearly, no single measure of forecasting success can give a complete picture and it is desirable to include, in addition to S, the CSI, POD, and FAR (at least) in any summary of forecasting verification (as in, for example, Goldsmith 1989). While it borders on being trite to draw this conclusion, its very triteness suggests that many verification efforts put too much emphasis on a single score to describe all of the information contained in a contingency table.

Acknowledgments. We would like to thank Mr. Ding Jincai (visiting NSSL from the Shanghai Meteorological Center) for several helpful discussions on this topic, as well as the staff members of the National Severe Storms Forecast Center for supplying us with

the data used in section 5 of this paper. The constructive criticisms of the anonymous reviewers were of considerable value in improving the presentation. Ms. Joan Kimpel's skillful drafting of the figures is also appreciated.

APPENDIX A

Derivation of the TSS from the Generalized Definition

We begin with the generalized definition of the TSS as the ratio of traces of **R**-matrices shown in (7). In the case of the numerator, we use the notation shown in Table 4 to find (in the 2×2 case) that

$$\begin{aligned}\mathrm{tr}(\mathbf{R}) &= \left[n_{11} - \frac{(n_{1.})(n_{.1})}{n_{..}}\right] + \left[n_{22} - \frac{(n_{2.})(n_{.2})}{n_{..}}\right] \\ &= (n_{..})^{-1}[(n_{..}n_{11} - n_{1.}n_{.1}) + (n_{..}n_{22} - n_{2.}n_{.2})],\end{aligned} \tag{A1}$$

while

$$\begin{aligned}\mathrm{tr}(\mathbf{R}^*) &= \left(n_{.1} - \frac{(n_{1.})(n_{.1})}{n_{..}}\right) + \left(n_{.2} - \frac{(n_{2.})(n_{.2})}{n_{..}}\right) \\ &= (n_{..})^{-1}[(n_{..}n_{.1} - n_{1.}n_{.1}) + (n_{..}n_{.2} - n_{2.}n_{.2})].\end{aligned} \tag{A2}$$

Using (A1) and (A2) in (7) and making use of the fact that in the case of perfect forecasts, $n_{1.} = n_{.1}$ and $n_{2.} = n_{.2}$, it can be seen that

$$\mathrm{TSS} = \frac{[n_{..}(n_{11} + n_{22}) - (n_{1.}n_{.1} + n_{2.}n_{.2})]/n_{..}}{[n_{..}(n_{.1} + n_{.1}) - (n_{.1}^2 + n_{.2}^2)]/n_{..}}. \tag{A3}$$

Expanding the numerator and denominator, noting that $n_{.1} + n_{.2} = n_{..}$ and that $n_{.1}^2 + n_{.2}^2 = (n_{.1} + n_{.2})^2 - 2n_{.1}n_{.2}$, (A3) becomes

$$\begin{aligned}\mathrm{TSS} &= \frac{[2(n_{11}n_{22} - n_{12}n_{21})]/n_{..}}{2(n_{.1}n_{.2})/n_{..}} \\ &= \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{.1}n_{.2}}.\end{aligned} \tag{A4}$$

From (A4) it is easy to confirm that

$$\mathrm{TSS} = \frac{n_{11}}{n_{.1}} - \frac{n_{22}}{n_{.2}} \equiv \mathrm{POD} - \mathrm{POFD}, \tag{A5}$$

using the notation from Table 4 and the definitions in Table 2. This confirms that (2) can be derived from the more general TSS definition (7).

APPENDIX B

Derivation of the Relation Between the TSS and Heidke S Scores

In order to show things a bit more clearly, we will revert to the original notation for the 2×2 contingency table (Table 1). As before, we begin with the most general definition of the TSS as the ratio of $\mathrm{tr}(\mathbf{R})$ to $\mathrm{tr}(\mathbf{R}^*)$. In this case, $E_{11} = (N)^{-1}[(x + z)(x + y)]$ and $E_{22} = (N)^{-1}[(z + w)(y + w)]$. Therefore,

$$\begin{aligned}R_{11} &= x - \frac{(x + z)(x + y)}{N} = \frac{1}{N}(xw - yz) \\ &= w - \frac{(z + w)(y + w)}{N} = R_{22}.\end{aligned} \tag{B1}$$

