An Information-Theoretical Score of Dichotomous Precipitation Forecast

MAJID FEKRI

Department of Atmospheric and Oceanic Sciences, McGill University, Montréal, Québec, and Meteorological Research and Development, Pelmorex Media Inc., Oakville, Ontario, Canada

M. K. YAU

Department of Atmospheric and Oceanic Sciences, McGill University, Montréal, Québec, Canada

(Manuscript received 10 June 2015, in final form 6 January 2016)

ABSTRACT

This study presents an information-theoretical score (ITS) with an emphasis on desirable and undesirable mutual information between a series of dichotomous forecast and observation. As ITS makes use of the same contingency table as traditional scores, the performance of threat score (TS), equitable threat score (ETS), and true skill statistics (TSS) are compared with ITS using three different approaches. First, a hypothetical forecast setup is employed to investigate the responses of the scores to bias, phase error, and event frequency. It was found that the desirable mutual information portion of ITS \((C_1)\) is closer to TSS, and the undesirable mutual information portion of ITS \((C_2)\) reveals the presence of biases and random errors in the forecast. There is also a similarity between ITS and ETS. Second, the sensitivities of ITS and ETS to forecast bias tendency are examined analytically using the critical performance ratio (CPR). It is shown that ITS has a more dynamical response to incremental bias. By increasing the bias, the CPR value of ITS increases more rapidly than that of ETS indicating a higher resistance to hedging. Third, the skill scores on two sets of operational forecasts are applied with respect to a mosaic of observed radar reflectivity. The results show that ITS remains more consistent in its evaluation of skills at different thresholds compared to other scores.

1. Introduction

Several forecast skill scores are being used in the verification of precipitation forecasts, and each of them has its pros and cons. There has been extensive research to understand and compare the behavior of these scores and to identify their sensitivity to different errors (Hamill 1999; Accadia et al. 2003; Baldwin and Kain 2006; Ahijevych et al. 2009; Gilleland et al. 2009; Mittermaier and Roberts 2010). A number of recurrent questions exist about the true skill of precipitation forecasts, including the sensitivity of the scores to bias, location error, and event frequency.

Bias can skew scores. As a result it is possible for a wet-biased forecast with a larger bias to score better than a slightly more skillful but unbiased forecast. This higher score is believed to be a consequence of the inflated coverage of the precipitation and not an indication of actual forecast skill. To take into account the correct placement of precipitation, Mesinger (2008) and Brill and Mesinger (2009, hereafter BM09) introduced the bias-adjusted threat score.

The location or displacement error affects significantly the accuracy of the forecast (Casati 2010; Jolliffe and Stephenson 2012). Since precipitation patterns entail complex geometrical shapes and they evolve over a vast range of temporal and spatial scales, it is difficult to pinpoint a specific source of error. However, one can investigate how each skill score penalizes false alarms and misses due to the displacement of the forecast fields (Gneiting and Raftery 2007; Gilleland et al. 2009).

The “event frequency” or “base rate” is defined as the average coverage of observed rain events over the verification domain. Gandin and Murphy (1992) first introduced the notion of “equitable” score, which is independent of the domain size or event frequency. The score yields the same expected value for all random and constant forecasts and does not depend on the relative frequency of events. At the beginning it was thought that equitable scores should consist of a linear combination of the elements of
the contingency table. However, Hogan et al. (2010) showed that such restriction is unnecessary. The main concept in regard to event frequency lies in the consistency of forecast skill measures when dealing with rare events and common events.

Stephenson et al. (2008) introduced the extreme dependency score (EDS), a logarithmic measure designed to maintain a meaningful limit for rare events. The main disadvantages of EDS are its strong dependence on event frequency (Ghelli and Primo 2009) and its susceptibility to hedging by bias. Hogan et al. (2009) introduced symmetric EDS (SEDS), which is harder to hedge but still depends on event frequency and does not have a fixed range. A more advanced score in the same category is the symmetric extremal dependence index (SEDI), which has improved characteristics (Ferro and Stephenson 2011). We will discuss and compare some properties of the SEDI with the ITS score in section 6.

Parallel to these efforts, information-theoretical approaches proposed logarithmic measures to evaluate the skill of operational forecasts (Roulston and Smith 2002). The information-theoretical analysis of the forecasts provides a measure of the entropy of forecast compared to the entropy of the observation. In such approaches, the mutual information between forecast and observation is regarded as a measure of forecast skill (Ahrens and Walser 2008). Random correlations happen due to chance, but mutual information indicates the fraction of systematic similarity and dissimilarity between two random sets. Further support of the logarithmic scores for forecast evaluation was provided by Benedetti (2010). Nevertheless, previous research did not consider the conventional scores and the information-based scores alongside each other. The reason is that in conventional scores, the goal is to identify objective similarities between forecast and observation, while in information-based scores the mutual information includes every organized similarity that can be desirable or undesirable.

