Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events

CHRISTOPHER A. T. FERRO AND DAVID B. STEPHENSON

National Centre for Atmospheric Science, University of Exeter, Exeter, United Kingdom

(Manuscript received 20 September 2010, in final form 28 January 2011)

ABSTRACT

Verifying forecasts of rare events is challenging, in part because traditional performance measures degenerate to trivial values as events become rarer. The extreme dependency score was proposed recently as a nondegenerating measure for the quality of deterministic forecasts of rare binary events. This measure has some undesirable properties, including being both easy to hedge and dependent on the base rate. A symmetric extreme dependency score was also proposed recently, but this too is dependent on the base rate. These two scores and their properties are reviewed and the meanings of several properties, such as base-rate dependence and complement symmetry that have caused confusion are clarified. Two modified versions of the extreme dependency score, the extremal dependence index, and the symmetric extremal dependence index, are then proposed and are shown to overcome all of its shortcomings. The new measures are nondegenerating, base-rate independent, asymptotically equitable, harder to hedge, and have regular isopleths that correspond to symmetric and asymmetric relative operating characteristic curves.

1. Introduction

Extreme weather events such as high wind speeds, heavy precipitation, or high temperatures can have severe impacts on society. Improving predictions of such events therefore has a high priority in national weather services, and an important part of this activity is to determine whether or not prediction quality is improved when prediction systems are updated. Assessing the quality of predictions of extreme weather events, however, is complicated by the fact that measures of forecast quality typically degenerate to trivial values as the rarity of the predicted event increases. The drive to improve predictions of extreme events and the associated difficulties of measuring the quality of such predictions has generated a growing interest in better ways of verifying forecasts of extreme events.

In this paper we consider the problem of verifying deterministic forecasts of rare binary events. Forecasts that state whether or not daily rainfall accumulations will exceed a high threshold provide one example. A set of such forecasts is commonly displayed in a 2×2 contingency table, such as Table 1.

Many summary statistics of contingency tables have been proposed as measures of forecast performance (Mason 2003). Popular examples include the hit rate,

$$H = \frac{a}{a+c};$$

the false-alarm rate,

$$F = \frac{b}{b+d};$$

and the odds ratio,

$$\mathrm{OR} = \frac{ad}{bc}.$$

We can illustrate the difficulty of verifying forecasts of extreme events with a set of precipitation forecasts considered previously by Stephenson et al. (2008). The forecasts are 6-h rainfall accumulations taken directly from the old 12-km mesoscale version of the Met Office Unified Model (Davies et al. 2005) at the grid point nearest to Eskdalemuir in Scotland between 1 January 1998 and 31 December 2003. The observations are 6266 corresponding rain gauge measurements from the

Corresponding author address: C. Ferro, Exeter Climate Systems, Mathematics Research Institute, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, United Kingdom. E-mail: c.a.t.ferro@exeter.ac.uk

DOI: 10.1175/WAF-D-10-05030.1

TABLE 1. A contingency table representing the frequencies of forecast–observation pairs for which the event and nonevent were forecasted and observed. Entries are also written in terms of the sample size, n; base rate, p; hit rate, H; and false-alarm rate, F.

	Event observed	Nonevent observed	
Event forecasted	$a = Hpn$	$b = F(1-p)n$	$a + b$
Nonevent forecasted	$c = (1-H)pn$	$d = (1-F)(1-p)n$	$c + d$
	$a + c = pn$	$b + d = (1-p)n$	n

p=(a+c)/n

Eskdalemuir observatory and are plotted opposite the forecasts in Fig. 1.

Suppose that the event of interest corresponds to 6-h rainfall exceeding the threshold marked in Fig. 1 and that the event is forecasted to occur if the forecasted rainfall exceeds the same threshold. The elements of the contingency table are then the numbers of points in the four quadrants of Fig. 1. If we construct a contingency table for each of several different thresholds, then we can examine how verification measures change as we move to rarer events. Figure 2 shows that the hit rate and false-alarm rate decrease toward zero and the odds ratio increases toward infinity as the events become rarer. Stephenson et al. (2008) demonstrated that such behavior is common: verification measures such as these typically degenerate to trivial values as the definition of the event is changed to become increasingly rare. This happens because entries a, b, and c in the contingency table tend to decay to zero at unequal rates (Ferro 2007).

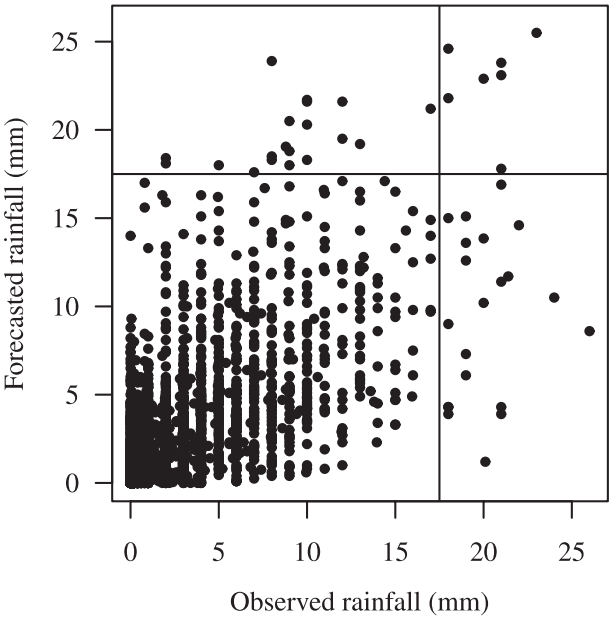

FIG. 1. Forecasted 6-h rainfall accumulations against observations at Eskdalemuir.

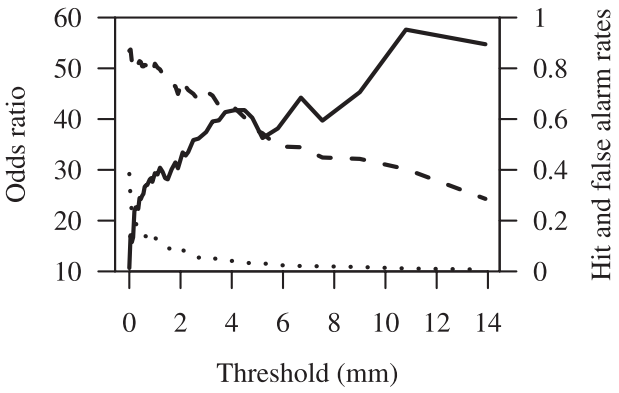

FIG. 2. Odds ratio (OR, solid line), hit rate (H, dashed line), and false-alarm rate (F, dotted line) against threshold (mm) for the Eskdalemuir precipitation forecasts.

Stephenson et al. (2008) proposed a new verification measure, the extreme dependency score or EDS, for summarizing the performance of deterministic forecasts of rare binary events. Instead of degenerating, the EDS converges to a meaningful limit for rare events. We define the EDS in section 2 and then discuss its advantages. Some undesirable properties of the EDS have been noted recently in the literature and we review these criticisms in section 3, while also clarifying the meaning of some properties that have caused confusion elsewhere in the literature. An alternative version of the EDS, the symmetric extreme dependency score or SEDS, was proposed recently by Hogan et al. (2009) in an attempt to overcome some of the shortcomings of the EDS. We discuss the SEDS in section 4 and show that it also suffers from some drawbacks. Motivated by these results, we introduce in section 5 two new measures that overcome all of the undesirable features of the EDS and SEDS. These measures are the extremal dependence index,

$$\mathrm{EDI} = \frac{\log F - \log H}{\log F + \log H}, \tag{1}$$

and the symmetric extremal dependence index,

$$\mathrm{SEDI} = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}. \tag{2}$$

We illustrate the various measures throughout with idealized and operational forecasting examples, and conclude with a summary in section 6.

