The Response of Performance Metrics for Binary Forecasts to Hedging that Approaches Random Change

KEITH F. BRILL

Hydrometeorological Prediction Center, National Centers for Environmental Prediction, Camp Springs, Maryland

MATTHEW PYLE

Environmental Modeling Center, National Centers for Environmental Prediction, Camp Springs, Maryland

(Manuscript received 5 November 2009, in final form 19 February 2010)

ABSTRACT

Critical performance ratio (CPR) expressions for the eight conditional probabilities associated with the 2×2 contingency table of outcomes for binary (dichotomous ''yes'' or ''no'') forecasts are derived. Two are shown to be useful in evaluating the effects of hedging as it approaches random change. The CPR quantifies how the probability of detection (POD) must change as frequency bias changes, so that a performance measure (or conditional probability) indicates an improved forecast for a given value of frequency bias. If yes forecasts were to be increased randomly, the probability of additional correct forecasts (hits) is given by the detection failure ratio (DFR). If the DFR for a performance measure is greater than the CPR, the forecast is likely to be improved by the *random* increase in yes forecasts. Thus, the DFR provides a benchmark for the CPR in the case of frequency bias inflation. If yes forecasts are decreased randomly, the probability of removing a hit is given by the frequency of hits (FOH). If the FOH for a performance measure is less than the CPR, the forecast is likely to be improved by the random decrease in yes forecasts. Therefore, the FOH serves as a benchmark for the CPR if the frequency bias is decreased. The closer the FOH (DFR) is to being less (greater) than or equal to the CPR, the more likely it may be to enhance the performance measure by decreasing (increasing) the frequency bias. It is shown that randomly increasing yes forecasts for a forecast that is itself better than a randomly generated forecast can improve the threat score but is not likely to improve the equitable threat score. The equitable threat score is recommended instead of the threat score whenever possible.

1. Introduction

This note systematically applies the bias sensitivity analysis method of Brill (2009) to the eight conditional probabilities associated with binary (dichotomous ''yes'' or ''no'') forecasts and examines implications for random or near-random changes to forecasts. The critical performance ratio (CPR; Brill 2009) quantifies the sensitivity of a performance measure for binary forecasts to either increasing or decreasing the frequency bias. The CPR is the minimum increase in probability of detection (POD) per increase in frequency bias for the performance measure to indicate improved forecasts. In the opposite case, the CPR gives the maximum allowed decrease in POD per decrease in frequency bias for the performance measure to indicate improvement. Hereafter, the word ''bias'' is taken to mean frequency bias, the ratio of the yes forecast count or area to the yes observed count or area. The CPR may be used to assess the potential benefits of ''hedging'' (e.g., Marzban 1998; Brill 2009) by deliberately altering the bias of a forecast (changing otherwise intended yes and no forecasts) based on knowledge of past performance assessed by some measurement or score along with bias.

The CPR analysis applies to verification of past forecasts. There are two ramifications of this: 1) results are relevant to the verification of future forecasts only to the extent that past performance is indicative of future performance for similar meteorological situations, and 2) the observed event frequency (base rate) in the CPR analysis may be treated as a constant. The latter condition

Corresponding author address: Keith F. Brill, NCEP/HPC, W/NP32, NOAA Science Center, Rm. 410B-2, 5200 Auth Rd., Camp Springs, MD 20746-4304.
E-mail: keith.brill@noaa.gov

DOI: 10.1175/2010WAF2222381.1

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

assures that the ratio of the change in POD to a change in bias is the same as the change in correct yes forecasts (hits) to the total change in yes forecasts (Brill 2009). In other words, the CPR expresses a hit fraction for added or removed yes forecasts. In keeping with the first ramification, the CPR analysis is applied to address the ''what if'' questions associated with hedging: What if the forecaster had changed some no forecasts to yes forecasts to increase the bias? Or, what if the forecaster had changed some yes forecasts to no, thereby decreasing the bias? Specifically, how might a performance measure respond to such deliberate modifications, and what condition must be satisfied for that performance measure to give an indication of improvement?

Consider a hypothetical collection of yes–no forecasts and observed results completely populating the 2×2 contingency table from which some performance measure S is computed. Further, suppose that the CPR for S is computed and found to have the value 0.10. What if the forecaster had added additional yes forecasts by changing some of the no forecasts? The CPR analysis indicates that if more than 10% of those added yes forecasts turned out to be hits, then S would show improvement. On the other hand, what if the forecaster had changed some of the yes forecasts to no forecasts? The CPR analysis says that if more than 10% of those changed forecasts turned out to be hits, then S would indicate a degraded forecast. There is an interesting asymmetry apparent from this example: If yes forecasts are increased, only 10% of them need to be correctly chosen to keep S from indicating degraded performance. However, if no forecasts are increased (by decreasing yes forecasts), 90% of them must be chosen correctly.