The information-theoretical score (ITS) that we investigate in this paper is a measure of the joint information between forecast and observation with an acute distinction between desirable and undesirable mutual information. Our objective is to study the conventional skill scores and ITS in the same context by introducing two skill indexes \( C^+ \) and \( C^- \). These two measures divide the mutual information into two parts. The \( C^+ \) index indicates desirable similarity and \( C^- \) represents the undesirable similarity (or dissimilarity) between forecast and observation. A comparison between ITS and some other well-known scores in a hypothetical test and an analytical approach reveals their sensitivities to bias, displacement, and event frequency. Additionally, the same scores are applied to two sets of operational precipitation forecasts and a set of radar observations as an example of real forecast verification.

The organization of this paper is as follows. Section 2 contains a brief description of the theory, basic assumptions, and governing equations. In section 3, a hypothetical forecast situation adopted from Baldwin and Kain (2006) is used to investigate the sensitivity of ITS to bias, displacement, and event frequency. We compare our results to those of the threat score (TS, also known as critical success index), the equitable threat score (ETS), and the true skill statistics (TSS). In section 4, the analytical method of Brill (2009) and BM09 is applied to compare the ITS with the ETS. We obtained an analytical formulation of critical performance ratio (CPR) for the ITS and examined its performance at various event frequencies and biases. In section 5, the different scores are applied to NWP forecasts and radar rainfall estimates to provide an example of real forecast evaluations. Finally, a discussion and conclusion is given in section 6.

2. Formulation of scores

Consider a dichotomous forecast consisting of rain and no-rain events. Let us define the rain events as the accumulation of precipitation over a desired threshold during a certain amount of time. Consider also a dichotomous observation that monitors rain events over the same area and the same period. The binary nature of these dichotomous forecasts and observations allows us to regard them as two random series of zeroes and ones. To compare these two random series, we need an evaluation method to measure and to quantify their similarity.

We consider information as the logarithm of the inverse probability of an event. In general, the probability of a random event is hard to define and to calculate. However, when large samples of events are available, we can use the frequency of occurrence of an event as its probability (Watanabe 1969). As such, in a dichotomous spatial forecast over a large domain, the number of rainy grid points over the total number of grid points in space is defined as the probability of rain \( p_r \) for that particular forecast. A similar argument can be applied to a no-rain event with the probability \( 1 - p_r \). As the total information of a closed set of events is also called the entropy function and is calculated by averaging the information of all possible events, the entropy function \( S(\mathbf{F}) \) for any set of binary forecast data \( \mathbf{F} \) can be written as

\[
S(\mathbf{F}) = -p_r \log(p_r) - (1 - p_r) \log(1 - p_r). \tag{1}
\]
To be consistent, the logarithm in (1) is the logarithm with base 2 and the information is given in units of bytes. 

Now, consider two binary vectors \( \mathbf{O} \) and \( \mathbf{F} \) with length \( n \) representing observation and forecast at \( n \) grid points in space or \( n \) steps in time. Let a new set \( \mathbf{G} \) be defined as the conjunction of \( \mathbf{O} \) and \( \mathbf{F} \). The total entropy of \( \mathbf{G} \) must be less than or equal to the sum of the entropies of \( \mathbf{O} \) and \( \mathbf{F} \). The equality holds if and only if the binary sets \( \mathbf{O} \) and \( \mathbf{F} \) are totally random and statistically independent. Therefore, the difference between the entropy of \( \mathbf{G} \) and the sum of the entropies of \( \mathbf{O} \) and \( \mathbf{F} \) can be regarded as the interdependence of \( \mathbf{O} \) and \( \mathbf{F} \) (Watanabe 1969). This interdependence is also called the mutual information or joint information \( J \), which is always greater than or equal to zero:

\[
J(\mathbf{O}, \mathbf{F}) = S(\mathbf{O}) + S(\mathbf{F}) - S(\mathbf{G}). \tag{2}
\]

The traditional approach in forecast evaluation consists of using a \( 2 \times 2 \) contingency table (Table 1) that lists the number of grid points with correct hits \( a \), false alarms \( b \), misses \( c \), and correct negatives \( d \). These variables correspond directly to the four probabilities of the four possible events in \( \mathbf{G} \). Hence, the entries \( a, b, c, \) and \( d \) can be used to calculate the joint information between forecast and observation. The entropy of the observation \( S(\mathbf{O}) \), the entropy of the forecast \( S(\mathbf{F}) \), and the entropy of \( \mathbf{G} \) can be written as

\[
S(\mathbf{O}) = -(a + c) \log(a + c) - (b + d) \log(b + d), \tag{3}
\]

\[
S(\mathbf{F}) = -(a + b) \log(a + b) - (c + d) \log(c + d), \tag{4}
\]

\[
S(\mathbf{G}) = -a \log a - b \log b - c \log c - d \log d. \tag{5}
\]