2. Extreme dependency score

Following Coles et al. (1999), the EDS was defined by Stephenson et al. (2008) as

$$\mathrm{EDS} = \frac{2\log[(a+c)/n]}{\log(a/n)} - 1.$$

The EDS can also be rewritten in the following form (Primo and Ghelli 2009), which will be useful for our treatment later:

$$\mathrm{EDS} = \frac{2\log p}{\log(Hp)} - 1 = \frac{\log p - \log H}{\log p + \log H}, \quad (3)$$

where H is the hit rate and $p = (a + c)/n$ is the base rate, the relative frequency with which the event was observed to occur. Rare events therefore correspond to low base rates.

The EDS is designed to measure the dependence between the forecasts and observations in such a way that it will converge to a meaningful limit for rare events. We explain later in this section that, in order to achieve this meaningful limit, it is necessary to separate out the dependence from any bias. Consequently, the EDS should not be calculated for raw forecasts. Rather, the EDS should be calculated only after recalibrating the forecasts so that the number of forecasted events ($a + b$) equals the number of observed events ($a + c$) in Table 1. If the event is forecasted to occur when a continuous forecast variable exceeds a threshold u, and is observed to occur when a continuous observation variable exceeds a threshold v, then the forecasts can be recalibrated by choosing u and v to be the upper p quantiles of the forecasted and observed variables, respectively (Ferro 2007; Stephenson et al. 2008). When forecasts are recalibrated in this way, the EDS converges to a meaningful limit in the interval $(-1, 1]$ as the base rate decreases. This convergence holds under quite weak conditions on the joint distribution of the forecasts and observations, which imply that $a/n \sim \kappa p^{1/\eta}$ for small p, where $\kappa > 0$ and $0 < \eta \leq 1$ (Ledford and Tawn 1996; Coles et al. 1999; Ferro 2007). Consequently, EDS $\to l = 2\eta - 1$ as $p \to 0$. One way to interpret this limit is in terms of the rate at which the number of hits, a, in Table 1 decays to zero (Stephenson et al. 2008). In particular, a decays at a rate of $p^{2/(1+l)}$ as $p \to 0$, and so

- if $l > 0$, then a decreases slower than p^2;
- if $l = 0$, then a decreases at the same rate as p^2; and
- if $l < 0$, then a decreases faster than p^2.

The expected value of a for calibrated, random forecasts is np^2 because np events are observed and events are forecasted randomly with probability p. The threshold $l = 0$ therefore separates forecasts with extremal dependence that is stronger than for random forecasts ($l > 0$) from those with extremal dependence that is weaker than for random forecasts ($l < 0$).

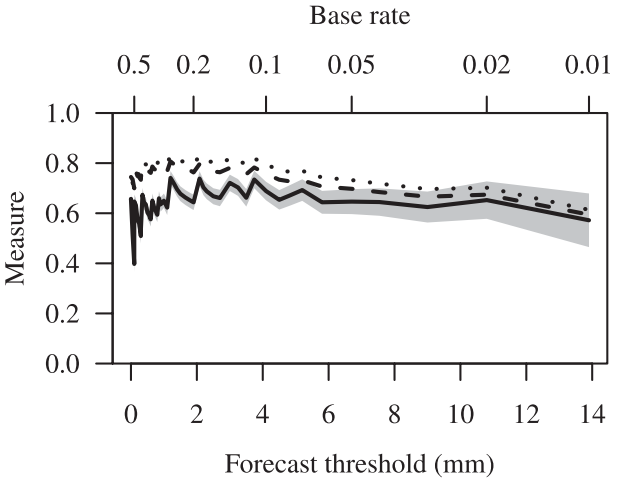

FIG. 3. EDS (solid line) with approximate 95% confidence intervals (gray shading) against forecast threshold (mm) and base rate for the Eskdalemuir precipitation forecasts. EDI (dashed line) and SEDI (dotted line) are also shown.

If the EDS is calculated without recalibrating the forecasts, then it may still converge to a nontrivial limit, but only under stronger conditions on the joint distribution of the forecasts and observations than we needed for the recalibrated case above. For example, for uncalibrated forecasts with $(a + b)/n = q \neq p$, stronger conditions can be imposed to ensure that a/n behaves like $\kappa(pq)^{1/(2\eta)}$ when p is small (Ramos and Ledford 2009). If, in addition, the frequency bias q/p converges to a positive constant β as $p \to 0$, then $a/n \sim \kappa' p^{1/\eta}$ as before, where $\kappa' = \kappa\beta^{1/(2\eta)}$. In this case, the EDS still converges to $2\eta - 1$ and the limit remains meaningful. In other cases, however, the limiting value of the EDS depends on how the bias changes as the base rate decreases, and degenerate limits are possible. This is why the EDS should not be calculated for uncalibrated forecasts of rare events. When the EDS is calculated after recalibrating forecasts, then the bias of the raw forecasts can also be reported in order to provide a more complete description of forecast performance.

We close this section by calculating the EDS for the precipitation forecasts in Fig. 1. The forecasts were recalibrated and the EDS was calculated for base rates ranging from 0.01 to 0.99. The results are plotted in Fig. 3 with approximate 95% confidence intervals of the form EDS $\pm$ 2s, where s is an estimate of the standard error of the EDS. As in Stephenson et al. (2008), s was obtained by fixing n and p, assuming that a is the number of hits in np independent cases, and then employing the delta method (e.g., Davison and Hinkley 1997, p. 45) to obtain

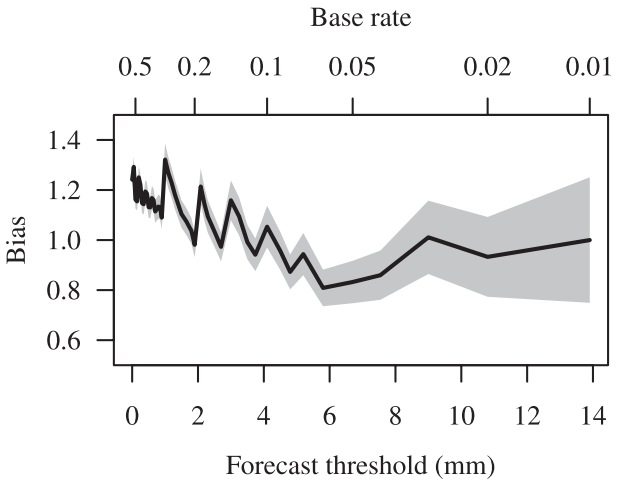

FIG. 4. Bias (solid line) with approximate 95% confidence intervals (gray shading) against forecast threshold (mm) and base rate for the Eskdalemuir precipitation forecasts.

$$s = \frac{2|\log p|}{H(\log p + \log H)^2}\sqrt{\frac{H(1-H)}{pn}}.$$

The graph in Fig. 3 differs slightly from Fig. 5 of Stephenson et al. (2008) because a different range of base rates is considered here, and because Stephenson et al. (2008) added some random noise to the forecasts and observations to mitigate the effects of the discretization of the precipitation totals. Nonetheless, the gross features are similar: the EDS is always positive and converges to a value near two-thirds as the base rate decreases, indicating good skill at forecasting heavy rainfall totals. The oscillations of the EDS at low thresholds are due to the fact that the observations are typically recorded to the nearest millimeter (see Fig. 1) and that the data are denser at lower thresholds, which means that only small changes in the threshold are required for the elements of the contingency table to change. The frequency bias, $(a + b)/(a + c)$, is shown in Fig. 4 and indicates that rainfall events are overforecasted by approximately 20% at low thresholds but that the bias decreases until events are underforecasted by approximately 10% for thresholds greater than 4 mm.