Another useful example to consider is the problem of comparing two different forecasting systems or models having different bias characteristics. If the system having the higher bias is to be bias corrected for the comparison, then choosing a performance metric with a high CPR will likely favor the model undergoing bias deflation. For example, if the CPR for the performance metric is 0.45, then up to 45% of the yes forecasts changed in bias correcting the overbiased system can be hits without degrading the indicated performance. Unless the bias correction results in bias equality between the two systems, the overbiased system could gain an advantage in the comparison after bias correction.

The preceding examples demonstrate the quantitative meaning of the CPR. The CPR analysis cannot reveal definitively whether or not a performance measure would have indicated improvement had the bias been changed, but it gives the condition that must be met for such an indication of improvement. The CPR analysis also reveals an important aspect of the bias sensitivity of the performance measures: the more resistant a measure is to hedging by inflating the bias, the less resistant it is to hedging by deflating the bias, and vice versa.

Of specific interest in this note are the eight conditional probabilities associated with the 2×2 contingency table for dichotomous forecasts and their CPR formulas. As denominated by Doswell et al. (1990), the conditional probabilities are probability of detection (POD), false alarm ratio (FAR), frequency of misses (FOM), probability of a null event (PON), frequency of hits (FOH), probability of false detection (POFD), detection failure ratio (DFR), and frequency of correct null forecasts (FOCN). The nomenclature of Doswell et al. (1990) is not unique: The FOH is also known as postagreement and is identified as the true positive ratio (TPR) by Jolliffe and Stephenson (2003). The POFD is sometimes called the false alarm rate (here FAL; see Jolliffe and Stephenson 2003). The DFR is termed the conditional miss rate by Stephenson (2000) and the miss ratio (MR) by Jolliffe and Stephenson (2003). It is beyond the scope of this note to propose a standard terminology for these conditional probabilities. As long as it is understood that the probabilities are conditional on either forecast events or observed events (not both) and are not marginal probabilities, the Doswell et al. (1990) nomenclature serves adequately. CPR evaluations for these conditional probabilities are of interest because at least two of them, POD and FAR [already examined by Brill (2009) but included here for completeness], are sometimes used directly as performance measures and because two of them (DFR and FOH) are shown here to provide benchmarks for the CPR of any performance measure.

It is useful to have some measure of how the number of hits could change by the random conversion of no to yes forecasts to serve as a benchmark for the CPR. A goal of this work is to demonstrate that the DFR serves this purpose. Likewise, it is useful to have a similar benchmark for random conversion of yes forecasts to no forecasts. It is shown herein that the FOH serves this purpose. It is complementary to consider the Clayton skill score (CSS; Wilks 2006), CSS = FOH − DFR, which measures the width of the ''window'' between these two benchmarks.

In this note, the concept of benchmarks for the CPR is applied to investigate the response of the threat score (TS), equitable threat score (ETS), and CSS to the limiting case of random increases or decreases in yes forecast points or area. Section 2 presents the 2×2 contingency table in the form useful for this discussion, derives the CPR formulas for the eight conditional probabilities associated with the 2×2 contingency table, and discusses the DFR and FOH benchmarks. Section 3 applies

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

TABLE 1. Contingency table in terms of cell frequencies a, b, c, and d along with the corresponding expressions in terms of the bias, B; probability of detection, P; and base rate, α.

Events	Observed	Not observed	Total
Forecast	$a = \alpha P$	$b = \alpha(B - P)$	$a + b = \alpha B$
Not forecast	$c = \alpha(1 - P)$	$d = 1 - \alpha \times (B + 1 - P)$	$c + d = 1 - \alpha B$
Total	$a + c = \alpha$	$b + d = 1 - \alpha$	1

the results of section 2 to TS, ETS, and CSS, providing an example of the DFR and FOH as benchmarks for the CPR in the verification of high-resolution quantitative precipitation forecasts (QPFs) from a National Centers for Environmental Prediction (NCEP) regional model. Finally, section 4 presents a brief summary.

2. The DFR and FOH benchmarks for CPR

In the literature, the 2×2 contingency table appears in various forms. The table is written here in terms of both the frequencies of occurrence in each cell (a, b, c, and d, such that $a + b + c + d = 1$) and the parameters utilized by Brill (2009): POD, P; frequency bias, B; and event frequency or base rate, α. Both forms of the contingency table are shown together in Table 1. With the help of Table 1, the formulas for the eight conditional probabilities (see Doswell et al. 1990) along with formulas for the TS, ETS, and CSS are expressed in Table 2 as functions of a, b, c, and d and as functions of P, B, and α.