Two useful quantities that represent desirable mutual information and undesirable part were suggested by Watanabe (1969) as measures of direct similarity \( C^+ \) and dissimilarity \( C^- \) of binary sets and can be written as

\[
C^+ = a \log \left( \frac{a}{a + b} \right) + d \log \left( \frac{d}{d + b} \right), \tag{6}
\]

\[
C^- = b \log \left( \frac{b}{(b + a)(b + d)} \right) + c \log \left( \frac{c}{(c + d)(c + a)} \right). \tag{7}
\]

where the logarithm is the logarithm with base 2, and \( C^+ \) and \( C^- \) are in units of byte. In case of perfect similarity between forecast and observation \( a + d = 1 \) and \( b, c = 0 \), \( C^+ \) then becomes maximum and \( C^- = 0 \) (0 log 0 = 0). On the other hand, when forecast and observation are exact opposites of each other \( b + c = 1 \) and \( a, d = 0 \), \( C^+ = 0 \), and \( C^- \) then becomes a maximum. Although \( C^+ \) and \( C^- \) can be positive or negative numbers, the sum of them is always a nonnegative value because it is equal to the total joint information between forecast and observation. By adding (6) and (7), the joint information between forecast and observation is obtained in (2), thus

\[
J(\mathbf{O}, \mathbf{F}) = S(\mathbf{O}) + S(\mathbf{F}) - S(\mathbf{G}) = C^+ + C^- \tag{8}
\]

As the maximum possible value of joint information is always less than or equal to the entropy of observation, it is convenient to normalize the joint information score using the entropy of observation with the following form:

\[
0 \leq J(\mathbf{O}, \mathbf{F}) \leq S(\mathbf{O}), \tag{9}
\]

\[
\text{ITS} = \frac{J(\mathbf{O}, \mathbf{F})}{S(\mathbf{O})} \leq 1. \tag{10}
\]

In summary, we emphasize that ITS is a dimensionless, nonnegative value equal to or smaller than 1 and it represents the ratio of the mutual information to the observation information. ITS relies on the same traditional contingency table, and false alarms and misses equally affect the total mutual information score of the forecast. The definitions and the mathematical formulas for the scores are listed in Table 2. The \( C^+ \) and \( C^- \) indexes divide mutual information into similarity and dissimilarity measures. A positive \( C^+ \) indicates a systematic similarity between prediction and observation that contributes positively to the total score. A negative \( C^- \) represents a situation where the similarities are random, and dissimilarities are prominent. Note that \( C^+ = 0 \) is the threshold below which the forecast has no desired skill. A positive \( C^- \) shows a systematic dissimilarity between forecast and observation that contributes positively to the score but must come from a serious error in the forecast. A negative \( C^- \) negates some of the total joint information and indicates the magnitude of the random errors. In the next section, the sensitivity of ITS to bias and displacement error is studied and discussed.

### 3. Hypothetical test

This test is a modified version of the hypothetical forecast situation proposed by Baldwin and Kain (2006, hereafter BK06) designed to illustrate the properties of the scores. BK06 used two circles inside a reference square to represent rain events and simulate bias and displacement errors. Here, we represent forecast and
observed rain events as two squares inside a larger reference square. Since the main idea is to approximate rain events by simple geometrical forms, we found that circles or squares both produce the desired results. The advantage of using squares over circles is that the calculation of overlapped areas becomes much easier. Figure 1 shows four possible hypothetical scenarios. The reference square depicts the whole domain, and smaller squares represent the rain events from observation and forecast, respectively. The reference square is assumed to be a unit square. The observation and forecast squares can change size and can be displaced in space. For each scenario, the elements of the $2 \times 2$ contingency table (Table 1) are calculated analytically as functions of the distance $D$ between the centers of the two squares with sizes $d_1$ and $d_2$. We fix the size of observed rain event at a certain event frequency $\alpha = d_1^2$ and change the relative size of the forecast coverage $d_2^2$ as a measure of the forecast bias $B = (d_2/d_1)^2$. Similar to BK06, we consider three cases of different rain event frequencies: a rare event with $\alpha = 0.04$, a common event with $\alpha = 0.16$, and a very common event with $\alpha = 0.36$. For each case, the $2 \times 2$ contingency table is calculated as a function of displacement $D$ and bias $B$ and then scores of forecast skills are computed. We mention that BK06 has discussed the results for the TS, ETS, and TSS scores. In this study, our focus is on the properties of ITS in comparison with the other traditional scores.

### a. Rare event

Figures 2a–d show the contour plots of TS, ETS, TSS, and ITS as functions of previously defined displacement $D$ and bias $B$ for a rare observed rain event ($\alpha = 0.04, d_1 = 0.2$). The best score in every case occurs at $B = 1$ and $D = 0$ where observation and forecast squares perfectly overlap. For a rare event, in the region where bias is around unity $0.5 \leq B \leq 1.5$, all the scores decrease rapidly by increasing the displacement $D$. Scores are more sensitive to $D$ in a rare event because a greater portion of area of correct hits $a$, false alarms $b$, and misses $c$ is affected by small displacements. As expected, at bias values $<1$ and at small displacements $(D < 0.1)$ where the forecast square is completely contained within the observed square, the contours become parallel to the $D$ axis. This region is at the bottom-left corner of the figures and corresponds to the scenario in Fig. 1d. Similarly, the region at the bottom right of the figures is where the larger forecast square encloses the whole observation, similar to the scenario in Fig. 1c. In this region, all of the scores become insensitive to $D$, but as $B$ increases the area of hits remains constant and the number of false alarms increases. At a very small displacement (e.g., $D = 0.05$), TS and ETS (Figs. 2a,b) decrease faster than ITS (Fig. 2d) when $B$ becomes larger than 1. However, TSS (Fig. 2c) becomes less sensitive to $B$ because it is the only score that does not penalize false alarms. At larger displacement ($0.05 \leq D \leq 0.25$) and larger biases ($1.0 \leq B$), the sensitivity of ITS and TSS to the bias becomes negligible, and the contours are slightly parallel to the $B$ axis.