3. Shortcomings of the EDS

In the previous section we reviewed the EDS and pointed out its desirable property of converging to a meaningful limit for rare events. Several shortcomings of the EDS have been noted recently in the literature. We discuss these criticisms below and add some new observations of our own.

a. Base-rate dependence

The notion of verification measures that are base-rate independent has existed for over a century but uncertainty over its meaning still arises in the weather forecasting community. The phrase itself may in fact be relatively recent and the same idea has been given several different labels. For example, Swets (1988) advocated measures that are "independent of event frequencies," Woodcock (1976) referred to "trial independence," and Yule (1912, p. 586f.) advocated measures that are "unaffected by selection." The common definition used by all of these authors is the following one: *a verification measure is base-rate independent if it can be written as a function of only the hit rate and false-alarm rate.*

We know of only limited discussions in the weather forecasting literature of why this is a sensible definition and useful property, so we provide a fuller discussion here before commenting on the EDS specifically.

The starting point is to appreciate that the numbers of observed events and nonevents in a contingency table are beyond the control of the forecasting system being assessed and therefore should not affect the assessment of forecast skill (Mason 2003, p. 41). To understand the implications of this idea, first note that the skill of a forecasting system must be defined with respect to a particular forecasting problem, which is identified with a particular population of events and nonevents. For example, we might wish to know the skill of a system for forecasting whether or not daily rainfall totals at Exeter in southwest England exceed 25 mm, in which case the population might comprise daily exceedances from all days in recent decades. To quantify skill, we obtain a sample from the population and calculate summary measures for the contingency table of corresponding forecast–observation pairs. Importantly, this sample must be representative of the population of interest; otherwise, we would be measuring the skill for a different forecasting problem. For example, if we sampled daily rainfall exceedances from only winters, then we would obtain a different impression of skill than if we sampled from all seasons.

From these ideas it follows that we should seek summary measures that are insensitive to changes in the numbers of observed events and nonevents in the sample as long as the sample otherwise remains representative of the population of interest. This is taken to mean that, however the numbers of events and nonevents in the sample are determined, the sampled events must be representative of the events in the population and the sampled nonevents must be representative of the nonevents in the population. In addition to this insensitivity, measures should be sensitive to other changes in sample

design, and also to changes in the sampled population and forecasting system, since these factors can affect forecast skill.

So, which measures are insensitive in this sense? Under the conditions of the previous paragraph, we can think of the two columns of Table 1 as separate samples, with one representing the population of events and one representing the population of nonevents. While the frequencies of hits and misses in the first column vary with the total number of events, the proportions of hits and misses among the observed events are typically close to the corresponding proportions in the population regardless of however many events are sampled. Both of these proportions are given by the hit rate H. Similarly, the analogous proportions in the second column, which are given by the false-alarm rate F, are largely unaffected by the number of nonevents that are sampled. The hit rate and false-alarm rate are therefore insensitive to the numbers of events and nonevents. Moreover, any other insensitive measure can be written as a function of H and F because, together with the numbers of observed events and nonevents, they define the entire contingency table. Finally, note from Table 1 that knowing the numbers of observed events and nonevents is equivalent to knowing the sample size and base rate, so the measures that are insensitive to both the sample size and base rate are those that can be written as a function of H and F only. This is why such measures are called base-rate independent.

Medical screening provides a helpful analogy. Consider the task of diagnosing whether or not a patient has a disease (the observation) based on the result of a diagnostic test (the forecast). The analog of the base rate in this case is the prevalence of the disease in the population, and the analog of the hit rate is the probability of a positive test result for patients who do have the disease. This probability is just a property of the diagnostic test procedure that will remain constant however many people happen to contract the disease.

Base-rate-independent measures are particularly useful for monitoring forecast performance over time because they are not unduly influenced by variations in the numbers of events and nonevents that are observed. Base-rate-dependent measures, on the other hand, may vary over time because of changes in the base rate only. If we use a base-rate-dependent measure, then we cannot tell if changes in its value are due to changes in skill or to changes in the base rate. If we use a base-rate-independent measure, however, then we know that any change in its value is due to a change in skill.

Mason (2003, p. 47f.) categorizes several popular measures as either base-rate dependent or base-rate independent. In addition to the hit rate and false-alarm rate, for example, the odds ratio is also base-rate independent (Stephenson 2000) because it can be written as

$$\mathrm{OR} = \frac{H(1 - F)}{F(1 - H)}. \tag{4}$$

One example of a base-rate-dependent measure is the frequency bias. Primo and Ghelli (2009) and Ghelli and Primo (2009) noted that the EDS is also base-rate dependent.

Let us illustrate the idea of base-rate dependence with an artificial numerical example. Suppose that a forecasting system produces the contingency table shown in Table 2. Here, $p = 0.1$, $H = 0.55$, $F = 0.05$, and EDS = 0.59. Suppose now that forecasts are made for a second time period in which the sampled population is the same but the base rate happens to be $p = 0.3$. The data in Table 3 exemplify a case in which the forecasting system remains unchanged. The hit rate and false-alarm rate are the same as before but now EDS = 0.34, reflecting its dependence on base rate. The data in Table 4, on the other hand, exemplify a case in which the forecasting system is changed in such a way that its forecasts are unbiased in the second period. Here, the hit rate and false-alarm rate increase to $H = 0.65$ and $F = 0.15$, and EDS = 0.47, reflecting the change in performance of the forecasts as well as the change in base rate. These calculations are summarized in Table 5.

We close this section by addressing two misunderstandings about base-rate dependence that we have noticed in the verification community.

1) The definition of base-rate independence does not mean that base-rate-independent measures cannot also be written in a form that involves the base rate: $H = a/(a + c) = a/(np)$, for example. A measure that cannot be written as a function of only H and F, however, is base-rate dependent.

TABLE 2. An artificial set of unbiased forecasts with base rate 0.1.

	Event observed	Nonevent observed
Event forecasted	55	45
Nonevent forecasted	45	855
	100	900

TABLE 3. An artificial set of biased forecasts with base rate 0.3.

	Event observed	Nonevent observed
Event forecasted	165	35
Nonevent forecasted	135	665
	300	700

TABLE 4. An artificial set of unbiased forecasts with base rate 0.3.

	Event observed	Nonevent observed
Event forecasted	195	105
Nonevent forecasted	105	595
	300	700

TABLE 5. Values of four verification measures for the data in Tables 2–4.

	p	H	F	EDS	SEDS	EDI	SEDI
Table 2	0.1	0.55	0.05	0.59	0.59	0.67	0.71
Table 3	0.3	0.55	0.05	0.34	0.56	0.67	0.71
Table 4	0.3	0.65	0.15	0.47	0.47	0.63	0.66

2) There are many situations in which H and F will change in tandem with the base rate, but only if whatever causes the base rate to change also changes the forecast skill. For example, if we change to assessing a forecasting system in winter rather than in summer and different physical processes predominate in the two seasons, then the population represented by the sample changes and both the base rate and skill may change (see also Hamill and Juras 2006). As before, if we use a base-rate-dependent measure, then we cannot tell if changes in its value are due to changes in skill or to changes in the base rate, but if we use a base-rate-independent measure, then we know that any change in its value is due to a change in skill.

   Another example arises in the verification of forecasts of extreme events. Recall Fig. 2 in which we plotted three verification measures against the precipitation threshold used to define the event. See Göber et al. (2004) for similar examples. As the threshold increases, the definition of the event changes. Therefore, the base rate changes but so does the forecast skill: both the population and the forecasting system are being changed, so there is no reason to expect the skill to remain constant. Instead, as Fig. 2 illustrates, most measures degenerate to trivial values as rarer events are considered, but this is not due to base-rate dependence: even base-rate-independent measures such as H can decay to zero. Measures degenerate because they quantify aspects of forecast quality for which it is intrinsically hard to maintain the same level of performance as events become rarer. (Of course, maintaining a nonzero hit rate for rare events is possible in theory. Investigating why forecasting systems typically fail to do so would be an interesting exercise.) The EDS, on the other hand, measures the rate at which forecast performance degenerates and therefore need not degenerate itself.

b. Hedging

We have seen that the EDS is base-rate dependent. A second criticism of the EDS is that it can be hedged (Primo and Ghelli 2009; Ghelli and Primo 2009; Brill 2009). There is no consensus in the literature about what is meant by hedging for deterministic forecasts (Jolliffe 2008) and so we clarify below the senses in which the EDS is hedgable.

