Sensitivity of Several Performance Measures to Displacement Error, Bias, and Event Frequency

MICHAEL E. BALDWIN AND JOHN S. KAIN

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 19 September 2005, in final form 29 November 2005)

ABSTRACT

The sensitivity of various accuracy measures to displacement error, bias, and event frequency is analyzed for a simple hypothetical forecasting situation. Each measure is found to be sensitive to displacement error and bias, but probability of detection and threat score do not change as a function of event frequency. On the other hand, equitable threat score, true skill statistic, and odds ratio skill score behaved differently with changing event frequency. A newly devised measure, here called the bias-adjusted threat score, does not change with varying event frequency and is relatively insensitive to bias. Numerous plots are presented to allow users of these accuracy measures to make quantitative estimates of sensitivities that are relevant to their particular application.

1. Introduction

Performance measures, such as threat score, are widely used as summary measures of forecast quality. There are a variety of measures from which to choose depending on the type of forecast being issued, whether continuous or categorical, probabilistic or deterministic. For example, the Environmental Modeling Center (EMC), a part of the National Centers for Environmental Prediction (NCEP), has primarily used equitable threat and bias scores to quantify the performance of precipitation forecasts from numerical guidance (e.g., Mesinger 1996; Rogers et al. 2001). Typical performance measures provide information on a single aspect of forecast quality, such as forecast *accuracy*. Accuracy was defined by Murphy (1993) as the degree of correspondence between the forecasts and observations. Accuracy is one of the many aspects of forecast quality that can be obtained from the joint distribution of forecasts and observations. Previous research (e.g., Murphy and Winkler 1987; Murphy 1991; Brooks and Doswell 1996) emphasized the dangers of failing to perform a complete analysis of the joint distribution in order to properly diagnose the verification information. However, such a distributions-oriented approach is rarely used in practice due to the complexity and high dimensionality of the joint distribution of forecasts and observations, particularly when multiple forecasting systems are compared.

When selecting an accuracy measure, one must understand the sensitivity characteristics of the score. To what types of errors is the score most sensitive? Does the score encourage biased forecasts? Are false alarms punished more (or less) than missed events? Does the score behave differently for rare events than for more common events? The validity of verification information depends upon such characteristics. Sensitivities of accuracy measures have been considered by several researchers in the past. Mason (1989) examined the sensitivity of the threat score (critical success index) to the observed event frequency as well as the decision threshold, which can be related to the bias. In the framework of his analysis (probability of detection and probability of false detection held constant), the threat score was found to be highly sensitive to these factors. Specifically, commonly occurring events resulted in higher threat scores than rare events, and the threat score was maximized for bias values greater than one (overforecasting). Hamill (1999) discussed the implications of the sensitivity to bias in determining confidence intervals

Corresponding author address: Michael E. Baldwin, OU/CIMMS, 1313 Halley Circle, Norman, OK 73069.
E-mail: mbaldwin@ou.edu

for the threat score. Modifications to the threat score have been proposed in order to reduce the sensitivity to forecast bias (Schaefer 1990; Mesinger and Brill 2004). These modifications were intended to make the threat score more *equitable* (Gandin and Murphy 1992) so that the modified forms do not reward over- or underforecasting of an event, implying that the score should be maximized when the bias is equal to one. However, Marzban (1998) analyzed over a dozen different performance measures and showed that none of them were equitable for rare events under realistic forecasting conditions.

While equitable scores are desirable for some applications, they may be inappropriate measures of forecast *value* (Murphy 1993), which depends on how forecasts are used and how different outcomes affect the user of the forecast information. Forecast value was defined by Murphy (1993) as the benefits of forecast information to a user of the forecast. Each user will have a different level of sensitivity to false alarms and missed events, depending on their individual situation. For certain situations, a biased forecast may, in fact, be more valuable than an unbiased forecast. For example, users that are especially sensitive to missed events will generally find forecasts with a high bias more valuable than unbiased forecasts. The opposite is true for users sensitive to false alarm errors; forecasts with low bias generally provide more value than unbiased forecasts. Thornes and Stephenson (2001) provided an example of the complicated relationship between forecast bias, accuracy, and value for a winter weather forecasting situation. The cost–loss situation for a city deciding whether or not to treat slippery roads was analyzed for two competing forecast providers. Thornes and Stephenson (2001) found that a forecast provider with a bias of 1.8 resulted in greater economic value to the city than a forecast with a bias of 1, even though various accuracy measures showed the unbiased forecast to be preferred.

In the analysis of verification results, the issue of the sensitivity of a particular score to bias is often raised. When comparing competing forecast systems using any of a number of accuracy measures, one is forced to consider whether a higher score is the result of superior performance, or perhaps a characteristic "prejudice" of the accuracy measure in question. It may be impossible to modify a system in order to produce unbiased forecasts, and arbitrary adjustments to the forecasts may be difficult to justify. Information related to the degree of bias sensitivity for a particular performance measure is necessary for proper interpretation of verification information. Of course, bias is one of many aspects of forecast quality (Murphy 1993) that should be considered when analyzing the performance of forecast systems.

More recently, several researchers have emphasized the high degree of sensitivity of performance measures to spatial scale as well as methods of representing observations and forecasts at identical locations (Tustison et al. 2001; Gallus 2002; Mass et al. 2002; Accadia et al. 2003; Gong et al. 2003; Weygandt et al. 2004). These factors are particularly important when comparing forecasts that contain realistic detail (high resolution) with those that are smooth/coarse (low resolution).

To provide more meaningful information, one may desire to decompose forecast errors into separate independent factors. For example, Murphy (1996) decomposed scores related to the mean square error into bias, reliability, and resolution components. For spatial forecasts in particular, one could consider several components of forecast error using a phenomenological or object-oriented approach. Ebert and McBride (2000), following the general idea of Hoffman et al. (1995), designed a technique to decompose errors in forecasts of specific precipitation events into components due to displacement, amplitude, and shape errors. Nachamkin (2004) used event compositing techniques in order to analyze various spatial error characteristics, such as displacement and amplitude errors, for particular mesoscale phenomena.

In this paper, we address spatial forecast errors such as those associated with quantitative precipitation forecasts. The sensitivity of several measures of accuracy to event frequency, bias, and displacement errors will be examined using a hypothetical forecasting situation that can be systematically manipulated. The equitability of each measure will also be examined for a variety of forecasting conditions. The hypothetical forecasting situation is described in section 2. The performance measures are defined in section 3. Analysis of the results is found in section 4, followed by concluding remarks in section 5.

