Equitability revisited: why the "equitable threat score" is not equitable

Article

Published Version

Equitability Revisited: Why the "Equitable Threat Score" Is Not Equitable

ROBIN J. HOGAN

Department of Meteorology, University of Reading, Reading, United Kingdom

CHRISTOPHER A. T. FERRO, IAN T. JOLLIFFE, AND DAVID B. STEPHENSON

School of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom

(Manuscript received 25 August 2009, in final form 14 October 2009)

ABSTRACT

In the forecasting of binary events, verification measures that are "equitable" were defined by Gandin and Murphy to satisfy two requirements: 1) they award all random forecasting systems, including those that always issue the same forecast, the same expected score (typically zero), and 2) they are expressible as the linear weighted sum of the elements of the contingency table, where the weights are independent of the entries in the table, apart from the base rate. The authors demonstrate that the widely used "equitable threat score" (ETS), as well as numerous others, satisfies neither of these requirements and only satisfies the first requirement in the limit of an infinite sample size. Such measures are referred to as "asymptotically equitable." In the case of ETS, the expected score of a random forecasting system is always positive and only falls below 0.01 when the number of samples is greater than around 30. Two other asymptotically equitable measures are the odds ratio skill score and the symmetric extreme dependency score, which are more strongly inequitable than ETS, particularly for rare events; for example, when the base rate is 2% and the sample size is 1000, random but unbiased forecasting systems yield an expected score of around −0.5, reducing in magnitude to −0.01 or smaller only for sample sizes exceeding 25 000. This presents a problem since these nonlinear measures have other desirable properties, in particular being reliable indicators of skill for rare events (provided that the sample size is large enough). A potential way to reconcile these properties with equitability is to recognize that Gandin and Murphy's two requirements are independent, and the second can be safely discarded without losing the key advantages of equitability that are embodied in the first. This enables inequitable and asymptotically equitable measures to be scaled to make them equitable, while retaining their nonlinearity and other properties such as being reliable indicators of skill for rare events. It also opens up the possibility of designing new equitable verification measures.

1. Introduction: What is equitability and why is it desirable?

To assess objectively the skill of a sequence of n yes–no forecasts (e.g., whether or not a tornado will occur or whether a rain rate will exceed a certain threshold), one first defines the 2×2 contingency table (e.g., Mason 2003) containing the total number of correct forecasts of occurrence (or "hits") a, the number of incorrect forecasts of occurrence b, the number of incorrect forecasts of nonoccurrence c, and the number of correct forecasts of nonoccurrence d,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \text{hits} & \text{false alarms} \\ \text{misses} & \text{correct negatives} \end{pmatrix}, \quad (1)$$

such that $a + b + c + d = n$. For a particular set of weather conditions, the observed frequency of occurrence or "base rate" is fixed at $p = (a + c)/n$. A verification measure S is then defined as a function of the elements of the contingency table.

It has been known since the time of Finley (1884) that there are pitfalls to avoid in the design of the measure one uses to quantify the degree of skill (Murphy 1996). In particular, the measure should not encourage a forecaster to "hedge," that is, to issue a forecast that differs from his or her "true belief" or "judgment" in order to improve either the score that is being used to assess the forecast, or its expectation. This definition has been implicit or explicit in discussions of hedging since the

Corresponding author address: Dr. Robin Hogan, Dept. of Meteorology, Earley Gate, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: r.j.hogan@reading.ac.uk

DOI: 10.1175/2009WAF2222350.1

1950s (see references in Jolliffe 2008). However, it should be noted that hedging has also been used to denote a somewhat different pattern of behavior (Marzban 1998; Brill 2009), namely issuing a forecast that has a frequency bias, $B \neq 1$ [where bias $B = (a + b)/(a + c) = 1$], in order to improve the value of the score being used to assess the forecast, compared to its value when the forecast is unbiased ($B = 1$). The difference between the two may be illustrated simply as follows. The first definition of hedging includes the concept of ''hedging toward climatology,'' that is, changing one's forecast rate of occurrence to match the observed rate (because one's true belief has a bias). This is the sense of the term that was explored at length by Stephenson (2000). The second definition would not consider such behavior to be hedging. Indeed, if the original biased forecast was rewarded by a particular verification measure for its bias, then the second definition might judge hedging to be *not* to change one's forecast rate of occurrence to match the observed rate, since it would be being rewarded by issuing a biased forecast even though that was the forecaster's true belief. We stick to the older and more intuitive definition of hedging.