The orientation of a performance measure or conditional probability, $S(P, B, \alpha)$, is determined by how it changes to indicate an improved forecast. If S increases with forecast improvement, then it is positively oriented; otherwise, it is negatively oriented (Wilks 2006). A stipulation for the CPR analysis is that $(\partial S/\partial P) > 0$ for positively oriented performance measures and <0 for negatively oriented performance measures (Brill 2009). This latter condition could serve as the definition for the orientation of a performance measure (F. Mesinger 2008, personal communication), keeping in mind that the partial derivative is taken in the usual way, that is, with the other independent variables (B and α) held constant.

The general formulation for the CPR as derived by Brill (2009) is

$$\mathrm{CPR} = -\frac{\left(\frac{\partial S}{\partial B}\right)}{\left(\frac{\partial S}{\partial P}\right)}. \tag{1}$$

The formulas for S, the CPR, and the constituent partial derivatives are shown in Table 2. While there are eight conditional probabilities, there are only four unique formulas of the CPR for these, and there is an equal split between positive and negative orientations. Four of the eight conditional probabilities have the trivial values of 0 or 1. For two of them, DFR and FOH, the CPR formula is identical to the formula for the conditional probability itself. These two have special significance as benchmarks for random changes to forecasts and are examined in more detail in the following discussion.

The probability of randomly hitting additional observations by increasing yes forecasts is the number of observations not hit divided by the total number of no forecasts. In Fig. 1, this is the ratio of the unshaded hatched observed area to the entire unshaded area outside of the yes forecast area within the total rectangular verification domain. In terms of Table 1, this is given by

$$\frac{c}{c + d} = \frac{\alpha(1 - P)}{1 - \alpha B} = \mathrm{DFR}. \tag{2}$$

As discussed in Brill (2009), if α is constant, the CPR for a performance measure is a ratio of the change in correct yes forecasts (hits, h) to the change in yes forecasts (f), $\Delta h/\Delta f$. If yes forecasts are increased and the ratio of new hits to new yes forecasts exceeds the CPR, the performance measure indicates improvement. Since the DFR gives the probability of hits for the random increase in yes forecasts, the number of additional hits expected by chance is $\Delta h_A = \Delta f_A \times \mathrm{DFR}$, where the subscript denotes the addition of yes forecasts. Therefore, the DFR can be compared directly to the CPR for any performance measure. If the DFR exceeds the CPR, it is likely that a random increase in yes forecasts will lead to an indication of improvement by the performance measure. Therefore, the DFR may serve as a lower bounding benchmark for the CPR. The greater the CPR minus DFR difference, the more difficult successful hedging by increasing the bias is likely to be for such forecasts. Since hedging (as defined in the introduction) is based on past verification, a performance measure whose CPR greatly exceeds the DFR can be chosen to discourage hedging by increasing the bias.

If yes forecasts are randomly changed to no forecasts to decrease the bias, Fig. 1 shows that the probability of randomly selecting a hit for removal is the number of yes forecasts that are hits divided by the total number of yes forecasts. In terms of Table 1, this probability is

$$\frac{a}{a + b} = \frac{P}{B} = \mathrm{FOH}. \tag{3}$$

Like the DFR in Eq. (2), the FOH CPR, derived using Eq. (1), is seen to have the same formula as FOH itself. The number of hits expected to be lost by randomly decreasing yes forecasts is $\Delta h_R = \Delta f_R \times \mathrm{FOH}$, where the

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

TABLE 2. Table of performance measure (PM) formulas, partial derivatives with respect to B and P, and the CPR formulas. Here, "Or" refers to the orientation (see text). All other symbols and abbreviations are defined in the text.