2. Hypothetical forecast situation

The spatial forecast situation, such as forecasting accumulated precipitation greater than a specific threshold, will be modeled using a simple hypothetical example. In Fig. 1, forecast and observed regions of a specific event are represented by circular shapes. Since this is a dichotomous (yes–no) forecast, it can be verified through the use of a 2×2 contingency table (Table 1). In this paper, we will examine the sensitivity of several performance measures to variations in bias B and displacement D errors. Therefore, the various elements of the contingency table must be computed as a func-

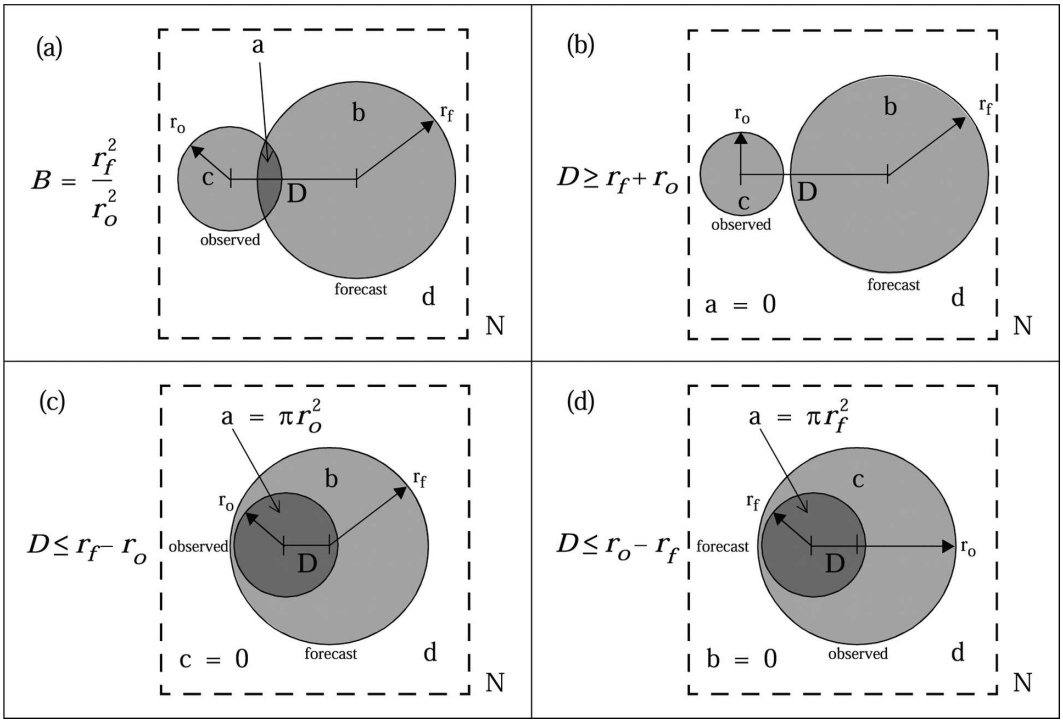

FIG. 1. Hypothetical examples of spatial forecasts and associated observed regions. The radius of each observed circle is r_o, the radius of each forecast circle is r_f, and the distance between the centers of the observed and forecast circles is denoted by D. The total verification domain N is indicated by the dashed outer square, the area of which is held fixed at $N = 1$. The area of correct "yes" forecasts is denoted by a. The region of correct "no" forecasts is d. The region that is forecast but not observed (false alarm) is denoted by b. The area that is observed but not forecast (missed event) is indicated by c. (a) A situation with partial overlap between the forecast and observed circles, with a bias indicated by B. (b) A scenario with no overlap. (c) A situation where the forecast completely envelopes the observed area, with no missed events. (d) A scenario where the forecast is entirely within the observed area, with no false alarms.

tion of B and D. Circular shapes are chosen to allow for the simple calculation of the forecast and observed areas, in addition to the overlap area indicating the region of correct forecasts (hits). To simplify the problem further, the total area of the verification domain $N = a + b + c + d$ is set to 1. As a result, the elements of the contingency table become fractional areas relative to the entire verification domain. While the shape of the overall verification domain is arbitrary, it is convenient to consider a square domain consisting of sides of unit length. In this case, the displacement error and circle radii can be considered fractions of the length scale of the verification domain. The lengths of the observed circle r_o and forecast circle r_f radii are allowed to vary. The centers of the two circles are separated by D. The radius of the observed circle determines the fraction of the verification domain in which the specified event is observed. In this study, this fractional area also represents the event frequency P. The square of the ratio of the forecast radius to the observed radius determines B. To allow for comparison of results with different observed event frequencies, the displacement errors will be normalized relative to the radius of the observed circle, that is, the normalized displacement error $D' = D/r_o$.

TABLE 1. Contingency table for a given event.

		Observed		
		Yes	No	Total
Forecast	Yes	a	b	$a + b$
	No	c	d	$c + d$
	Tot	$a + c$	$b + d$	1

Given the areas of the forecast and observed regions, the area of their intersection, and the area of the verification domain, each element of the contingency table can be computed analytically. The area of a lens created by two intersecting circles (Weisstein 2005) represents the area of correct forecasts, denoted by *a:*

$$a = r_o^2 \cos^{-1}\left(\frac{D^2 + r_o^2 - r_f^2}{2Dr_o}\right) + r_f^2 \cos^{-1}\left(\frac{D^2 + r_f^2 - r_o^2}{2Dr_f}\right) - \frac{1}{2}\sqrt{(r_f + r_o - D)(D + r_f - r_o)(D + r_o - r_f)(D + r_f + r_o)}. \quad (1)$$

Using the formula for the bias ($r_f = \sqrt{Br_o}$), this equation can be rewritten in terms of displacement, bias, and the observed circle radius:

$$a = r_o^2 \cos^{-1}\left[\frac{D^2 + (1 - B)r_o^2}{2Dr_o}\right] + Br_o^2 \cos^{-1}\left[\frac{D^2 + (B - 1)r_o^2}{2D\sqrt{B}r_o}\right] - \frac{1}{2}\sqrt{[r_o(\sqrt{B} + 1) - D][D + r_o(\sqrt{B} - 1)][D + r_o(1 - \sqrt{B})][D + r_o(\sqrt{B} + 1)]}. \quad (2)$$