Hedging can be defined as issuing a forecast that differs from one's judgment. Unless a forecaster is certain about the future, a deterministic forecast will differ from his judgment and, in this sense, all deterministic forecasts are hedged forecasts (Jolliffe 2008) and all verification measures for deterministic forecasts can be hedged.

The notion of consistency (Murphy and Daan 1985) provides another way to define hedging for deterministic forecasts. A verification measure is said to be consistent with a particular rule for converting probabilistic beliefs into deterministic forecasts if the forecaster will optimize their expected score by following that rule. For forecasts of binary events, any measure is consistent with a rule of the form ''forecast the event when your belief exceeds a specific threshold'' (Mason 2003). The value of this optimal threshold depends on the measure and, possibly, on the entries in the contingency table, but can be computed. So all verification measures for forecasts of binary events are consistent with some rule. If a forecaster is directed to employ a specific rule to produce deterministic forecasts but the forecasts are evaluated using a measure that is inconsistent with that rule, then the measure could be hedged by disregarding the directive and employing the rule with which the measure is consistent. In such a situation, we may say that the measure is hedgable. To find the optimal threshold for the EDS, suppose that a forecaster's belief that the event will occur is a probability q and that the entries in the contingency table are all nonzero. If the event were to be forecasted, then the number of hits, a, will be incremented by 1 with probability q and the number of false alarms, b, will be incremented by 1 with probability $1 - q$. The forecaster's expected value of the EDS is therefore

$$q\left\{\frac{2\log[(a+c+1)/(n+1)]}{\log[(a+1)/(n+1)]} - 1\right\} + (1-q)\left\{\frac{2\log[(a+c)/(n+1)]}{\log[a/(n+1)]} - 1\right\}.$$

Similarly, if the event were not forecasted, then the expected value is

$$q\left\{\frac{2\log[(a+c+1)/(n+1)]}{\log[a/(n+1)]}-1\right\} + (1-q)\left\{\frac{2\log[(a+c)/(n+1)]}{\log[a/(n+1)]}-1\right\}.$$

The former is greater than the latter if and only if $q > 0$. Thus, the optimal threshold for the EDS is zero, and so the EDS is consistent with the rule "always forecast the event." This rule is unlikely ever to be issued as a directive and therefore the EDS will be hedgable whenever directives are employed.

Another way to think about hedging is to determine whether or not there exist "unskillful" modifications of the forecasts that guarantee an improvement in the value, or expected value, of the verification measure. This is related to equitability (Jolliffe 2008) but equitability ensures only that the expected score cannot be improved by choosing one set of random forecasts over another; the score may still be improved by other choices of unskillful forecasts. We return to equitability later in this section. Another type of unskillful modification is to switch forecasts randomly from events to nonevents or vice versa (Stephenson 2000). The EDS is prone to hedging in this sense because the EDS attains its optimal value of 1 when $H = 1$, and this can be achieved by always forecasting the event (Primo and Ghelli 2009). Reassigning all forecasts of nonevents to forecasts of events therefore ensures that EDS = 1. A general approach to constructing measures that are not hedgable in this sense has yet to be advanced, but a necessary condition for positively oriented, base-rate-independent measures is that the measure should be strictly increasing in the hit rate and strictly decreasing in the false-alarm rate. To see that this is a necessary but not sufficient condition for preventing hedging, suppose that we switch forecasts of events to nonevents with probability α. Then, the hit rate and false-alarm rate are both strictly decreasing in α. The derivative of a base-rate-independent measure S with respect to α can be written as

$$\frac{\partial S}{\partial \alpha} = \frac{\partial S}{\partial H}\frac{\partial H}{\partial \alpha} + \frac{\partial S}{\partial F}\frac{\partial F}{\partial \alpha}.$$

Therefore, if S is strictly decreasing in both H and F, or strictly decreasing in one and constant in the other, then S is strictly increasing in α and hedging is possible. Similarly, if we switch forecasts of nonevents to events with probability α, then hedging is possible if S is strictly increasing in both H and F, or strictly increasing in one and constant in the other. Assuming that S is not constant in both H and F, then the derivatives of S with respect to H and F must be of opposite signs, and for positively oriented measures we should require S to be strictly increasing in H and strictly decreasing in F, rather than vice versa.

The derivative of the EDS (3) with respect to H is

$$\frac{-2\log p}{H(\log H + \log p)^2},$$

which, as required, exceeds zero when $p < 1$. The EDS does not depend on the false-alarm rate, however, and so it is prone to overforecasting, as we have seen (Primo and Ghelli 2009).

Hedgable measures have also been defined by Marzban (1998) as those measures that cannot be optimized for unbiased (calibrated) forecasts. The EDS is optimized if and only if $c = 0$ and $a \neq 0$ so that $H = 1$ and $p \neq 0$. This is achieved for perfect forecasts that have no bias, but can also be achieved for biased forecasts by always forecasting the event, as noted by Brill (2009) and Hogan et al. (2009).

c. *Regularity*

Signal detection theory (Swets 1988) makes a useful distinction between the actual performance of a set of forecasts and the potential performance of the forecasting system (Harvey et al. 1992). So far we have considered measures of actual performance, summary measures of a single contingency table that can usually be written as functions of H, F, and, in the case of base-rate-dependent measures, p. Signal detection theory is based on the idea that the event is forecasted if a decision variable exceeds a decision threshold. Figure 1 provides an example: the forecasted rainfall is the decision variable and the event is forecasted if a threshold is exceeded. A contingency table then reflects the performance of the forecasting system for a particular decision threshold. The potential performance of the forecasting system, on the other hand, is considered to be independent of the decision threshold. Instead, the potential performance is determined by two frequency distributions: the distribution of the decision variable prior to events being observed, and the distribution of the decision variable prior to nonevents being observed. These two distributions are usually displayed as a relative operating characteristics (ROC) curve, which is the graph of the hit rate against the false-alarm rate as the decision threshold is varied over the range of the decision variable (e.g., Mason 1982; Mason and Graham 1999). The empirical ROC curve for the forecasts in Fig. 1 is shown in Fig. 5. A ROC curve encapsulates the potential performance of the forecasting system and each point on the curve identifies the actual performance of the forecasts for a particular decision threshold.

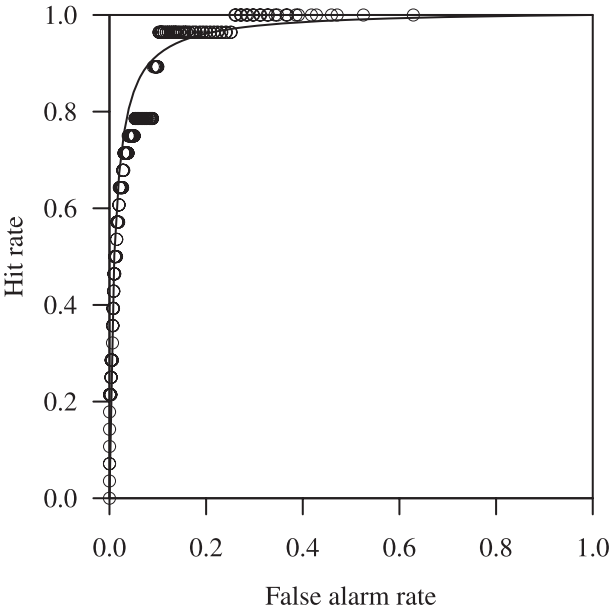

FIG. 5. The empirical ROC curve (circles) for the forecasts of Eskdalemuir precipitation exceeding 17.5 mm. An isopleth (solid line) of the odds ratio is also shown.