PM	Or	$S(a, b, c, d)$	$S(B, P)$	$\partial S/\partial B$	$\partial S/\partial P$	CPR
POD	+	$a/(a+c)$	P	0	1	0
FAR	−	$b/(a+b)$	$1-P/B$	P/B^2	$-1/B$	P/B
FOM	−	$c/(a+c)$	$1-P$	0	-1	0
PON	+	$d/(b+d)$	$\frac{1-\alpha B-\alpha+\alpha P}{1-\alpha}$	$\frac{-\alpha}{1-\alpha}$	$\frac{\alpha}{1-\alpha}$	1
FOH = TPR	+	$a/(a+b)$	P/B	$-P/B^2$	$1/B$	P/B
POFD = FAL	−	$b/(b+d)$	$\frac{\alpha(B-P)}{1-\alpha}$	$\frac{\alpha}{1-\alpha}$	$\frac{-\alpha}{1-\alpha}$	1
DFR = MR	−	$c/(c+d)$	$\frac{\alpha(1-P)}{1-\alpha B}$	$\frac{\alpha^2(1-P)}{(1-\alpha B)^2}$	$\frac{-\alpha}{1-\alpha B}$	$\frac{\alpha(1-P)}{1-\alpha B}$
FOCN	+	$d/(c+d)$	$\frac{1-\alpha B-\alpha+\alpha P}{1-\alpha B}$	$\frac{\alpha^2(P-1)}{(1-\alpha B)^2}$	$\frac{\alpha}{1-\alpha B}$	$\frac{\alpha(1-P)}{1-\alpha B}$
TS	+	$\frac{a}{a+b+c}$	$\frac{P}{B+1-P}$	$\frac{-P}{(B+1-P)^2}$	$\frac{B+1}{(B+1-P)^2}$	$\frac{P}{B+1}$
ETS	+	$\frac{a-a_r}{a+b+c-a_r}$, $a_r=(a+b)(a+c)$	$\frac{P-\alpha B}{B+1-P-\alpha B}$	$\frac{P(2\alpha-1)-\alpha}{(B+1-P-\alpha B)^2}$	$\frac{B+1-2\alpha B}{(B+1-P-\alpha B)^2}$	$\frac{P+\alpha-2\alpha P}{B+1-2\alpha B}$
CSS	+	$\frac{ad-bc}{(a+b)(c+d)}$	$\frac{P-\alpha B}{B(1-\alpha B)}$	$\frac{2\alpha PB-P-\alpha^2B^2}{B^2(1-\alpha B)^2}$	$\frac{1}{B(1-\alpha B)}$	$\frac{P+\alpha^2B^2-2\alpha PB}{B(1-\alpha B)}$

subscript denotes the removal of yes forecasts. If the FOH is less than the CPR for a performance measure, it is likely that a random decrease in yes forecasts will result in an indication of improvement by that measure. Therefore, the FOH may be considered an upper bounding benchmark for the CPR: the larger the FOH minus CPR difference for a performance measure, the greater the likely difficulty of successful hedging by decreasing the bias.

In the preceding discussion, the FOH is presented as an upper bounding value for the CPR, while the DFR is described as a lower bounding value for the CPR. In general, it is expected that FOH > DFR, and the CSS is positive. Under what condition is the opposite true, giving a negative CSS? The requirement is

$$\frac{P}{B} < \frac{\alpha(1-P)}{1-\alpha B}. \tag{4}$$

Since $1-\alpha B > 0$, this reduces to the following inequality:

$$\alpha P < \alpha^2 B. \tag{5}$$

From Table 1, αP is a, the frequency of hits on the verification domain ("domain hit frequency" hereafter). The probability of a chance hit is the product of the forecast frequency and the observed frequency, which, with reference to Table 1, is $\alpha^2 B$. Therefore, by Eq. (5), forecasts having a domain hit frequency less than that associated with a random forecast will have CSS negative and FOH < DFR.

3. Application

This section applies the results of the previous section to the TS, ETS, and CSS by deriving what conditions characterize a forecast that could be improved by randomly increasing or decreasing yes forecasts. For our purposes here, random change is considered to be the extreme limiting case for hedging forecasts. If the bias is increased by randomly changing no to yes forecasts, then the score improves if

$$\mathrm{DFR} = \frac{\Delta h_A}{\Delta f_A} = \frac{\alpha(1-P)}{1-\alpha B} > \mathrm{CPR}. \tag{6}$$

On the other hand, if the bias is decreased by randomly changing yes forecasts to no forecasts, then the score improves if

$$\mathrm{FOH} = \frac{\Delta h_R}{\Delta f_R} = \frac{P}{B} < \mathrm{CPR}. \tag{7}$$

For the TS to indicate improvement for a random increase in yes forecasts, applying Eq. (6) using the CPR formula for TS from Table 2 yields

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

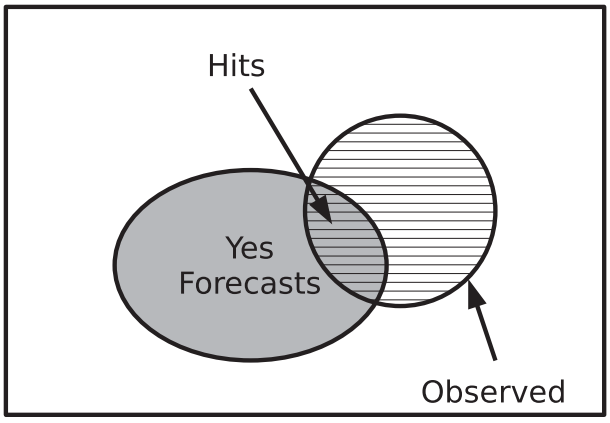

FIG. 1. Schematic showing yes forecasts (shaded), yes observations (hatched), hits (hatched and shaded), and no forecasts (all unshaded including hatched unshaded).