In general, the total area of the observed circle is $a + c = \pi r_o^2$, and since N = 1, the event frequency $P = a + c$. The area of the forecast circle is $a + b = \pi r_f^2 = B\pi r_o^2$. Therefore, the b and c contingency table elements can be written in terms of D and B and r_o, where a is defined in Eq. (2) above. The d element of the contingency table can be easily computed as a remainder [$d = 1 + a - \pi r_o^2(B + 1)$]. Given that the total domain area is fixed at 1.0, the largest observed circle that could be considered would be $r_o = \sqrt{1/\pi} \cong 0.56$. For a fixed bias, as displacement error increases, the area of overlap will decrease in size until it reaches zero at $D = r_o + r_f$ (Fig. 1b). For a fixed displacement error, as the bias increases, the overlap area will increase until the forecast circle completely envelops the observed circle (Fig. 1c). In this scenario, the area of overlap is equal to the area of the observed circle; therefore, $a = \pi r_o^2$. In addition, there are no missed events and therefore $c = 0$. On the other hand, if the forecast circle is completely enveloped by the observed region (Fig. 1d), the overlap area is equal to the size of the forecast circle, $a = \pi r_f^2$. In this situation there are no false alarms and therefore $b = 0$.

3. Definition of accuracy measures

Several commonly used scores as well as a recently introduced measure of forecast accuracy will be examined. The definitions of the scores in terms of the elements of the 2 × 2 contingency table along with the possible range of values for each are found in Table 2. The elements of the contingency table have been computed in terms of B and D for the hypothetical situation described in section 2. Since many of these scores are commonly used, only a brief description and associated references will be provided here. Probability of detection (POD) is simply the fraction of the observed region that was correctly predicted, which can vary from zero to one. Threat score (TS) is the fraction of the union of observed and forecast areas that was correctly forecast [Gilbert (1884) called this the ratio of verification], which also varies from 0 to 1. Equitable threat score (ETS) adjusts the threat score in order to remove the expected size of the correct forecast area due to random chance [Schaefer (1990) originally called this the Gilbert skill score]. This adjustment allows the score to fall below zero; the minimum ETS (= −1/3) is found when the b and c elements of the 2 × 2 contingency table are both equal to 0.5. The true skill statistic (TSS) is equivalent to the probability of detection minus the probability of false detection [POFD = $b/(b + d)$]. POFD is the ratio of the false alarm area to the section of the domain that did not observe the event, which is also known as the false alarm rate. This should not be confused with the false alarm ratio [FAR = $b/(a + b)$], which is the fraction of the forecast area that did not observe the event. While Doswell et al. (1990) used the term true skill statistic, Stephenson (2000) called this measure the Peirce skill score, in honor of its original discovery by Peirce (1884). In addition, Richardson (2000) called this the Kuipers score, and showed that in the simple cost–loss ratio decision model, TSS is equivalent to the maximum relative economic value that can be obtained from a forecast system if the cost–loss ratio is equal to the event frequency, and ranges from −1 to 1. The odds ratio skill score (ODDS) is a function of the odds ratio (= ad/bc) and varies from −1 to 1 (Stephenson 2000).

Another accuracy measure has been proposed by Mesinger and Brill (2004), known as the bias-adjusted threat score (TSA). TSA adjusts the threat score to account for the impact of bias, with the stated goal of providing information on "the accuracy of the *place-*

TABLE 2. Definitions of performance measures.

Score	Definition	Range
Probability of detection	$\text{POD} = \frac{a}{a+c}$	$0 \le \text{POD} \le 1$
Threat score	$\text{TS} = \frac{a}{a+b+c}$	$0 \le \text{TS} \le 1$
Equitable threat score	$\text{ETS} = \frac{a - a_{\text{rand}}}{a+b+c-a_{\text{rand}}}$, $a_{\text{rand}} = (a+b)(a+c)$	$-\frac{1}{3} \le \text{ETS} \le 1$
True skill statistic	$\text{TSS} = \frac{a}{a+c} - \frac{b}{b+d} = \text{POD} - \text{POFD}$	$-1 \le \text{TSS} \le 1$
Bias-adjusted threat score	$\text{TSA} = \frac{(a+c)^{1/B} - c^{1/B}}{(a+c)^{1/B} + c^{1/B}}$, $B = \frac{a+b}{a+c}$	$0 \le \text{TSA} \le 1$
Odds ratio skill score	$\text{ODDS} = \frac{ad-bc}{ad+bc}$	$-1 \le \text{ODDS} \le 1$

ment of forecast events." To adjust the threat score, a relationship between the number of forecasts and the number of correct forecasts was developed. For spatial forecasts, Mesinger and Brill (2004) assumed that the change in the number of correct forecasts per unit change in the number of forecasts was proportional to the size of the missed event region. This assumption results in an exponential relationship between the number of correct forecasts (a) and the number of forecasts issued ($F \equiv a + b$):

$$a(F) = (a + c)[1 - \exp(-\alpha F)]. \quad (3)$$

The parameter α can be determined given sample verification data from a biased forecasting system and can be thought of as an inherent measure of performance. After α is determined and assuming that it remains constant for a particular forecasting system, one can estimate the number of correct forecasts corresponding to an unbiased forecast by setting $F = a + c$ in Eq. (3). Using this bias-adjusted number of correct forecasts, the bias-adjusted threat score can be calculated, and is defined in Table 2 in terms of the elements of the 2×2 contingency table. TSA is maximized for $c = 0$ (POD = 1). For an unbiased forecast ($B = 1$), TSA and TS are equivalent, since $b = c$ when $B = 1$.

4. Results

Results are presented as a function of event frequency, P. The rare event scenario is considered first, followed by common, then very common events.

a. Rare event

Figure 2 shows how the accuracy measures vary as a function of B and D for a rare observed event ($P \cong 0.03$, $r_o = 0.1$). As expected, for a fixed D', POD (Fig. 2a) increases as B increases, since the forecast circle enlarges until it eventually "swallows" the observed circle at POD = 1. For a fixed bias, POD decreases as D' increases, because the overlap area quickly decreases to zero as the circles move away from each other. The four scenarios portrayed in Fig. 1 produce regions within the POD contour plot that contain specific characteristics. For instance, POD contours are parallel to the D' axis in the region of the plot where D' and B are small (lower-left quadrant in Fig. 2a). In this region the forecast circle is completely enveloped within the observed circle (e.g., Fig. 1d). In this situation, the b element of the contingency table is zero, and POD = B. On the other hand, POD = 1 for large values of B and small values of D' (lower-right quadrant). This is the region where the observed circle is contained within the forecast circle (e.g., Fig. 1c). Contours of many of the other accuracy measures in this region are parallel to the D' axis (Fig. 2). For relatively large values of D', there is no overlap between the forecast and observed circles (e.g., Fig. 1b) and POD = 0 (upper-left quadrant). In the region where POD varies between zero and one and the contours are not parallel to the D' axis, all four elements of the contingency table are nonzero (as in Fig. 1a).