The ITS can be better understood by examining the contributions of $C^+$ and $C^-$ (Figs. 2e,f). In this rare event, the normalized $C^+$ (Fig. 2e) is positive almost everywhere and behaves very similar to TSS. Here $C^-$ (Fig. 2f) is the difference between ITS and $C^+$ and it reveals how the ITS score is affected by random errors. Note that $C^-$ is zero at $B = 1$ and $D = 0$, but becomes increasingly negative with a small increase in displacement reaching a minimum at about $(D = 0.1)$.

---

**Table 2. Definition of various scores including ITS.**

<table>
<thead>
<tr>
<th>Score Type</th>
<th>Equation</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threat score (TS)</td>
<td>$TS = \frac{a}{a + b + c}$</td>
<td>$0 \leq TS \leq 1$</td>
</tr>
<tr>
<td>Equitable threat score (ETS)</td>
<td>$ETS = \frac{a - a_{rand}}{a + b + c - a_{rand}}$</td>
<td>$\frac{1}{5} \leq ETS \leq 1$</td>
</tr>
<tr>
<td>$a_{rand} = (a + b)(a + c)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True skill statistics (TSS)</td>
<td>$TSS = \frac{a}{a + c} - \frac{b}{b + d}$</td>
<td>$-1 \leq TSS \leq 1$</td>
</tr>
<tr>
<td>Information-theoretical score (ITS)</td>
<td>$ITS = \frac{C^+ + C^-}{S(O)}$</td>
<td>$0 \leq ITS \leq 1$</td>
</tr>
<tr>
<td>$C^+ = a\log \frac{a}{(a+b)(a+c)} + d\log \frac{d}{(d+b)(d+c)}$</td>
<td>$-1 \leq \frac{C^+}{S(O)} \leq 1$</td>
<td></td>
</tr>
<tr>
<td>$C^- = b\log \frac{b}{(b+a)(b+d)} + c\log \frac{c}{(c+d)(c+a)}$</td>
<td>$-1 \leq \frac{C^-}{S(O)} \leq 1$</td>
<td></td>
</tr>
<tr>
<td>$S(O) = -(a + c)\log(a + c) - (b + d)\log(b + d)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**b. Common event**

Figure 3 shows performance scores for a more common event ($\alpha = 0.16$) where the square of observation is twice as large as the rare event ($d_1 = 0.4$). In this case, by increasing the displacement, the scores decrease, but the rate of change shows less sensitivity to displacement compared to the rare event (Figs. 3a–d). The size of the observed square is larger than that of a rare event, and the overlap between the forecast and observed squares vanishes at relatively large displacements. The normalized $C^+$ (Fig. 3c) is again similar to the TSS (Fig. 3c). The dotted line in Fig. 3d shows the region of the minimum score where the ITS becomes very close to zero. As shown, the ITS maintains higher sensitivity to the displacement because the correlation at larger displacement increases the number of false alarms and misses and is penalized by $C^-$ as having a random dissimilarity. Here $C^-$ (Fig. 3f) has again a minimum at ($D \approx 0.1$) but with a larger magnitude than the rare event case.

We remark that at the upper-right quadrant of the figures, the observation square and the forecast square overlap again despite the large distance between them. At very large bias, the forecast square becomes large enough that it overlaps the observation square to generate some hits. The ETS and TSS (Figs. 3b,c) both become negative in this region. The TS (Fig. 3a) remains insensitive but ITS (Fig. 3d) becomes larger than zero. In this particular region, the fact that $C^-$ becomes positive while $C^+$ takes on...
negative values indicates that the information theoretical score is associated with systematic dissimilarity and not the usual systematic similarity.

c. Very common event

Figure 4 shows the forecast performance scores for a very common rain event with (\(\alpha = 0.36\)) and (\(d_1 = 0.6\)). In contrast to the two previous cases, the observation square is larger, and the probability of the rain event is higher than before. At very high bias values (\(B > 1.7\)), the sum of the areas of observation and forecast becomes more than the reference area. As a result, the score values are insensitive to bias values exceeding 1.7 and the contours become parallel to the \(B\) axis. The
overlap behavior of the forecast and observation at the upper-right quadrant when displacement and bias are large is similar to the situation of the common event.

Figures 4b, 4c, and 4e show that the ETS, TSS, and $C^+/S(O)$ plots are very similar particularly in the positive regions of TSS and $C^+/S(O)$. In comparison to the less common events (Figs. 3a and 2a), Fig. 4a shows that the spread of the TS contours is much larger along the $D$ axis with less sensitivity to displacement when the frequency of rain event $\alpha$ increases. Although scores like ETS, TSS, and $C^+/S(O)$ are sensitive to event frequency (cf. Figs. 2, 3, and 4b,c,e), the ITS score is the least affected (cf. Fig. 4d with Figs. 3d and 2d). This behavior can be anticipated because ITS is the mutual information between forecast and observation divided by the total entropy of the observation. For most practical event

---

**Fig. 3.** Various performance scores as a function of bias and displacement errors for a common event ($\alpha = 0.16$): (a) TS, (b) ETS, (c) TSS, (d) ITS, (e) $C^+/S(O)$, and (f) $C^-/S(O)$. 

frequencies ($\alpha < 0.5$), reducing (increasing) the event frequency corresponds to reducing (increasing) the entropy of observation, which occurs in the denominator of the ITS formula. Therefore, the value of ITS is essentially maintained at various event frequencies.