According to signal detection theory, measures of the skill of a forecasting system should be numerical summaries of the system's ROC curve. Note that the ROC curve and derived summary measures are base-rate independent because they depend on only hit rates and false-alarm rates, and cannot be hedged because the ROC curve is defined by all possible decision thresholds whereas hedging relates to choosing a particular decision threshold. One popular summary of ROC curves is the area under the curve (e.g., Mason and Graham 2002; Marzban 2004). An alternative way of summarizing ROC curves is to find a verification measure whose value is the same at each point on the system's ROC curve. If such a measure can be found, then the ROC curve is said to be an isopleth of the measure, which then provides a good summary of the system's skill (Swets 1986). An isopleth of the odds ratio is shown in Fig. 5. From the definition of the odds ratio (4), we find that the isopleth satisfying OR = k is the graph of the function

$$H = \frac{kF}{1 - (k+1)F}.$$

In Fig. 5 the isopleth of the odds ratio tends to lie above the points nearer to (0, 0) and below the points nearer to (1, 1) and is therefore a poor fit to the empirical ROC curve in this example. If an isopleth of a verification measure does provide a good fit to the ROC curve of a forecasting system, however, then that measure cannot be hedged by changing the decision threshold. The same measure may well provide a poor fit to the ROC curve of another system, in which case the measure could be hedged by the second system. In contrast, the area under the curve is unhedgable for all systems.

Almost all empirical ROC curves for real forecasting systems possess the following two properties: the curve connects the points (0, 0) and (1, 1), and otherwise remains strictly inside the unit square. The isopleths of verification measures that provide good summaries of ROC curves must therefore also satisfy these two properties. Verification measures for which this is true are called regular (e.g., Mason 2003, p. 62). The odds ratio is regular but the EDS is nonregular because its isopleths correspond to horizontal lines on ROC diagrams: EDS = k if and only if

$$H = p^{(1-k)/(1+k)},$$

which is constant in F.

d. Range

Now we highlight a drawback of the EDS that has been overlooked by previous authors. Both Coles et al. (1999) and Stephenson et al. (2008) stated that the EDS lies in the interval $(-1, 1]$. In fact, when the EDS is calculated for calibrated forecasts, its range of possible values depends on the base rate (Segers and Vandewalle 2004, p. 345). The upper bound of the EDS is always 1 but the lower bound is -1 only when $p \leq \frac{1}{2}$. The contingency table yields the inequality $c \leq (1 - p)n$, which implies that $a = pn - c \geq (2p - 1)n$ and therefore

$$\mathrm{EDS} \geq \frac{2\log p}{\log(2p - 1)} - 1$$

when $p > \frac{1}{2}$. Although this condition does not refer to rare events, we would like measures to have good properties for all base rates if possible. A measure can be difficult to interpret if its range of possible values depends on the base rate. For example, if a set of forecasts with $p = \frac{3}{4}$ achieves an EDS equal to its lowest possible value, -0.17, does that indicate a better or worse level of performance than forecasts with an EDS of -0.6 when $p = \frac{1}{4}$?

e. Equitability

Another desirable property of verification measures is equitability (Gandin and Murphy 1992). A measure is equitable if its expected value is the same for all random forecasts. Hogan et al. (2010) noted that many measures (including the so-called equitable threat score) are equitable only in the limit as the sample size n increases to infinity, and called this weaker property asymptotic

equitability. When the base rate is p and the event is forecasted to occur at random with probability q, the expected value of a is npq and so a/n converges to pq as $n \to \infty$. In this case, $H \to q$ and so EDS $\to (\log p - \log q)/(\log p + \log q)$. The limit of the EDS for random forecasts therefore varies with the forecast probability q and so the EDS is not asymptotically equitable. In fact, random forecasts with $q = 1$ (for which the event is always forecasted) maximize the EDS. However, if the random forecasts are recalibrated, then $H \to p$ as $n \to \infty$ and so EDS $\to 0$ always.

If an asymptotically equitable measure is also increasing in a for fixed values of $a + b$ and $a + c$, then, for large sample sizes, the measure will exceed the expected value for random forecasts if and only if the forecasts' performance is better than the expected performance of random forecasts. The expected score for random forecasts therefore provides a meaningful origin that separates better-than-random forecasts from worse-than-random forecasts. This property holds for the EDS when it is calculated for recalibrated forecasts: EDS > 0 if and only if $a > np^2$. Figure 3 shows that the EDS is always positive for our precipitation forecasts, indicating that they perform better than random forecasts for all base rates. For uncalibrated forecasts, zero is no longer a meaningful origin: if $q > p$, then the EDS can be positive for forecasts that are worse than random, while if $q < p$, then the EDS can be negative for forecasts that are better than random.

f. Complement symmetry

So far, we have identified five undesirable properties of the EDS: it is base-rate dependent, it has nonregular isopleths, its range changes with the base rate, and, if the EDS is used without recalibrating the forecasts, it is not asymptotically equitable and can be hedged. Sometimes it is impossible or undesirable to recalibrate forecasts (Hogan et al. 2009) and in such situations we suggest that the EDS should not be used: there is no guarantee of a meaningful limit for extreme events, and all five of the aforementioned drawbacks will apply. In the remainder of this section we discuss three more properties that have been advocated as desirable in the literature and that are not satisfied by the EDS. In these cases, however, we argue that there are no general reasons for preferring measures with these properties.

Measures that are invariant to relabeling the event as the nonevent and the nonevent as the event are called complement symmetric by Stephenson (2000). Relabeling in this way rearranges the elements of the contingency table from (a, b, c, d) to (d, c, b, a). If the original contingency table has base rate p, hit rate H, and false-alarm rate F, then the new table has base rate $1 - p$, hit rate $1 - F$, and false-alarm rate $1 - H$. The value of the EDS therefore typically changes after relabeling and so the EDS is not complement symmetric.

At first sight, complement symmetry is a desirable property: it seems unfair to change the skill of the system just because we decide to start calling events ''nonevents'' and nonevents ''events'' when the sampled population and forecasting system are unchanged. Here, it is important to distinguish between actual and potential levels of performance. We should expect actual performance to change after taking complements: the hit rate and false-alarm rate typically change and so the forecasts have a different quality. If we wish to summarize actual performance, then there is no reason, therefore, to use a complement symmetric measure. The potential performance of the forecasting system, on the other hand, should be unaffected by taking complements. We discussed earlier how the ROC curve encapsulates potential performance and that summaries of ROC curves can provide measures of potential performance. A popular example is the area under the ROC curve. On taking complements, hit rates and false-alarm rates are changed in such a way that a system's ROC curve is reflected in the negative diagonal, the line $H = 1 - F$. The area under the ROC curve is invariant to this reflection and so that measure of potential performance is invariant to taking complements.