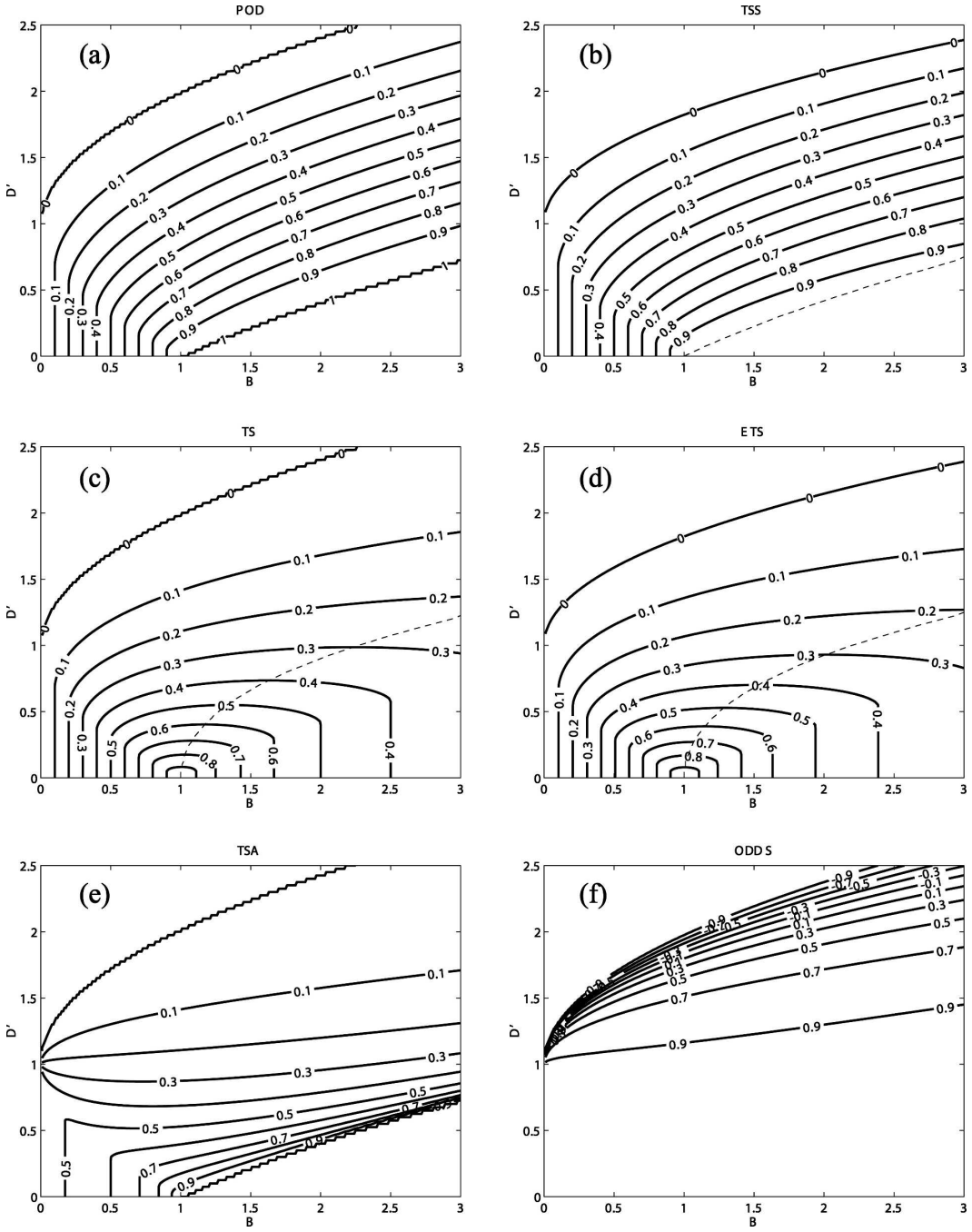

FIG. 2. For the rare event situation ($P \cong 0.03$, observed circle radius $r_o = 0.1$), accuracy measures as a function of bias (B) and normalized displacement error (D'): (a) POD, (c) TS, and (e) TSA; (b) TSS, (d) ETS, and (f) ODDS. Axis of maximum score value is indicated in each plot by a dashed line. Contour interval is 0.1 in each panel except for ODDS, which uses a 0.2 contour interval.

For this rare event scenario, TSS and POD are nearly identical (cf. Figs. 2a and 2b). TSS shows very little sensitivity to false alarms in this case, since POFD remains insignificant with increasing B for the range of values analyzed here. The axis of maximum TSS increases nearly linearly in B as D' increases, closely following the POD = 1 contour.

In this rare event situation, the expected size of the correct forecast region due to random chance is very small; therefore, TS and ETS (Figs. 2c and 2d) are prac-

tically equal. These scores are maximized at $D' = 0$ and $B = 1$. Generally, the scores drop as D' increases. For a fixed D', maximum scores are found at B values greater than one for all $D' > 0$, and the B that is associated with the maximum score increases with increasing D'. As with POD, contours of each of these scores are parallel to the D' axis when B and D' are relatively small (lower-left quadrant) because the forecast area is completely enclosed within the observed area. Similarly, the contours are vertical in the lower-right quadrant because the observed area is located entirely within the forecast area.

By comparing Figs. 2c and 2e, one can readily see the impact of the bias adjustment that distinguishes TSA from TS. TSA is generally larger than TS for $B < 1$, and lower than TS for $B > 1$. For the most part, the TSA contours slope in the positive B direction, indicating that forecasts with higher biases will generally be preferred by this measure. While TSA displays a dependence on B, this sensitivity is relatively small over certain regions of the D'–B phase space. Notably, compared to TS, TSA is considerably less sensitive to B in the region where B varies from about 0.5 to 1.5, especially where D' is close to 1. This region of the phase space is particularly relevant because operational forecast systems tend to be calibrated to produce B as close to one as possible. Outside of this region, however, TSA has some undesirable characteristics. For example, it is very sensitive to D' as it approaches its maximum value in the lower-right quadrant in Fig. 2e. As previously noted, TSA is maximized when POD = 1, which occurs in this quadrant (cf. Figs. 2c and 2e).