As an example, consider a more common rain event with $\alpha = 0.36$ and a less common event with $\alpha = 0.16$. Since the observational entropy $S(O)$ is larger in the former case, the denominator in the normalized ITS given by (10) increases and the ITS score remains relatively unchanged for the two events. However, the maximum possible entropy of observation $S(O)$ is one byte of information and it occurs at $\alpha = 0.5$ when rain and no-rain events are equally probable. At higher event frequencies,
\( S(O) \) decreases until it reaches zero at \( \alpha = 1 \). This is, however, a trivial case as it is raining everywhere. Another trivial case is the everywhere no-rain event with \( \alpha = 0 \). Again \( S(O) = 0 \) as there is no information in the observation. Excluding these two trivial cases, the equations for \( C^+, C^- \), and ITS \([6, 7, \text{and } 10]\) have no singularities in the nontrivial forecast and observation scenarios.

We note that the minimum of \( C^- \) in Figs. 2, 3, and 4 occurs at low displacement values \( (D \approx 0.1) \) regardless of event frequencies. Some researchers argue that the small phase error in the high-resolution NWP models contributes to the total amount of random forecast error. They evaluate model performance over varying spatial domains without penalizing near misses or near false alarms. Spatial forecast verification methods such as fractions skill score (FSS) serve that purpose (Mittermaier and Roberts 2010). Another method is to consider joint his-
tivation. Excluding these two trivial cases, the equations for

\[ f \equiv f(B, P). \]

According to BM09, the critical performance ratio CPR (denoted by \( \rho \)) is defined as the ratio of the partial derivative of scores:

\[ \rho = -\frac{\delta f}{\delta B}. \]

This ratio \( \rho \) indicates the tendency of increase or decrease in scores subject to small changes in bias or POD of forecasts. BM09 provided an analytical formulation of this quantity for various scores and examined their sensitivity to hedging. Here, we derive the CPR of ITS and investigate its specific properties relative to other scores.

Substituting \( B \) and \( P \) from (11) and (12) into the definition of ITS \([8–10]\), and taking partial derivatives leads to the formulation of ITS CPR. Since \( S(O) \) is only a function of \( \alpha \), its partial derivatives van-
ishes. Also \( S(F) \) can be expressed exclusively in terms of \( \alpha \) and \( B \), and its partial derivative with respect to \( P \) be-
comes nil. This modification simplifies the derivation of

\[ \rho_{(ITS)} = \frac{\partial S(F)}{\partial B} \times \frac{\partial S(G)}{\partial B}. \]

With some simplification, the final form of the ITS CPR in terms of \( B, P, \) and \( \alpha \) becomes

\[ \rho_{(ITS)} = \frac{\log\left\{ \frac{(B - P)(1 - \alpha B)}{B[1 - \alpha(B + 1 - P)]} \right\}}{\log\left\{ \frac{\alpha(B - P)(1 - P)}{P[1 - \alpha(B + 1 - P)]} \right\}}. \]

First, to confirm that the ITS CPR has real values, we show that the expressions inside the logarithms are always positive. According to the definition in (11) and (12), we expect that \( P \leq B \), therefore, \( B - P \geq 0 \). Moreover, \( [1 - \alpha(B + 1 - P)] = d \geq 0 \) is the correct negative element of the \( 2 \times 2 \) contingency table. Again, according to the definitions in (11) and (13), \( \alpha B = a + b \) is the event frequency in the forecast that is less than or equal to 1, thus \( 1 - \alpha B \geq 0 \). The special case of \( P = \alpha B \) leads to an ambiguous CPR value \( \rho_{(ITS)} = \log(1)/\log(1) \).
This ambiguity can be resolved by assuming \( P = x \alpha B \) and taking the limit of \( \rho_{\text{ITS}} \):

\[
\lim_{x \to 1} \rho_{\text{ITS}} = \alpha.
\]  

(18)

Therefore, ITS CPR has real values within the extent of the usual 2 \( \times \) 2 contingency table, but in the case that any of the variables \( B, P, \alpha, a, o, r \) are zero, it is not defined.

Second, to investigate the properties of ITS CPR, we consider three possible scenarios:

(i) \( P = \alpha B \): The numerator and denominator of (17) become both equal to zero but at the limit \( \rho_{\text{ITS}} = \alpha \).

(ii) \( P > \alpha B \): It is enough to show that both the numerator and denominator of (17) are positive, so \( \rho_{\text{ITS}} > 0 \).

(iii) \( P < \alpha B \): It is enough to show that both numerator and denominator of (17) are negative, so \( \rho_{\text{ITS}} > 0 \).