$$\frac{\alpha(1-P)}{1-\alpha B} > \frac{P}{B+1}, \tag{8}$$

which, upon simplification (since $1 - \alpha B > 0$), gives

$$\alpha P < \alpha^2 B + \alpha^2(1-P). \tag{9}$$

Therefore, Eq. (9) indicates that the TS can improve by randomly increasing yes forecasts even if the original forecast has a domain hit frequency exceeding the random hit frequency, but the domain hit frequency must not exceed the random frequency plus the second term on the rhs of Eq. (9), which could be small owing to the squared event frequency factor. For common events [$\alpha \sim 0.5$; see Baldwin and Kain (2006)], hedging the TS is more likely to succeed. On the other hand, if yes forecasts are randomly decreased, the condition for the threat score to improve is given by substitution of the CPR for TS from Table 2 into Eq. (7): $P/B < P/(B + 1)$, requiring $P < 0$, which is not possible. The TS is not likely to be improved by randomly changing yes forecasts to no forecasts.

Similar derivations for ETS and CSS using Eqs. (6) and (7) with CPR formulas from Table 2 [provided $\alpha < 0.5$ in the case of ETS, rarely a serious constraint; see Baldwin and Kain (2006) and Brill (2009)] result in the same conditions for either the random decrease or increase of yes forecasts for both scores:

$$\alpha P < \alpha^2 B. \tag{10}$$

The only likely way for ETS or CSS to indicate improvement for any random changes to the number or area of yes forecasts is for the original forecast to have a domain hit frequency less than the random hit frequency (both ETS and CSS less than zero). Thus, given forecasts having better than random skill, successful hedging to improve ETS or CSS itself requires some ''skill,'' although the practice of doing so usually is discouraged. Indeed, hedging is rarely done randomly, because the yes and no forecast points or areas changed are not randomly chosen; rather, they are selected based on the temporal or spatial proximity to points or areas existing in the original forecast. It is also important to note that Eq. (10) is not uniquely applicable to the ETS or CSS; it is also true for some other performance metrics (e.g., Peirce skill score and Heidke skill score). Derivations applying Eqs. (6) and (7) do not yield simple results using the CPR formulas for performance measures such as the odds ratio skill score [CPR formula derived by Brill (2009)] and the Mesinger (2008) bias-adjusted scores [CPR formulas derived by Brill and Mesinger (2009)]. In such cases, numerical or graphical methods may be applied to assess CPR behavior relative to the benchmarks. Treatment of these performance measures is reserved for future work and is beyond the scope of this note.

For a CSS that is appreciably greater than zero, the CPR benchmarks are far apart and not of much consequence; therefore, poor forecasts without regard to event frequency or misplaced yes forecasts of rare or highly localized events are of interest in revealing the behavior of the CPR relative to the benchmarks. High-resolution models often produce stochastically reasonable distributions of heavy precipitation, but fail to achieve proper placement of the accumulation areas relative to verifying observations. In fact, such placement errors have motivated other verification treatments known as spatial techniques (e.g., Gilleland et al. 2009; Mesinger 2008; Davis et al. 2006, among others). Here, verification of 36-h forecasts of precipitation exceeding the threshold of 3.0 in. (~76 mm) in a 24-h period from a mesoscale NCEP operational model is utilized to illustrate the previously discussed concepts.

The NCEP High-Resolution (Hires) Window model forecasts examined here are based on the Weather Research and Forecasting (WRF) model system's Nonhydrostatic Mesoscale Model (WRF-NMM; e.g., Janjić et al. 2005), utilizing a locally modified variant of WRF version 2.2 code. The horizontal grid spacing is 4 km, with 35 levels in the vertical and a model top pressure of 50 hPa. The model initial and lateral boundary conditions are generated by interpolation of NCEP's North American Mesoscale Model (NAM) output onto an integration domain covering the eastern two-thirds of the continental United States for the cases considered here. The physics options match those used by NAM (e.g., Rogers et al. 2005) with the exception that no