ODDS (Fig. 2f) behaves quite differently than the other accuracy measures. Values of ODDS are near their maximum across the entire range of B for all $D' < 1$. This is due to the fact that, for rare events, the 2×2 contingency table is dominated by the d element (correct nulls) of the contingency table. For $P \cong 0.03$, the a, b, and c elements are $O(10^{-2})$, while d remains close to 1.0. Therefore, $ad \gg bc$ and ODDS becomes approximately ~1. ODDS becomes very sensitive to displacement error as D' increases above 1, dropping off quickly to the minimum value of the score for D' greater than ~2.

b. Common event

Figure 3 displays the various accuracy measures as a function of B and D' for a relatively common observed event ($P \cong 0.28$, $r_o = 0.3$). Since the observed circle has grown in size, in the upper-right quadrant of the figures where the bias and normalized displacement errors are large, the union of the observed and forecast circles would result in a total area larger than the fixed verification domain size ($N = 1$). Therefore, accuracy measures cannot be computed in this region.

As in the rare event case, for a fixed D′, POD (Fig. 3a) increases with increasing B. In fact, POD, TS, and TSA (Figs. 3a, 3c, and 3e) behave in exactly the same manner as in the rare event case (Figs. 2a, 3c, and 3e). This is not surprising, since the scores are plotted in normalized displacement error space. For example, consider the situation where the observed and forecast circles are the same size and are separated by the length of the observed circle radius (Fig. 4, $B = 1$, $D' = 1$). In terms of absolute displacement error (D), the larger circles are farther apart than the smaller circles. However, in this situation, the relative arrangement of circles remains constant as long as B and D' are held fixed. Therefore, the fraction of the observed area that is correctly forecast (POD ≅ 0.39), as well as the fraction of the union of observed and forecast areas that is correct (TS ≅ 0.24), will remain unchanged for a given B and D' as the size of the observed circle changes. Since TSA is a function of POD and B, it also remains unchanged for a given B and D' as P changes.

In contrast, ETS, TSS, and ODDS all show a sensitivity to P. For example, while TS and ETS behave similarly for rare observed events, ETS is quite different from TS in the relatively common event situation (cf. Fig. 3d to Fig. 3c). This difference in behavior is due to the fact that the expected amount of correct random forecasts increases with event frequency. The axis of maximum TS slopes in the positive B direction, while the maximum ETS scores are found consistently near $B = 1$ for a wide range of D' values, indicating a high degree of equitability for ETS at this level of P. ETS has also become more sensitive to D' as P has increased. For a fixed B, ETS decreases more rapidly with increasing D' in the common event case (Fig. 3d) than in the rare event case (Fig. 2d).

In contrast to the rare event scenario, TSS behaves very differently from POD in this relatively common event analysis (cf. Figs. 3a and 3b). This difference occurs because the effect of POFD becomes significant in the computation of TSS, since the size of the observed "no" region (denominator of POFD) is considerably smaller in this situation. The value of B that produces the maximum TSS score increases from $B = 1$ to $B \cong 1.5$ as D' increases from zero to 0.8, then decreases to zero bias at $D' \sim 1.2$. Similarly to ETS, the sensitivity of TSS to D' increases in the common event case.

ODDS (Fig. 3f) behaves quite differently from the other scores, with little variation in score for fixed D' as

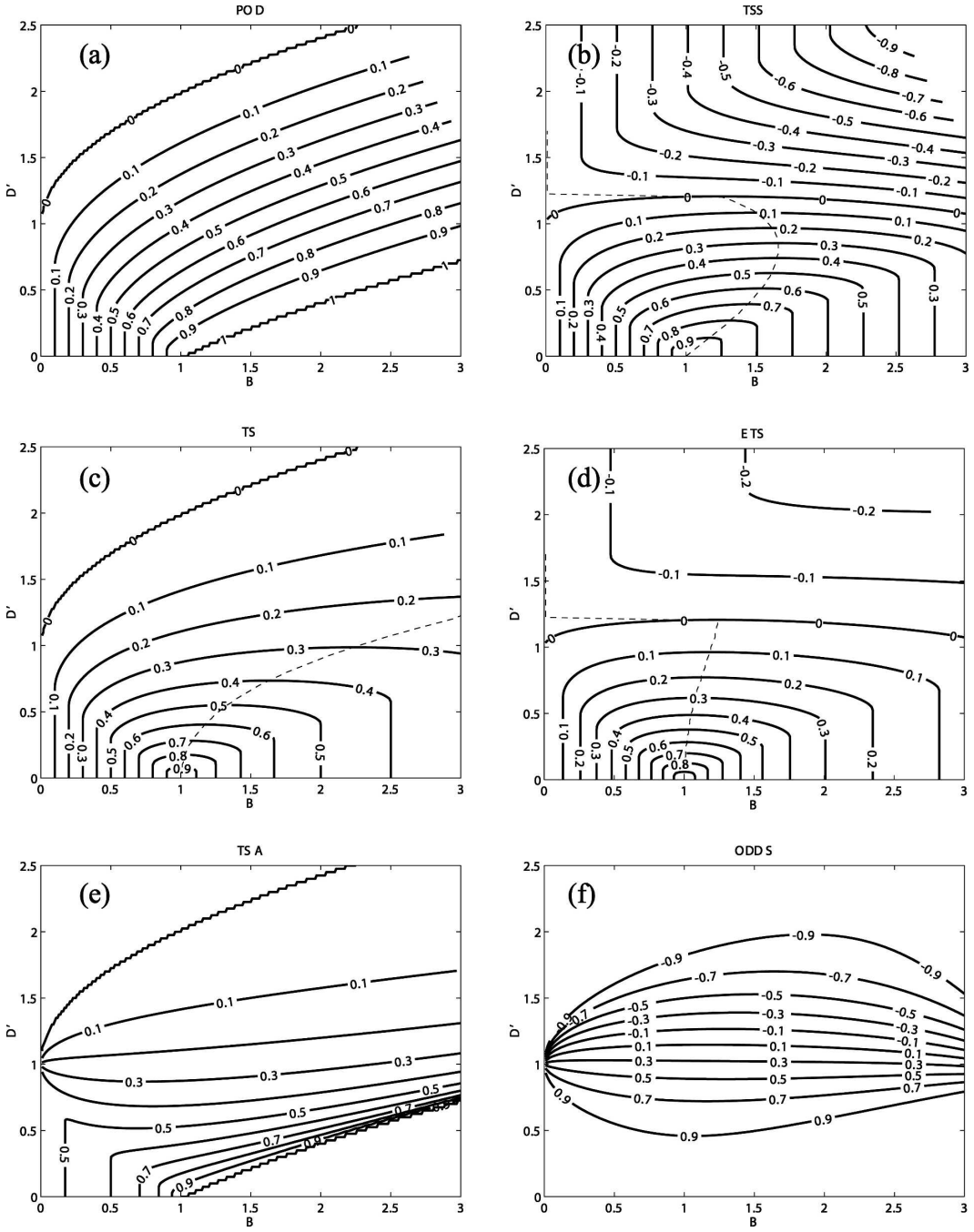

FIG. 3. As in Fig. 2 except for the relatively common event situation ($P \cong 0.28$, observed circle radius $r_o = 0.3$).