Therefore, the ITS CPR is always positive, or in other words, it is positively oriented. As explained in BM09, comparing CPR to the small changes of \( P \) relative to the small change in \( B \) specifies the condition for improvement of scores. The quantity \( \Delta P / \Delta B \) can be explained as the fraction of hit that is added or removed by changing a fraction of bias. Therefore, for a positively oriented score at a constant frequency \( \alpha \), when bias is increasing \( \Delta B > 0 \), improvement in score happens if the ratio of added hits beats CPR:

\[
\frac{\Delta P}{\Delta B} > \rho.
\]  

(19)

On the other hand, when bias is decreasing \( \Delta B < 0 \), the ratio of removed hits has to be smaller than CPR to allow for improvement of the score:

\[
\frac{\Delta P}{\Delta B} < \rho.
\]  

(20)

Figure 5 shows CPR contours of ITS as a function of \( B \) and \( P \) for three different event frequencies \( \alpha = 0.04 \) (Fig. 5a), \( \alpha = 0.16 \) (Fig. 5b), and \( \alpha = 0.32 \) (Fig. 5c). Bias values are plotted in a range from the proximity of zero to \( B = 4 \). The comparison between event frequencies shows that more common events are associated with higher ITS CPR values and sharper CPR gradients along the \( B \) and \( P \) axes. The blank area in the top left of these figures indicates the limit of \( P \leq B \) beyond which \( P \) is impossible to increase. Similarly, in the right bottom of Fig. 5c, the blank area corresponds to the limit \( 1 - aB \geq 0 \) beyond which the bias is undefined. The dotted vertical line indicates an unbiased situation where \( B = 1 \). As a general trend, ITS CPR values along this line increase by increasing the POD, which means that it becomes more difficult to hedge the score at higher POD by increasing bias values. A similar trend exists for other scores including ETS, and for some bias-adjusted ETS as in Fig. 1 of BM09. The similarity between bias-adjusted ETS CPR and ITS CPR is in their curved contours crossing the \( B = 1 \) line.

The slanting dash–dotted lines in Figs. 5a–c indicate \( P = aB \). The case of \( P = aB \) requires special treatment, because, it specifies the expected score of random forecasts. The area above (below) this line designates
forecasts that are performing more (less) skillfully than a random forecast. A point symmetry exists in ITS CPR values with respect to the midpoint of this line at $P = aB = 1/2$, which corresponds to the special case of $a = b = c = d = 1/4$, where the entropy of conjunction $S(G)$ is maximum. The symmetry in ITS CPR is a consequence of the fact that scoring less than a random forecast becomes as difficult as scoring higher than the random forecast. ITS is always zero (minimum) along the line $P = aB$, and starts to increase as forecast becomes farther from this line in either direction.

Figures 6a–c show contours of CPR for ITS and ETS as function of event frequency $\alpha$ and POD, at constant bias values $B = 0.9$ (Fig. 6a), $B = 1.0$ (Fig. 6b), and $B = 1.1$ (Fig. 6c). The dotted line again indicates the $P = aB$ criterion. The horizontal axis includes event frequencies in a range from the proximity of zero to $\alpha = 0.5$. In general, CPR values of ETS and ITS both increase as $P$ and $\alpha$ increase, meaning that it becomes more difficult to improve forecasts by increasing bias at higher scores and event frequencies.

The ETS CPR in Fig. 6b corresponds to $B = 1.0$, which is also available in Fig. 3a of Brill (2009). However, the ITS CPR contours are provided here for comparison. Along any line drawn outward from the origin of Fig. 6b, ITS CPRs are smaller and have smaller gradient than ETS CPR, except along the $P = aB$ line, where they become identical and equal to $\alpha$.

Figure 6a shows a dry-biased forecast with $B = 0.9$. In comparison with Fig. 6b that represents the unbiased forecast, both ETS and the ITS CPRs have larger values and greater gradients. However, the relative increase in the magnitude and gradient of the ITS CPR is more noticeable, especially at higher POD values. To visualize the differences we consider a line passing from the origin to the top-right corner of the figure. In Fig. 6a, the line will pass through larger ITS CPR contours compared to those of Fig. 6b. In a dry-biased forecast, improvement of score necessitates an increase in bias $\Delta B > 0$, therefore, the condition of (19) applies, which means that larger CPR values indicate a higher resistance of the score to hedging.

Figure 6c shows the case of a forecast that has a wet bias with $B = 1.1$. In this case, improvement in scores happens in the direction of $\Delta B < 0$, which corresponds to the condition of (20), where smaller CPR values indicate the resistance of score to hedging. Again consider a line passing from the origin to the top-right corner of the figure. This time, the line in Fig. 6c will cross the lower ITS CPR contours compared to those of Figs. 6b and 6a.

Comparison of Figs. 6a–c shows that in both ETS and ITS scores, CPR contours respond to changes in bias by

**Fig. 6.** Contours of CPR for ITS and ETS as function of event frequency $\alpha$ and POD, at constant bias values (a) $B = 0.9$, (b) $B = 1.0$, and (c) $B = 1.1$. The dotted line indicates the $P = aB$ criteria for a random forecast.
resisting against the direction of bias improvements. The ETS CPR contours change slightly in the space of $P$ and $a$, but the ITS CPR changes markedly and shows a dynamic resistance to hedging by bias.