Now consider a verification measure $S(H, F)$ with an isopleth that corresponds to the system's ROC curve. If the ROC curve is symmetric about the negative diagonal, then a little geometry shows that $S(1 - F, 1 - H) = S(H, F)$ and so the measure will be invariant to taking complements. If the ROC curve is not symmetric about the negative diagonal, however, the measure will not be invariant to taking complements. The measure is still an appropriate summary of potential performance, but evaluating the measure after taking complements would not provide a good summary of potential performance. This is because the reflection of the system's ROC curve will not correspond to an isopleth of $S(H, F)$. Instead, the reflected ROC will be an isopleth of the measure $S^*(H, F) = S(1 - F, 1 - H)$ and so we would need to evaluate S^* for the complementary events in order to obtain a measure of potential performance. If we wish to summarize potential performance using a measure whose isopleth corresponds to the system's ROC curve, then the measure must be chosen so that the isopleth matches the ROC curve even if the curve is asymmetric about the negative diagonal, in which case a complement asymmetric measure will be necessary.

g. Transpose symmetry

Hogan et al. (2009) criticize the EDS because, when calculated for biased forecasts, it is not invariant to

transposing the contingency table (interchanging elements b and c), which amounts to switching the roles of the observations and the forecasts. Hogan et al. (2009) also claim that transpose symmetric measures are more difficult to hedge. However, the relationship between hedging and transpose symmetry is unclear: the measure a/n for example is transpose symmetric but is optimized by always forecasting the event, while the Peirce skill score, $H - F$, is transpose asymmetric but is unhedgable in the sense of Stephenson (2000). Transpose symmetry is appropriate if both types of forecasting error, misses (c) and false alarms (b), are to be penalized equally but there appear to be no other reasons for requiring measures of forecast performance to be transpose symmetric.

h. Linearity

Hogan et al. (2009) also introduce a concept of linearity, which requires that the difference

$$S(a + 1, b - 1, c - 1, d + 1) - S(a, b, c, d)$$

should be invariant to the values of a, b, c, and d; see also Hubálek (1982). This property enables a half-life of forecast quality to be defined without ambiguity but other motivations for this property are unclear. Furthermore, measures that have nondegenerate limits for extremes require nonlinear transformations of the elements in the contingency table and are therefore unlikely to satisfy this notion of linearity. Indeed, the EDS is nonlinear (see Hogan et al. 2009). Nonetheless, measures that are approximately linear may be preferable to measures that are very nonlinear.

4. Symmetric EDS

In the previous section we showed that the EDS has several undesirable properties. Hogan et al. (2009) developed a new version of the EDS, the symmetric extreme dependency score or SEDS, which overcomes some of these problems. We discuss SEDS in this section, noting its advantages and remaining disadvantages.

The SEDS is defined as

$$\text{SEDS} = \frac{\log[(a+b)(a+c)/n^2]}{\log(a/n)} - 1,$$

and can also be written as

$$\begin{aligned}\text{SEDS} &= \frac{\log(pq)}{\log(Hp)} - 1 \\ &= \frac{\log q - \log H}{\log p + \log H},\end{aligned} \tag{5}$$

where $q = (a + b)/n$ is the relative frequency with which the event was forecasted. SEDS differs from EDS in (3) by replacing the base rate p with q in the numerator. As a result, SEDS > EDS if and only if $q < p$. If the forecasts are recalibrated so that $q = p$, then SEDS equals EDS.

The primary aim of Hogan et al. (2009) was to obtain a measure that can be used for uncalibrated forecasts, that is transpose symmetric, and that retains a meaningful limit as the base rate tends to zero. SEDS is transpose symmetric because it is symmetric in b and c. SEDS also has a meaningful limit, but only in certain circumstances. For example, if the frequency bias q/p converges to a positive constant as the base rate tends to zero, then SEDS has the same limit as EDS because

$$\begin{aligned}\text{SEDS} &= \frac{\log(q/p) + \log p - \log H}{\log p + \log H} \\ &= \frac{\log(q/p)}{\log(pH)} + \text{EDS}\end{aligned}$$

and $\log(pH) \to -\infty$ as $p \to 0$. If the bias does not converge to a positive constant, then the limiting value of SEDS depends on how the bias changes with the base rate. This compromises the interpretation of SEDS and is why we recommend calculating EDS for only recalibrated forecasts.

SEDS does enjoy some advantages over EDS. We show in appendix A, for example, that SEDS is asymptotically equitable and more difficult to hedge than EDS. On the other hand, SEDS is still base-rate dependent, has a range that depends on the base rate, and is nonregular. These latter properties are demonstrated in appendix A too and a summary is provided in Table 6. Properties of the equitable threat score, which typically degenerates to zero with the base rate (Stephenson et al. 2008), are also included in Table 6 for comparison (Mason 2003, 52–54).

If SEDS is calculated for uncalibrated forecasts, then its standard error can be estimated by

$$s_{\text{SEDS}} = \frac{|\log p + \log q|}{H(\log p + \log H)^2}\sqrt{\frac{H(1 - H)}{pn}}.$$

This is obtained via the delta method, previously employed for EDS in section 2.

For the reason given earlier, we do not recommend calculating SEDS for uncalibrated forecasts if the aim is to understand the extremal dependence between the forecasts and the observations. When the forecasts are recalibrated, SEDS equals EDS and so we do not calculate SEDS for the precipitation forecasts in Fig. 1. Let us instead calculate SEDS for the forecasts in Tables 2–4. Results are summarized in Table 5. From Table 2 we obtain SEDS = EDS = 0.59 because the forecasts are calibrated. For the uncalibrated forecasts in Table 3 with

TABLE 6. Properties of five verification measures.

	ETS	EDS	SEDS	EDI	SEDI
Nondegenerate limit	×	✓	✓	✓	✓
Base-rate independent	×	×	×	✓	✓
Nontrivial to hedge	✓	×	✓	✓	✓
Regular	×	×	×	✓	✓
Fixed range [−1, 1]	×	×	×	✓	✓
Asymptotically equitable	✓	×	✓	✓	✓
Meaningful origin	✓	×	✓	✓	✓
Complement symmetric	✓	×	×	×	✓
Transpose symmetric	✓	×	✓	×	×

the same hit rate and false-alarm rate but greater base rate, we obtain SEDS = 0.56. This is a less dramatic reduction than that experienced by EDS, which decreases to 0.34 for these data, but still illustrates the dependence of SEDS on the base rate. For the calibrated forecasts in Table 4, we obtain SEDS = EDS = 0.47 once more, reflecting the changes in hit rate, false-alarm rate, and base rate.

5. Extremal dependence indices

In the previous section we showed that, although SEDS is asymptotically equitable and more difficult to hedge than EDS for uncalibrated forecasts, SEDS is still base-rate dependent, nonregular, and has a range that depends on the base rate. We have also argued that SEDS should be calculated for only recalibrated forecasts if the purpose is to understand extremal dependence, in which case SEDS is identical to EDS. In this section we propose two new measures that avoid all of the shortcomings of EDS. Again, we recommend that the measures are calculated for recalibrated forecasts only. The difference between these two new versions of EDS is that one is complement symmetric and the other is complement asymmetric.

The first new measure is the extremal dependence index or EDI (1). The reasoning behind this definition is as follows. To obtain a base-rate independent measure, the measure should be a function of F and H only. Since, for recalibrated forecasts, $F = p(1 - H)/(1 - p)$ behaves like p as $p \to 0$, we can consider replacing p with F in the definition of EDS (3). Thus, we obtain a base-rate-independent measure that has the same meaningful limit as EDS for recalibrated forecasts.

EDI also overcomes other disadvantages of EDS. We show in appendix B, for example, that EDI is regular, asymptotically equitable, more difficult to hedge than EDS, and always has range [−1, 1]. It is neither transpose symmetric nor complement symmetric. These properties are summarized in Table 6.

The second new measure is the symmetric extremal dependence index or SEDI (2). This is similar to EDI but includes terms $\log(1 - F)$ and $\log(1 - H)$. Since F and H both decay to zero as $p \to 0$, these extra terms play a negligible role asymptotically and therefore SEDI has the same meaningful limit as EDS and EDI for recalibrated forecasts. Including the $\log(1 - F)$ and $\log(1 - H)$ terms merely makes SEDI complement symmetric. Otherwise, SEDI shares the same properties as EDI, as shown in appendix B and summarized in Table 6. The base-rate independence of EDI and SEDI is illustrated numerically in Table 5.