B changes for a wide range of D'. In fact, for the range of D' between zero and one, ODDS reaches a *minimum* value near $B = 1$, as opposed to a maximum value for the other accuracy measures. The ODDS minimum near $B = 1$ occurs because the denominator in the ODDS formula ($ad + bc$) reaches a relative maximum in this region of the D'–B phase space for the common event situation.

c. Very common event

Figure 5 displays the various accuracy measures as a function of bias and displacement errors for a very common observed event ($P \cong 0.50$, $r_o = 0.4$). Since the observed circle covers slightly more than one-half of the verification domain, accuracy measures cannot be computed over a considerable portion of the upper-

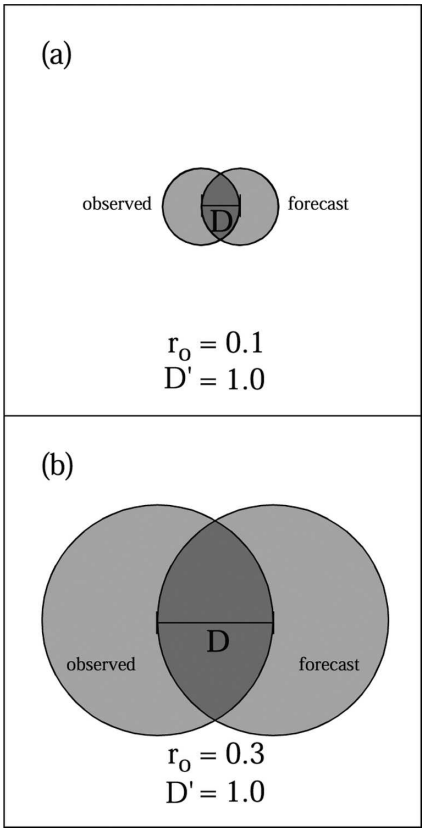

FIG. 4. Examples of holding B and D' fixed while changing P (or r_o): (a) $P \cong 0.03$, $r_o = 0.1$, $B = 1$, and (b) $P \cong 0.28$, $r_o = 0.3$, $B = 1$.

right quadrant of the figures. For reasons discussed previously, the behavior of POD, TS, and TSA (Figs. 5a, 5c, and 5e) is the same as it was with smaller values of P.

In this very common event situation, TSS and ETS (Figs. 5b and 5d) display quite different behavior than they showed in the rare (Figs. 2b and 2d) and relatively common (Figs. 3b and 3d) event cases. In this situation, the ETS and TSS scores are maximized for B values less than one for all $D' > 0$. The value of B that is associated with the maximum score decreases with increasing displacement error, until it reaches $B = 0$ at $D' \cong 1$. ETS is very sensitive to the expected amount of correct random forecasts a_r in the very common event scenario. Specifically, when $P \cong 0.50$, a_r is ~50% of the forecasted area. Therefore, in order for the numerator in the ETS formula to increase with B for a given D', more than 50% of the additional forecast area resulting from an increase in B must be correct. In addition, since $P \cong 0.50$, $a + c \cong b + d$; therefore, TSS simplifies to $(a - b)/P$. Similarly, TSS will increase with B only when more than half of the additional forecast area is correct. This can only occur in those situations where the observed area nearly encompasses the forecasted area (small D' and B), which explains why the maximum ETS and TSS scores are found in this region. The sensitivity of ETS and TSS to D' has increased further in this very common event case; that is, the spacing between contours in the D' direction continues to shrink as P increases (e.g., cf. Figs. 2d, 3d, and 5d).

Once again, ODDS (Fig. 5f) behaves differently than the other accuracy measures. Values of ODDS are near their maximum across the entire range of B for all $D' < 0.5$. In this region, either the forecast area is surrounded by the observed area ($b = 0$, lower left) or POD = 1 ($c = 0$, lower right), resulting in ODDS = 1. Beyond this region, many of the ODDS contours slope in the negative B direction, as do the ETS and TSS contours, indicating a prejudice by these measures toward forecasts with smaller B values in this very common event situation.

5. Discussion

The results demonstrate that POD, TS, and TSA display consistent behavior as P changes, while other measures, such as TSS and ETS, show considerable sensitivity to event frequency. For example, when B and D' are held constant at 1.0 so that the degree of overlap between forecast and observed regions does not change (as in Fig. 4), as P is varied from 0.03 to 0.28 TSS changes from 0.37 to 0.15, while TS remains fixed at 0.24. At first glance, these results appear to contradict the work of Mason (1989) who showed that TSS was independent of P while TS was quite sensitive to it. These differences can be explained by contrasting Mason's experimental design to that used here. Our strategy is to examine the sensitivity to P without changing the relative spatial configuration of forecast and observed areas. For a given B and D', P is changed simply by modifying the scale of the combined forecast and observed circles relative to the fixed verification domain (see Fig. 4). In this approach, the relative arrangement of the forecast and observed circles is not altered; therefore, the ratio of the overlap area to the union of the forecast and observed areas (TS) remains unchanged. Similarly, POD, TSA, and FAR do not change for a given B and D' as P varies. However, POFD does change with P under these conditions.