5. Application in real forecast and observations

In this section, two sets of deterministic precipitation forecasts are compared to a set of radar observations. The first set of forecasts is the WRF-based model run by the Center for Analysis and Prediction of Storms (CAPS) for the 2008 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT). The model was run at a resolution of 4 km for 30 h starting at 0000 UTC of each day and continuing to 0600 UTC of the next day. There are 10 members of the ensemble, each differs in its boundary conditions, initial conditions, and microphysical schemes. Among the members of the ensemble forecast, we chose the forecast without assimilation of radar data. Further details about the forecast configurations are available in Xue et al. (2008).

The second set of data is a GEM-based forecast from the Canadian Meteorological Centre (CMC) operational forecasts at 15-km horizontal resolution and 58 variable vertical levels up to 10 hPa. The assimilation process for this set of forecast includes satellite radiance, but no radar reflectivity is used. Other characteristics of the model are presented in detail by Mailhot et al. (2006). Note that the names GEM and WRF and the comparisons in this paper refer to the specific set of forecasts as described in this section and do not represent these NWP models in a general sense.

The observation data are a multiradar mosaic of reflectivity observations covering a vast region of North America. The reflectivity mosaic of the National Severe Storms Laboratory (NSSL) was available at 1-km horizontal resolution for the constant altitude plane position indicator (CAPPI) maps at 2.5 km (Zhang et al. 2005). The 6-h rain rates were obtained by accumulating the instantaneous rain rates at every 15 min during the periods between 0000, 0600, 1200, and 1800 UTC.

The domain of this study is confined to the central and eastern continental United States from 103° to 78°W in longitude and between 32° and 45°N in latitude. Data were available during three months of spring 2008: from 16 April to 6 June 2008 including a total of 540 h of precipitation that included 90 cases of 6-h accumulations. We considered each 6-h accumulation as an independent set of rain event in space. The values of rain accumulation are in units of millimeters, and we remapped them to a common 16-km resolution over the domain of study by linear interpolation between the closest available grid points. A certain threshold defines the rain or no-rain event at every grid point in space and produces dichotomous forecasts and observations. Contingency tables are formed by comparing forecasts to observations over the 2D domain for different thresholds. Various skill scores including ITS, $C^+$, and $C^-$ indexes are calculated based on these contingency tables. Finally, the scores are averaged over the whole time series at different thresholds of rain accumulation. Figures 7a and 7b show various performance scores of two independent precipitation forecasts (WRF based and GEM based) compared to the radar observations averaged over 90 cases of 6-h accumulations. Figure 7 demonstrates the behavior of different scores as a function of event thresholds or precipitation intensities.

As can be seen in Figs. 7a and 7b, the forecast skill of both models decreases rapidly by increasing the event threshold over $10 \text{mm (6h)}^{-1}$. Although GEM has a lower skill at higher thresholds, it shows a slightly better skill score at the lowest threshold $0.2 \text{mm (6h)}^{-1}$. Meanwhile, $C^-$ shows more negative scores at lower thresholds indicating that there are more random errors associated with small-scale features. From our companion study, we know that GEM forecasts are less skewed and have a smaller bias, a smaller error variance, and a distribution that is closer to the radar observations (Fekri 2015). Therefore, the $C^-$ index shows that there are more random errors in GEM compared to WRF. These errors are not systematic bias or skewness of distributions, but they appear as random errors that go undetected with traditional contingency scores. The sign of a genuine skill advantage in ITS of a forecast is the increase in $C^+$ accompanied by either a more negative $C^-$ or a less negative $C^-$. The former corresponds to the improved ITS where both forecasts have low skills and large random errors. The latter would indicate a more remarkable improvement in skill and a possible change in the predictability of the forecasted event. It would require a forecast to be more accurate and less affected by random errors at the same time.

We performed the hypothesis test of Hamill (1999) to investigate the effect of ITS on the calculation of statistical significance. When comparing relatively similar forecast models such as our WRF-based and GEM-based forecasts, we do not expect differences to reach statistical significance levels. However, we assume a null hypothesis such as “GEM-based forecasts are better than WRF-based forecasts” to calculate and compare the statistical significance ranges of the ETS and ITS.

Figure 7c shows the lines of average ETS (red) and ITS (blue) scores of WRF (solid) and GEM (dashed) forecasts. Dotted lines above (below) the averaged WRF score indicate the level that GEM has to reach in order to prove (disprove) the null hypothesis with a
statistical significance. At each threshold, the resampling pool was obtained by creating 2000 random combinations of 6-hourly scores from the two models. The differences in scores were calculated between each pair to create a sampling distribution. This distribution indicates the number of random pairs that have certain score difference. A significance level $\alpha = 0.05$ was used to delineate the distribution at positive and negative ends. The positive (negative) intervals were added to the average WRF scores to indicate the upper (lower) boundary beyond which the difference in scores becomes statistically significant.

On average, the TSS has the highest absolute values in both WRF-based and GEM-based forecasts followed by $C^1$, TS, ETS, and ITS. In Figs. 7a and 7b, the ITS has the smallest value on average. However, the reduced magnitude of ITS scores reduces the temporal variance in the score differences and, therefore, increases the statistical confidence. For example, a comparison between the upper limits of statistical significance between the ETS and ITS scores at the lowest threshold [0.2 mm (6 h)$^{-1}$] in Fig. 7c shows that GEM ITS score needs a relatively smaller increase (from 0.17 to 0.20) to reach the statistical significance level than that of the GEM ETS (from 0.23 to 0.30). The former roughly corresponds to an 18\% increase in the relative average score magnitude while the latter requires a 30\% increase.