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

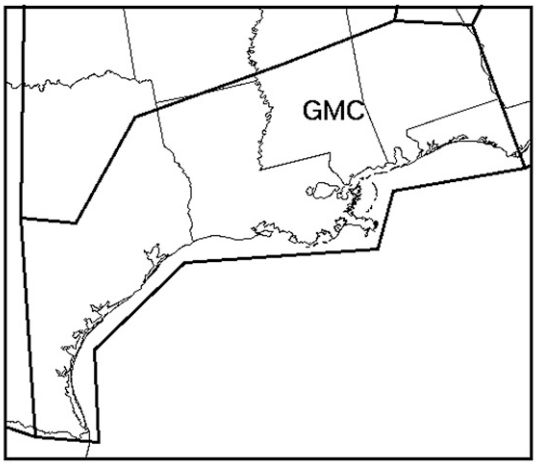

FIG. 2. Map showing the Gulf of Mexico coastal verification region of the continental United States.

parameterized convection is utilized for the Hires Window run.

Verification is performed against the NCEP/Climate Prediction Center's (CPC) gauge-only precipitation analysis (e.g., Shi et al. 2003), which provides a quality-controlled 24-h (1200 to 1200 UTC) total. Both the model forecast and the verification dataset are remapped onto a common 12-km grid to generate the statistics populating the elements in Table 1.

The verification selected for this demonstration is cumulative over the area labeled GMC in Fig. 2. Four valid dates (cases, valid at 1200 UTC) for the 36-h forecasts of 24-h precipitation accumulation are selected on the basis of ETS values and ordered (nonchronologically) from highest to lowest ETS for QPFs exceeding 3.0 in. The first two cases exhibit positive ETSs and are better than a random forecast, while the last two have negative ETS values and are worse than a random forecast. The first one of the last two of these forecasts is selected on the basis of both nonzero POD and negative ETS, which, like negative CSS, indicates performance worse than a random forecast. The worst-case forecast has zero POD. Figures 3a–c show outcomes using the TS, ETS, and CSS, respectively, in terms of actual scores, CPR values, and the benchmarks DFR and FOH as a function of the cases laid out along the abscissa in the order described above. The fifth point on the abscissa shows results of cumulative verification for the 1-month period from 22 March through 22 April 2009, the interval of time from which the individual cases are selected. Observed and forecast frequencies along with frequency bias for these cases and the cumulative verification are given in Table 3, revealing that the forecast frequency exceeds that observed for all but the fourth case, which is the worst forecast, having POD = 0. The cumulative verification also exhibits a high bias.

Figure 3a shows that the TS exceeds zero for all but the worst forecast, indicating some "skill" because there are some hits in each of the first three cases. The first case, a forecast valid at 1200 UTC 26 March 2009, shows the TS CPR value between the two benchmarks for a forecast that is better than random, indicating little likelihood of improving the TS by randomly increasing or decreasing yes forecasts. This also is true of the cumulative verification results shown by the last point on the abscissa. The second forecast valid at 1200 UTC 19 April 2009 is almost a random forecast, but is slightly better than random as shown by small positive values of ETS and CSS in Figs. 3b and 3c. This second case illustrates the behavior of the TS with respect to Eq. (9), because in Fig. 3a the DFR exceeds the CPR for the TS, indicating that the TS for this forecast, which has slightly better than random performance, is likely to be improved by a random increase in yes forecasts. The DFR exceeds the CPR for TS for the last two individual cases along the abscissa in Fig. 3a, indicating that the TS for these worse-than-random forecasts would likely be improved by a random increase in yes forecasts. On the other hand, the FOH benchmark for the decreasing bias never falls below the CPR in Fig. 3a, demonstrating the unlikelihood of improving the TS by randomly decreasing yes forecasts as shown mathematically above.

Figures 3b and 3c demonstrate the conclusion drawn from Eq. (10), which applies to both ETS and CSS. The CPR value is expected to fall between the two benchmark values for the better-than-random forecasts composing the first two cases. The ETS and CSS CPR values also lie between the two benchmark values after the benchmark values switch ordinate positions relative to each other for the two individual cases that exhibit worse-than-random performance. This position of the CPR value in the last two individual cases indicates that any random change to the number of yes forecasts is likely to result in an improved CSS or ETS. The CSS CPR lies nearly on the FOH value because the event frequencies are very low. From Table 2, it is clear that for very small event frequency values the CSS CPR can be nearly equal to the FOH value. Therefore, the CSS is less likely than ETS to show improvement by randomly decreasing yes forecasts of rare events when the forecast itself has less skill than a random forecast.