In contrast, Mason (1989) assumed that POD and POFD remained fixed when P varied. Although this strategy results in a constant TSS and ODDS with varying P, its implications for the relative spatial configuration of forecast and observed areas are not obvious. Holding POD and POFD constant while changing P requires that both B and D' change. This is difficult to

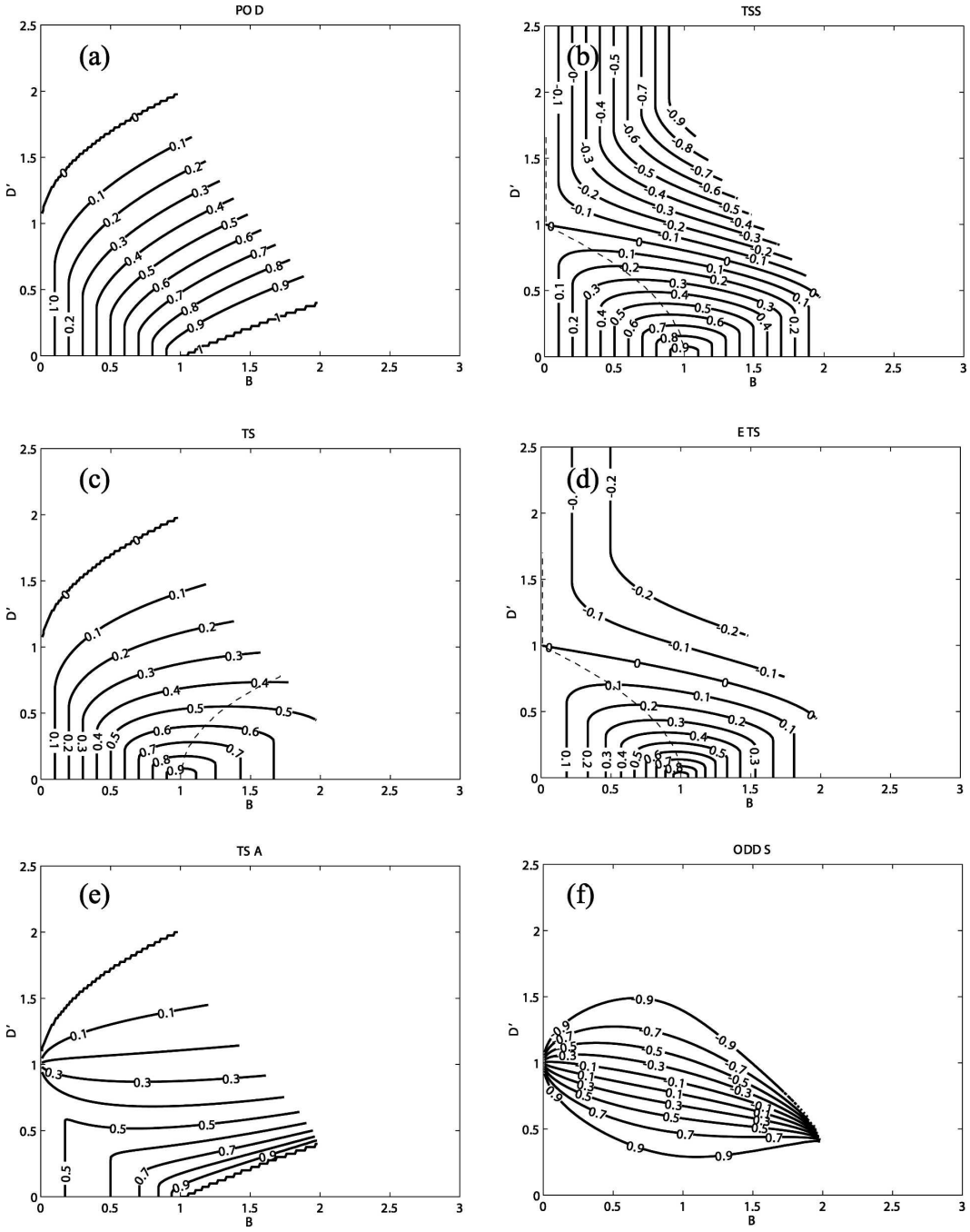

FIG. 5. As in Fig. 2 except for the very common event situation ($P \cong 0.50$, observed circle radius $r_o = 0.4$).

conceptualize, but is demonstrated graphically in Fig. 6. In this figure, POD and POFD maintain the same values (POD = 0.7, POFD = 0.1) as the observed circle increases in coverage. To generate these POD and POFD scores in the rare event situation (Fig. 6a; $P = 0.1$) the forecast circle must be larger than the observed circle ($B = 1.6$), with a D' of 0.78. In this case TS is 0.37. As the observed event increases in coverage (Fig. 6b; $P = 0.3$), the size of the region that does not observe the event decreases; therefore, the size of the forecast area must shrink as well in order to maintain a constant POFD. In this scenario B has decreased to 0.93. In addition, in order to keep POD constant with a smaller forecast circle, D' must decrease to 0.4. In this case, the threat score has increased to 0.57. These trends continue as the observed event increases in coverage (Fig.

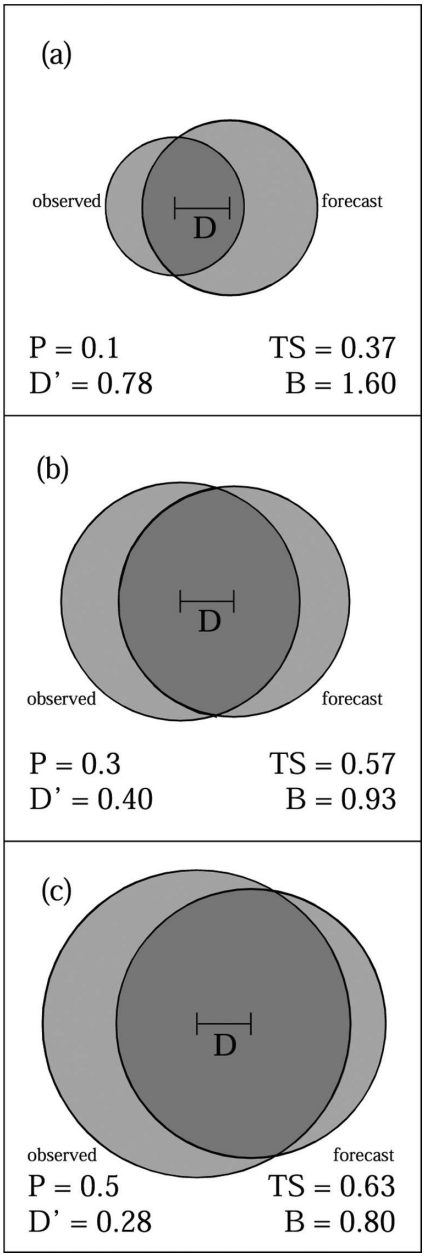

FIG. 6. Examples of holding POD (= 0.7) and POFD (= 0.1) fixed while changing P: (a) $P = 0.1$, $B = 1.6$, $D' = 0.78$; (b) $P = 0.3$, $B = 0.93$, $D' = 0.40$; and (c) $P = 0.5$, $B = 0.8$, $D' = 0.28$.