6. Discussion and conclusions

A fair score of precipitation forecast must remain consistent in various conditions, conform to the physical and statistical understanding of errors and skills based on tangible rules, and provide useful information about the forecast. In this study, a logarithmic score of dichotomous precipitation forecasts was presented and compared with other traditional scores. The basis for this score is the mutual information between observation and forecast with an emphasis on two possible fractions of information indicating systematic similarity and dissimilarities. We made comparisons with regards to bias, displacement, and event frequency and also studied the behavior of the scores under a hypothetical test and analytical sensitivity analysis.

An overall consideration of the results of the hypothetical test shows that the $C^+$ index of ITS has a comparable behavior with respect to other traditional scores, especially TSS. On the other hand, the $C^-$ index provides additional information about the inaccuracies of the forecast. For example, in case of a large phase error, although the hit rate increases by increasing the bias, the $C^-$ index indicates the unfavorable forecast skill. The accumulative contribution

![Fig. 7. Average scores of (a) WRF and (b) GEM forecasts based on various scores and (c) the statistical significance (dotted) of WRF (solid) vs GEM (dashed) according to ETS (red) and ITS (blue) scores.](http://journals.ametsoc.org/mwr/article-pdf/144/4/1633/4330832/mwr-d-15-0225_1.pdf)
of $C^+$ and $C^-$ results in an ITS score that in most cases of the hypothetical test showed resemblance with ETS. Although total ITS decreases as the error increases, the decrease of $C^+$ has a different interpretation than the negative contribution of $C^-$. 

One of the properties of traditional scores is their ambiguous physical meaning (Mesinger 2008). Usually, mathematical definitions and ad hoc changes are tailored to fit the expected functionality of each specific score without exact physical justifications. The ITS is different, in the sense that its formulation originates directly from information theory, and its meaning is understandable in terms of entropy. The baseline for null skill in ITS is the skill of a random forecast (Fig. 5). Furthermore, ITS is always in the expected range of $[0,1]$, and can be calculated from the same contingency table of other conventional scores. A comparison between the sensitivity of ETS and ITS to small changes in bias and hit rate indicated that both scores have a natural resistance to further change because of their increasing CPR. The results in Fig. 6 further demonstrated the existence of a higher resisting sensitivity for ITS against hedging by bias.

Another important issue is the effect of observed event frequency $\alpha$, because scores behave differently within the various range of event frequencies (Mason 1989; Baldwin and Kain 2006). In particular, when we consider forecast scores at multiple thresholds, the observed event frequency $\alpha$ becomes a function of the threshold to affect the consistency of the forecast evaluation (Mason 2008; Tartaglione 2010). In general, a comparison between Figs. 2d, 3d, and 4d show that the ITS is not as much affected by the observed event frequency compared to other scores. Applying different scores on operational forecasts also resulted in the same outcome about the resilience of ITS to event frequency. In Figs. 7a and 7b, the curve representing ITS score remains steady at lower thresholds more than any other score. The lower average magnitude of ITS scores contributes to their higher consistency and statistical confidence over a wide range of rain coverages.

Among the properties listed by Ferro and Stephenson (2011) for SEDI, there are two kinds of symmetries that are considered favorable in the context of score performances. First, the complement symmetry is the invariance of a score to reversing the event and nonevent in the contingency table. In complement symmetry, the elements $a$, $b$, $c$, and $d$ of the contingency table are replaced by $d$, $c$, $b$, and $a$, respectively. The ITS is complement symmetric by definition. Second, the transpose symmetry is the invariance of a score to changing the role of observation and forecast $(c,b)$. Again the $C^+$ and $C^-$ indexes of ITS are both transpose symmetric because false alarm and miss rates are treated equally by them. Evidently, by changing the role of forecast and observation, $S(O)$ and $S(F)$ would flip. We can still keep the transpose symmetry of ITS by normalizing the $C^+$ and $C^-$ indexes with respect to the forecast information $S(F)$ instead of observation $S(O)$. Aside from these symmetries, we showed that ITS has another important symmetry that no other score has, and that is the reflection symmetry with respect to the line $P = \alpha B$ that indicates the skill of the random forecast. On account of these advantages of the ITS, it is an attractive alternative measure for the evaluation of precipitation forecasts. While this study focused on precipitation, this measure would also be useful for other meteorological forecasts (e.g., for wind) that can be extreme and are often evaluated using contingency tables.

Acknowledgments. This study was mainly done as part of Majid Fekri’s Ph.D. thesis work at McGill University, supported by the NSERC/Hydro-Québec industrial research program. We gratefully thank the three anonymous reviewers for their comments and contributions. We also acknowledge the support and guidance of Luc Fillion.

REFERENCES


Fekri, M., 2015: Towards a better understanding of rain forecast error and skill. Ph.D. dissertation, Department of Atmospheric and Oceanic Sciences, McGill University, 239 pp.