Considering the cumulative verification, the positive ETS value in Fig. 3b (last point on the abscissa) indicates forecasts having skill over random performance with the

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

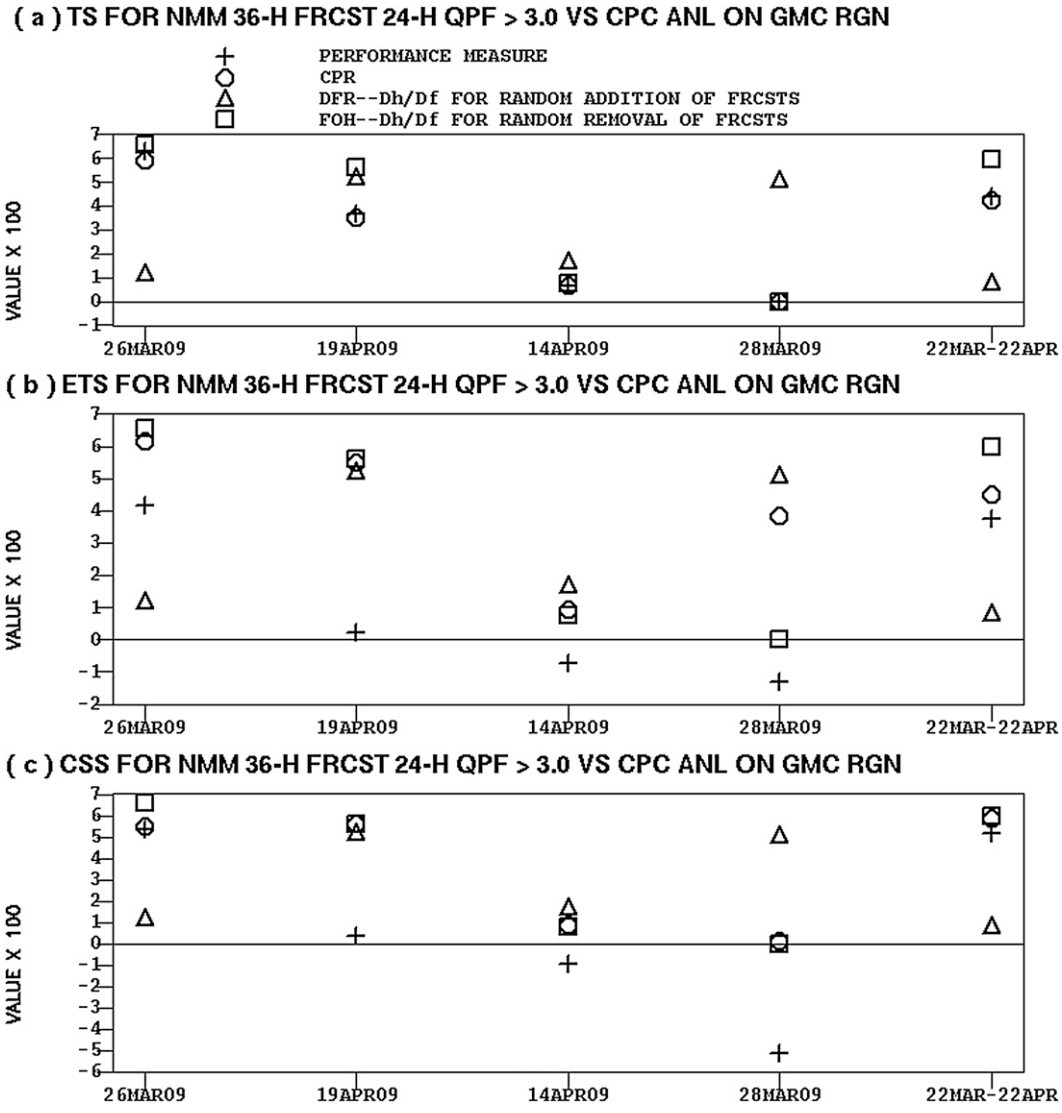

FIG. 3. (a) TS, (b) ETS, and (c) CSS (+ signs) for four selected dates (valid at 1200 UTC, displayed along the abscissa) and the accumulated 1-month verification (last point on the abscissa) from the operational high-resolution NMM verification record plotted against the ordinate on the left. The octagonal symbols plot the corresponding CPR values for each case. The DFR and FOH benchmarks are plotted as triangles and squares, respectively. All values plotted against the ordinate are scaled by 100.

CPR value falling well within the benchmarks, indicating little chance of improving the ETS by random changes in the number of yes forecasts. The positive CSS value for the cumulative verification in Fig. 3c also indicates better-than-random performance, but the CPR value for the CSS is only slightly less than the FOH value. This indicates that the CSS is close to being susceptible to improvement by random change of yes forecasts to no forecasts. The last column Table 3 indicates a substantial bias for the cumulative verification. The CPR analysis suggests that a nearly random bias correction is likely to improve the CSS for the cumulative verification.