6c; $P = 0.5$), B decreases to 0.8, D' decreases further to 0.28, and TS increases to 0.63. Although D' decreases with increasing P with fixed POD and POFD, it is interesting to note that the *absolute* displacement error remains nearly constant in this example ($D \approx 0.1$). Clearly, any conclusions regarding the sensitivity of TSS and TS to P depend upon the strategy used to vary P.

6. Summary

In this work, the sensitivity of five commonly used performance measures to displacement error, bias, and event frequency was analyzed for hypothetical forecasting scenarios. A newly developed measure was examined as well. The different scenarios were expressed using two circles: one representing the area in which the event was observed, the other the area where the event was predicted to occur. This simple model allowed for full control and manipulation of the event frequency, bias, and displacement errors as well as detailed analysis of the sensitivity of the scores to these factors.

The behavior of several scores (ETS, TSS, ODDS) changed considerably as event size/frequency changed. In general, for rare observed events, these accuracy measures were maximized for bias greater than one, indicating that the scores encouraged "hedging" toward overforecasting. This behavior implies that the scores were more sensitive to missed events than false alarms. For larger, relatively common events, these scores were shown to be less sensitive to bias, displaying nearly constant values across a wide range of biases for a given displacement error. This behavior indicates that these scores punished missed events and false alarms similarly. However, for very common events, these scores were maximized for bias values less than one. In this case, the scores were more sensitive to false alarms than missed events. These measures were also generally found to be more sensitive to displacement error as event frequency increased. When using ETS, TSS, or ODDS to compare multiple forecasts, users should be aware of the considerable sensitivity to P. Thus, the same event frequency should be used for all forecast systems, and the same geographical domain should be used for all forecasts since changing the domain size effectively changes P.

The sensitivities of POD, TS, and TSA to changes in bias and displacement error were shown to be independent of observed event frequency. For example, POD and TS both showed a sensitivity to bias, with higher bias resulting in higher scores. TSA showed less sensitivity to bias over a significant portion of the bias-displacement phase space, along with undesirable sensitivities in some other regions of this space.

The insensitivity of TS to changes in event frequency appears to contradict the work of Mason (1989). This apparent contradiction can be reconciled by contrasting Mason's experimental design to that used in this work. Our strategy was to maintain the relative spatial arrangement of the forecast and observed areas with

changing event frequency, while Mason (1989) maintained a constant POD and POFD with varying P. To keep POD and POFD fixed, the relative spatial configuration of the forecast and observed regions must change when event frequency changes. Any conclusions regarding the sensitivity of accuracy measures to event frequency depend upon the assumptions used in the experimental design.

Although the two-circle scenario utilized in this study is idealized, the general conclusions that it allows us to draw are not strongly dependent on geometry. The simple two-circle model allows us to clarify and conceptualize the sensitivities of various performance measures in a quantitative way. For example, most users of these scores have long recognized the sensitivity to bias, but the character of this sensitivity has remained poorly understood for too long. The plots contained herein provide a useful reference for specific applications. For example, if one can estimate P, D, and B for a given situation, he or she can estimate the sensitivity of scores to changes in any one of these variables by referring to the appropriate plots. In general, except for a few scores under very specific conditions, all of the performance measures analyzed in this work encouraged hedging of one form or another.

Every user of forecast information has a distinct amount of sensitivity to the different types of errors that can be realized in a dichotomous forecast. This sensitivity will vary depending upon the weather event, the "expenses" suffered by the user resulting from false alarms or missed events, etc. To provide information that is consistent with a user's impression of the value of a forecast, a performance measure must be sensitive to the various errors in a manner that is consistent with that particular user. Regardless of whether or not a score provides information that is consistent with value, it is important for users of verification information to understand the preferences and prejudices of the measures used to evaluate forecasting systems.

Acknowledgments. This manuscript benefited greatly from the constructive comments and suggestions provided in a preliminary review by Dr. Kimberly Elmore of the National Severe Storms Laboratory. Extensive discussions on verification-related issues with Dr. Fedor Mesinger (Environmental Modeling Center) and Keith Brill (Hydrometeorological Prediction Center) were very helpful and increased our understanding of many of the issues raised here. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA17RJ1227, U.S. Department of Commerce.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting,* **18,** 918–932.

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.

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.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

Gallus, W. A. J., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting,* **17,** 1296–1302.

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

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

Gong, X., A. G. Barnston, and M. N. Ward, 2003: The effect of spatial aggregation on the skill of seasonal precipitation forecasts. *J. Climate,* **16,** 3059–3071.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Hoffman, R. N., Z. Liu, J.-F. Louis, and C. Grassoti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.,* **123,** 2758–2770.

Marzban, C., 1998: Scalar measures of performance in rare event situations. *Wea. Forecasting,* **13,** 753–763.

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

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.,* **83,** 407–430.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.,* **77,** 2637–2650.

——, and K. Brill, 2004: Bias normalized precipitation scores. Preprints, *17th Conf. on Probability and Statistics,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, J12.6.

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: General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Mon. Wea. Rev.,* **124,** 2353–2369.

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

Nachamkin, J. E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.,* **132,** 941–955.

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

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

Rogers, E., T. Black, B. Ferrier, Y. Lin, D. Parrish, and G. DiMego, 2001: Changes to the NCEP Meso Eta Analysis and Forecast System: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3DVAR analysis. NWS Tech. Procedures Bull. 488, 15 pp. [Available online at http://www.emc.ncep.noaa.gov/mmb/mmbpll/eta12tpb/ or from Office of Meteorology, National Weather Service, 1325 East–West Highway, Silver Spring, MD 20910.]

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

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

Thornes, J. E., and D. B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteor. Appl.,* **8,** 307–314.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.,* **106,** 11 775–11 784.

Weisstein, E. W., cited 2005: Circle–circle intersection. *Mathworld*, Wolfram. [Available online at http://mathworld.wolfram.com/Circle-CircleIntersection.html.]

Weygandt, S. S., A. F. Loughe, S. G. Benjamin, and J. L. Mahoney, 2004: Scale sensitivities in model precipitation skill scores during IHOP. Preprints, *22d Conf. on Severe Local Storms,* Hyannis, MA, Amer. Meteor. Soc., CD-ROM, 16A.8.