If we consider the perfect forecast contingency table, but do not allow the expected values to change [see (4)], then

$$\begin{aligned}R_{11}^* &= (x + y) - \frac{(x + z)(x + y)}{N} \\ &= \frac{1}{N}(x + y)(y + w),\end{aligned} \tag{B2}$$

$$\begin{aligned}R_{22}^* &= (z + w) - \frac{(z + w)(y + w)}{N} \\ &= \frac{1}{N}(x + z)(z + w).\end{aligned} \tag{B3}$$

From (B2) and (B3) it is straightforward to show that

$$\mathrm{tr}(\mathbf{R}^*) = (N)^{-1}[y^2 + z^2 + 2xw + (x + w)(y + z)], \tag{B4}$$

so that from (B1–4) and (7), we have

$$\begin{aligned}\mathrm{TSS} &= \frac{(2/N)[xw - yz]}{(1/N)[y^2 + z^2 + 2xw + (x + w)(y + z)]} \\ &= \mathrm{S},\end{aligned} \tag{B5}$$

the latter equality following from (3), above. This demonstrates that in the case where the expected value matrix does not change, the TSS and S scores are exactly the same (at least for a 2×2 table).

In order to treat the general $k \times k$ case, we observe that $C = \mathrm{tr}(\mathbf{n})$ and $E = \mathrm{tr}(\mathbf{E})$. Therefore

$$\mathrm{S} = \frac{\mathrm{tr}(\mathbf{n}) - \mathrm{tr}(\mathbf{E})}{N - \mathrm{tr}(\mathbf{E})}, \quad \mathrm{TSS} = \frac{\mathrm{tr}(\mathbf{n}) - \mathrm{tr}(\mathbf{E})}{\mathrm{tr}(\mathbf{n}^*) - \mathrm{tr}(\mathbf{E}^*)}$$

which means, since $\mathrm{tr}(\mathbf{n}^*) = N$, that $\mathrm{S} \equiv \mathrm{TSS}$ whenever $\mathbf{E}^*$ is replaced with $\mathbf{E}$, even in the general $k \times k$ case.

REFERENCES

Bermowitz, R. J., and E. A. Zurndorfer, 1979: Automated guidance for predicting quantitative precipitation. *Mon. Wea. Rev.*, **107**, 122–128.

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

Donaldson, R. J., R. M. Dyer and M. J. Krauss, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints: *9th Conf. Severe Local Storms*, Norman, Oklahoma, Amer. Meteor. Soc., 321–326.

Doswell, C. A. III, and J. A. Flueck, 1989: Forecasting and verifying in a field research project: DOPLIGHT '87. *Wea. Forecasting,* **4,** 97–109.

——, D. A. Keller and S. J. Weiss, 1990: An analysis of the temporal and spatial variation of tornado and severe thunderstorm verification. Preprints, *16th Conf. Severe Local Storms,* Kananaskis Park, Alberta, Amer. Meteor. Soc., 294–299.

Flueck, J. A., 1987: A study of some measures of forecast verification. Preprints, *10th Conf. Probability and Statistics in Atmospheric Sciences,* Edmonton, Alberta, Amer. Meteor. Soc., 69–73.

Goldsmith, B. S., 1989: A comprehensive analysis of verification results for forecasts of precipitation type and snow amount. Preprints, *11th Conf. on Probability and Statistics in Atmospheric Sciences,* Monterey, California, Amer. Meteor. Soc., 150–155.

Hays, W. L., 1973: *Statistics for the Social Sciences* (2nd Ed.). Holt, Rinehart, and Winston, Inc., 954 pp.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärke-vorhersagen in Sturmwarnungsdienst. *Geogr. Ann. Stockh.,* **8,** 301–349.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

Murphy, A. H., and E. S. Epstein, 1967: A note on probability forecasts and "hedging". *J. Appl. Meteor.,* **6,** 1002–1004.

——, and H. Daan, 1985: Forecast Evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* Westview Press, 379–437.

Panofsky, H. A., and G. W. Brier, 1958: *Some Applications of Statistics to Meteorology,* Pennsylvania State University Press, p. 200 ff.

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

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