4. Concluding remarks

This work derives the critical performance ratio (CPR) expressions for the eight conditional probabilities associated with the 2×2 contingency table for dichotomous forecasts. For two of these conditional probabilities, the DFR and FOH, the CPR formula is identical to the formula for the conditional probability itself. These two represent the probability of adding a hit if yes forecasts are increased randomly in the case of DFR and the probability of removing a hit if yes forecasts are decreased randomly in the case of FOH. Ideally, any performance

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC

TABLE 3. Observed and forecast event frequencies giving the fraction of the verification domain covered by 24-h accumulations exceeding the 3.0-in. threshold followed by the frequency bias in the last row. The second–sixth columns hold values for the four cases and cumulative verification in the same order as along the abscissa in Fig. 3. All valid dates are in 2009.

Cases	26 Mar	19 Apr	14 Apr	28 Mar	22 Mar–22 Apr
Observed	0.0233	0.0528	0.0164	0.0502	0.0095
Forecast	0.2084	0.0890	0.0748	0.0177	0.0229
Bias	8.944	1.686	4.561	0.3526	2.410

measure applied to a forecast that is better than a random forecast must have DFR $<$ CPR $<$ FOH so that random changes to the number of yes forecasts are unlikely to cause the performance measure to indicate an improved forecast. This is certainly true of ETS and CSS as demonstrated herein. Hedging in general may be discouraged by choosing a performance measure whose CPR value typically falls well within the DFR and FOH benchmarks, optimally, midway between them.

Considering *only* random changes for forecasts that are worse than random (DFR exceeds FOH), performance measures may indicate improvement by only an increase or only a decrease in yes forecasts, or either one of these two changes, depending on the performance measure. To assess the performance of forecasts for rare events using performance measurements associated with a 2×2 contingency table, it is preferable to choose a performance measure that readily indicates worse-than-random performance. The ETS and CSS satisfy this requirement by taking on negative values for forecasts that are worse than random, but the TS cannot indicate worse-than-random performance. Although the TS has a long history of use (Baldwin and Kain 2006) and must be continued in many cases to demonstrate the long-term trend in forecast performance, this work suggests the TS should be abandoned as a decision-making metric in comparative evaluations of contemporary forecast systems. For verification situations in which it is not possible to populate the not-forecast–not-observed cell of the 2×2 contingency table, the TS continues to be a useful performance metric. Either the ETS or CSS may be used as a reasonable replacement for the TS, although doing so does not eliminate bias sensitivity. In all cases, the frequency bias must necessarily accompany any measurement of performance.

Acknowledgments. Funding from NCEP/HPC to provide computational tools and to cover the publication costs is much appreciated. NCEP/EMC is also gratefully acknowledged for contributing to the publication costs. The authors thank the anonymous reviewers whose suggestions for revisions substantially improved the comprehensibility of this note.

REFERENCES

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting,* **21,** 636–648.

Brill, K. F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting,* **24,** 307–318.

——, and E. Mesinger, 2009: Applying a general analytic method for assessing bias sensitivity to bias-adjusted threat and equitable threat scores. *Wea. Forecasting,* **24,** 1748–1754.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.,* **134,** 1772–1784.

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.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting,* **24,** 1416–1430.

Janjić, Z., T. Black, M. Pyle, E. Rogers, H.-Y. Chuang, and G. DiMego, 2005: High resolution applications of the WRF NMM. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 16A.4. [Available online at http://ams.confex.com/ams/pdfpapers/93724.pdf.]

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification. A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 240 pp.

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

Mesinger, F., 2008: Bias adjusted precipitation threat scores. *Adv. Geosci.,* **16,** 137–142.

Rogers, E., and Coauthors, 2005: The NCEP North American Mesoscale modeling system: Final Eta Model/analysis changes and preliminary experiments using the WRF-NMM. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 4B.5. [Available online at http://ams.confex.com/ams/pdfpapers/94707.pdf.]

Shi, W., E. Yarosh, R. W. Higgins, and R. Joyce, 2003: Processing daily rain-gauge precipitation data for the Americas at the NOAA Climate Prediction Center. Preprints, *19th Conf. on Interactive Information Processing Systems,* Long Beach, CA, Amer. Meteor. Soc., P1.6. [Available online at http://ams.confex.com/ams/pdfpapers/56719.pdf.]

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

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 630 pp.

Unauthenticated | Downloaded 01/31/23 10:50 PM UTC