The numerator of SEDI is

$$-\log\left[\frac{H(1 - F)}{F(1 - H)}\right],$$

a transformation of the odds ratio (4). SEDI can therefore be thought of as a normalized version of the log odds ratio, where the normalization transforms the odds ratio to fall in the interval [−1, 1] and ensures a meaningful limit as the base rate decreases to zero. This may be compared with the measure $Q = (\mathrm{OR} - 1)/(\mathrm{OR} + 1)$ proposed by Yule (1900), which also transforms the odds ratio to the interval [−1, 1] but which typically degenerates to either −1 or 1 for rare events.

EDS and EDI are equal if $F = p$ or $H = 1$, and otherwise satisfy the following relationship: EDI > EDS if and only if $F < p$, which is usually the case for low base rates. It is also possible to show that SEDI ≥ EDI if and only if $|H - 1/2| \le |F - 1/2|$, which is also usually the case for low base rates.

Let us compare EDI and SEDI with EDS for the recalibrated precipitation forecasts in Fig. 1. Further applications of the delta method show that an estimate of the standard error of the EDI for recalibrated forecasts is

$$s_{\mathrm{EDI}} = \frac{2\left|\log F + \dfrac{H}{1 - H}\log H\right|}{H(\log F + \log H)^2}\sqrt{\frac{H(1 - H)}{pn}},$$

and an estimate of the standard error of the SEDI is

$$s_{\mathrm{SEDI}} = \frac{2\left|\dfrac{(1 - H)(1 - F) + HF}{(1 - H)(1 - F)}\log[F(1 - H)] + \dfrac{2H}{1 - H}\log[H(1 - F)]\right|}{H\{\log[F(1 - H)] + \log[H(1 - F)]\}^2}\sqrt{\frac{H(1 - H)}{pn}}.$$

The values of EDI and SEDI are superimposed onto Fig. 3. The estimated standard errors of EDI and SEDI are close to the estimated standard errors of EDS for all base rates, but we suppress the confidence intervals for EDI and SEDI in the figure to preserve clarity. As expected, the scores satisfy the ordering SEDI > EDI > EDS for most thresholds and converge to the same limit at low base rates.

6. Conclusions

We have reviewed two existing measures for quantifying the performance of deterministic forecasts of rare binary events. EDS has several drawbacks, including being susceptible to hedging by overforecasting and being base-rate dependent. SEDS is harder to hedge than EDS but is still base-rate dependent. In the course of this review we have attempted to define and explain clearly the notions of base-rate dependence, hedging, and complement symmetry. We have also introduced two new measures that overcome all of the disadvantages of EDS and SEDS. One of the new measures is complement symmetric, and the other is complement asymmetric. We recommend that the new measures should be preferred to EDS and SEDS for examining the performance of rare-event forecasts. We emphasize that forecasts must be recalibrated before computing these measures if a clear understanding of forecast performance for rare events is desired.

The relative frequency of correct forecasts of the event typically behaves like αp^{β} for small base rates p, where $\alpha > 0$ and $\beta \geq 1$ are constants. The limiting values of our measures are informative for β but the scaling constant α may also be important. Information about both α and β can be obtained using the approach described by Ferro (2007).

Acknowledgments. We thank Robin Hogan and two anonymous reviewers for their comments on earlier versions of this paper, and members of the European Centre for Medium-Range Weather Forecasts Technical Advisory Committee Subgroup on Verification Measures for conversations about this work.

APPENDIX A

Properties of SEDS

We derive the properties of SEDS (see section 4) that are summarized in Table 6.

a. Base-rate dependence

SEDS is base-rate dependent because its value can change even when H and F are unchanged, as demonstrated by the numerical examples at the end of section 4.

b. Hedging

We saw that EDS is consistent with the directive "forecast the event when your belief exceeds zero," effectively "always forecast the event." SEDS, on the other hand, is consistent with a directive for which the belief threshold is a complicated function of the entries in the contingency table. This threshold is typically nonzero and therefore SEDS is consistent with a nontrivial directive.

SEDS is also less prone than EDS to hedging by random switching of forecasts. For example, if a proportion α of forecasts are switched randomly from forecasts of the event to forecasts of the nonevent (as in Stephenson 2000), then the entries in the contingency table become $(a - \alpha a, b - \alpha b, c + \alpha a, d + \alpha b)$ and SEDS becomes

$$\mathrm{SEDS}' = \frac{\log[(a+b)(a+c)/n^2] + \log(1-\alpha)}{\log(a/n) + \log(1-\alpha)} - 1.$$

Now, SEDS′ > SEDS if and only if

$$\{\log[(a+b)(a+c)/n^2] + \log(1-\alpha)\}\log(a/n) > \log[(a+b)(a+c)/n^2][\log(a/n) + \log(1-\alpha)],$$

and canceling terms common to both sides leaves

$$\log(1-\alpha)\log(a/n) > \log(1-\alpha)\log[(a+b)(a+c)/n^2].$$

Dividing through by the left-hand side and subtracting 1 shows that SEDS′ > SEDS if and only if SEDS < 0. In other words, random switching of forecasts from events to nonevents will improve SEDS if and only if SEDS < 0.

We noted earlier that EDS is strictly increasing in the hit rate but does not decrease as the false-alarm rate increases. In contrast, SEDS is strictly increasing in the hit rate and is also strictly decreasing in the false-alarm rate. To see this, note that the derivative of SEDS with respect to the false-alarm rate F is $(1 - p)/[q \log(Hp)]$, which is negative when $p < 1$ and zero when $p = 1$. The derivative of SEDS with respect to the hit rate H is $[\beta \log\beta - q \log(pq)]/[Hq(\log\beta)^2]$, where $\beta = Hp$ and $\max\{0, p + q - 1\} \leq \beta \leq \min\{p, q\}$. The denominator of this derivative is positive while the numerator is positive when $p < 1$ and zero when $p = 1$. The proof of this last statement is fairly straightforward but tedious. A simple approach is to consider three cases separately: first, when $p + q - 1 < 1/e < \min\{p, q\}$ and the numerator is minimized at $\beta = 1/e$; second, when $\min\{p, q\} < 1/e$ and the numerator is minimized at $\beta = \min\{p, q\}$; and third, when $p + q - 1 > 1/e$ and the numerator is minimized at $\beta = p + q - 1$. In all cases, it is possible to show that the minimum value achieved by the numerator is nonnegative.

These results all suggest that SEDS is harder to hedge than EDS. However, SEDS is only optimized for perfect, unbiased forecasts and so is hedgable in the sense of Marzban (1998).

c. Regularity

SEDS is nonregular. One way to see this is to use the identity $q = Hp + F(1 - p)$ to show that $H = p^{(1-\text{SEDS})/\text{SEDS}}$ when $F = 0$ and SEDS $\neq 0$. Therefore, the isopleths of SEDS typically fail to pass through the point (0, 0).

d. Range

The range of possible values of SEDS depends on the base rate. As for EDS, the maximum possible value of SEDS is always 1 but, unlike EDS, this maximum is obtained only for perfect forecasts with $b = c = 0$. To see this, note that SEDS ≤ 1 if and only if

$$\log\left[\frac{(a+b)(a+c)}{n^2}\right] \geq 2\log\left(\frac{a}{n}\right),$$

which holds if and only if $(a + b)(a + c) \geq a^2$. This is always true, with equality if and only if $b = c = 0$. Like EDS, the minimum possible value of SEDS depends on the base rate. Following an argument similar to that in section 3d, if $p + q \leq 1$, then the lower bound is -1, but

$$\text{SEDS} \geq \frac{\log(pq)}{\log(p+q-1)} - 1$$

if $p + q > 1$.

e. Equitability

Hogan et al. (2009) showed that SEDS is asymptotically equitable. For a contingency table with $a + b = qn$ and $a + c = pn$, the expected value of a is pqn for random forecasts, in which case SEDS $= \log(pq)/\log(pq) - 1 = 0$. SEDS is also increasing in a for fixed p and q so that SEDS exceeds zero if and only if the forecasts perform better than random forecasts.

f. Complement symmetry

SEDS is not complement symmetric because replacing (a, b, c, d) with (d, c, b, a) typically changes the value of SEDS.

g. Transpose symmetry

SEDS is transpose symmetric because it is symmetric in b and c.

h. Linearity

Hogan et al. (2009) showed that SEDS is approximately linear.

APPENDIX B

Properties of EDI and SEDI

We derive the properties of the new measures, EDI and SEDI, that are summarized in Table 6.

a. Base-rate dependence

Both EDI and SEDI are base-rate independent because they are functions of H and F only.

b. Hedging

As for SEDS, both EDI and SEDI are consistent with directives for which the belief thresholds are complicated functions of the entries in the contingency table. These thresholds are typically nonzero and therefore EDI and SEDI are consistent with nontrivial directives.

If a proportion α of forecasts are switched randomly from forecasts of the event to forecasts of the nonevent, then EDI becomes

$$\frac{\log F - \log H}{\log F + \log H + 2\log(1 - \alpha)},$$

which exceeds $(\log F - \log H)/(\log F + \log H)$ if and only if $F < H$. Therefore, random switching of forecasts from events to nonevents will improve EDI if and only if EDI < 0. Numerical experiments indicate that the same is true for SEDI, but we have no proof of this at present.

It is straightforward to show that, as for SEDS, both EDI and SEDI are strictly increasing in the hit rate and strictly decreasing in the false-alarm rate.

Finally, EDI = 1whenever $c = 0$ and a, b, and d are nonzero. Thus, EDI can be optimized for biased forecasts. In contrast, SEDI is undefined whenever one or more entries in the contingency table are zero. Therefore, SEDI only approaches its maximum value of 1 as the forecasts become close to perfect. These results all suggest that EDI and SEDI are both harder to hedge than EDS.

c. Regularity

Both EDI and SEDI are regular. For the isopleths of EDI, we have EDI $= k$ if and only if

$$H = F^{(1-k)/(1+k)},$$

a form of regular ROC curve known as a power ROC (e.g., Swets 1996, p. 75). These isopleths are asymmetric about the negative diagonal ($H = 1 - F$) unless EDI = 0. In other words, EDI is not complement symmetric. The regular ROC isopleths of SEDI $= k$ are defined implicitly by

$$[F(1 - H)]^{1-k} = [H(1 - F)]^{1+k}.$$

These isopleths are symmetric about the negative diagonal, and so SEDI is complement symmetric.

d. Range

Unlike EDS and SEDS, the range of EDI is always $[-1, 1]$ because

$$\begin{aligned}\mathrm{EDI} \le 1 &\Leftrightarrow \log F - \log H \ge \log F + \log H \\ &\Leftrightarrow 2\log H \le 0 \Leftrightarrow H \le 1,\end{aligned}$$

which is always true, and

$$\begin{aligned}\mathrm{EDI} \ge -1 &\Leftrightarrow \log F - \log H \le -\log F - \log H \\ &\Leftrightarrow 2\log F \le 0 \Leftrightarrow F \le 1,\end{aligned}$$

which is always true. Furthermore, EDI is maximized whenever $H = 1$ and minimized whenever $F = 1$. By a similar argument, SEDI always lies in the interval $[-1, 1]$ but, because it is undefined when any entry in the contingency table is zero, SEDI only approaches its maximum value as $H \to 1$ and $F \to 0$, and approaches its minimum value as $H \to 0$ and $F \to 1$.

e. Equitability

Like SEDS, both EDI and SEDI are asymptotically equitable. For random forecasts with $(a + b)/n = q \neq p$ and $a/n = pq$, we have $H = F = q$, which yields EDI = SEDI = 0. Furthermore, EDI and SEDI exceed zero if and only if $a/n > pq$ so that zero demarcates forecasts that are better than random and those that are worse than random.

f. Complement symmetry

EDI is not complement symmetric because replacing (a, b, c, d) with (d, c, b, a) typically changes the value of EDI. In contrast, SEDI is complement symmetric because replacing H with $1 - F$ and F with $1 - H$ leaves the measure unchanged.

g. Transpose symmetry

Neither EDI nor SEDI is transpose symmetric because switching b and c typically changes their values.

h. Linearity

Numerical experiments (not shown) similar to those in Hogan et al. (2009) indicate that EDI and SEDI are more nonlinear than SEDS.

REFERENCES

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

Coles, S., J. Heffernan, and J. Tawn, 1999: Dependence measures for extreme value analyses. *Extremes,* **2,** 339–365.

Davies, T., M. J. P. Cullen, A. J. Malcolm, M. H. Mawson, A. Staniforth, A. A. White, and N. Wood, 2005: A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Quart. J. Roy. Meteor. Soc.,* **131,** 1759–1782.

Davison, A. C., and D. V. Hinkley, 1997: *Bootstrap Methods and Their Application*. Cambridge University Press, 582 pp.

Ferro, C. A. T., 2007: A probability model for verifying deterministic forecasts of extreme events. *Wea. Forecasting,* **22,** 1089–1100.

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

Ghelli, A., and C. Primo, 2009: On the use of the extreme dependency score to investigate the performance of an NWP model for rare events. *Meteor. Appl.,* **16,** 537–544.

Göber, M., C. A. Wilson, S. F. Milton, and D. B. Stephenson, 2004: Fair play in the verification of operational quantitative precipitation forecasts. *J. Hydrol.,* **288,** 225–236.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.,* **132,** 2905–2923.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Hogan, R. J., E. J. O'Connor, and A. J. Illingworth, 2009: Verification of cloud-fraction forecasts. *Quart. J. Roy. Meteor. Soc.,* **135,** 1494–1511.

——, C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2010: Equitability revisited: Why the "equitable threat score" is not equitable. *Wea. Forecasting,* **25,** 710–726.

Hubálek, Z., 1982: Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biol. Rev. Cambridge Philos. Soc.,* **57,** 669–689.

Jolliffe, I. T., 2008: The impenetrable hedge: A note on proprety, equitability and consistency. *Meteor. Appl.,* **15,** 25–29.

Ledford, A. W., and J. A. Tawn, 1996: Statistics for near independence in multivariate extreme values. *Biometrika,* **83,** 169–187.

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

——, 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting,* **19,** 1106–1144.

Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

——, 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 37–76.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting,* **14,** 713–725.

——, and ——, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

Primo, C., and A. Ghelli, 2009: The affect of the base rate on the extreme dependency score. *Meteor. Appl.,* **16,** 533–535.

Ramos, A., and A. Ledford, 2009: A new class of models for bivariate joint tails. *J. Roy. Stat. Soc.,* **71B,** 219–241.

Segers, J., and B. Vandewalle, 2004: Statistics of multivariate extremes. *Statistics of Extremes: Theory and Applications,* J. Beirlant et al., Eds., John Wiley and Sons, 297–368.

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

——, B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.,* **15,** 41–50.

Swets, J. A., 1986: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychol. Bull.,* **99,** 181–198.

——, 1988: Measuring the accuracy of diagnostic systems. *Science,* **240,** 1285–1293.

——, 1996: *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Lawrence Erlbaum, 328 pp.

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

Yule, G. U., 1900: On the association of attributes in statistics: With illustrations from the material of the Childhood Society, &c. *Philos. Trans. Roy. Soc. London,* **194A,** 257–319.

——, 1912: On the methods of measuring association between two attributes. *J. Roy. Stat. Soc.,* **75,** 579–652.