The concept of hedging is most often considered in the case of probabilistic forecasts, but for deterministic forecasts of binary events it is a less useful concept. Jolliffe (2008) argued that it is difficult to conceive of a situation where a forecaster's true belief is not probabilistic, and therefore the fact that the forecaster only predicts occurrence or nonoccurrence (with apparently 100% confidence) means that all forecasts differ from his or her belief, and so are hedged in some sense. Murphy and Daan (1985) recognized this difficulty and defined the weaker property ''consistency'': if a probabilistic belief is converted into a deterministic forecast via some ''directive'' (e.g., forecast occurrence only if you believe the event has greater than a 50% probability of occurring), then a consistent verification measure is one that maximizes the score for a particular directive. We are then left with the problem of determining whether a directive is sensible.

An alternative desirable property, which is more amenable to analysis for binary forecasts, is that the measure assesses the performance of a forecaster or forecasting system relative to a random forecasting system. This is known as equitability and was defined rigorously by Gandin and Murphy (1992) in their seminal paper on the subject as follows:

1) An equitable verification measure awards all random forecasting systems, including those that always forecast the same value, the same expected score (denoted S_0 in this paper).
2) An equitable verification measure S must be expressible as the linear weighted sum of the elements of the contingency table; that is, $S = (S_a a + S_b b + S_c c + S_d d)/n + S_0$, where the weights $S_a, \ldots, S_d$ are independent of the individual elements of the contingency table, although they may depend on the base rate $p = (a + c)/n$. Gandin and Murphy (1992) expressed these weights in the form of a 2×2 ''scoring matrix.''

From these two requirements, Gandin and Murphy (1992) then proceeded to show that only the Peirce skill score (Peirce 1884), or a linear function of it, is truly equitable. However, it has since been claimed that several other measures are equitable, such as the Heidke skill score (Mason 2003), the odds ratio skill score (Stephenson 2000), and indeed the equitable threat score (ETS; Mason 2003). So what is the reason for this discrepancy? It can be shown from their definitions that these three measures do not satisfy requirement 2, so a possibility is that many authors equate equitability only with requirement 1, although we will show in the next section that two of these measures do not, in general, satisfy this requirement either. Before we discuss some of the previous work where apparently different definitions of equitability have been used, we step back and address a more fundamental question: why is equitability (as defined by Gandin and Murphy) desirable?

We start by considering requirement 1 and note that many people question the importance of assigning a constant baseline score to all random and constant forecast systems: ''if a measure can correctly *rank* the performance of different forecasting systems, then who cares about equitability?'' The answer to this is that it is easy to show that a measure that does not satisfy requirement 1 will sometimes incorrectly rank a random forecast above a forecast system with some skill. Consider two different random forecasters (e.g., one who forecasts occurrence with probability 0.1 and the other with probability 0.9), who are awarded different expected scores of S_1 and S_2, with $S_1 < S_2$. If a real forecaster predicts occurrence in the same proportion of his or her forecasts as the first random forecaster, but with positive skill such that the expected score awarded S is greater than S_1, then if S happens to be less than S_2, this forecaster will be incorrectly ranked below the second random forecaster. Moreover, such a measure would *reward* this forecaster if he or she were to abandon the physical principles on which forecasts were being made and instead issue random forecasts with a probability optimized for the measure in question. These properties are obviously undesirable and support the requirement for all random and constant forecasts to score the same. More generally, to what extent is requirement 1 related

to the inability to hedge? Given the arguments of Jolliffe (2008), we consider only the limited form of hedging described by Stephenson (2000): to randomly select a particular proportion of forecasts of nonoccurrence and change them to occurrence, or vice versa, in an attempt to improve one's score. By this definition, requirement 1 clearly prevents *random* forecasting systems from being able to hedge their forecasts. However, it does not guarantee that the same is true for the much wider class of forecasting systems in practice that do have some skill. At most, we can state that, because an equitable measure cannot be hedged by a random forecasting system, it *may* be less easy to hedge by a forecasting system with positive skill compared to an inequitable measure, but this is not guaranteed. As a point of clarity, Gandin and Murphy (1992) specified that the baseline "no skill" score, S_0, should be constant for a particular verification measure (and indeed $S_0 = 0$ for most measures). In principle, S_0 could be a function of p, a quantity over which the forecaster has no control, and a random forecaster would still be unable to hedge the measure. However, an additional justification for requirement 1 is that constant S_0 provides a single baseline above which a forecaster's performance can be said to be superior to the expected performance of a class of naïve forecasters, such as random forecasters. This makes subsequent interpretation of the measure easier, especially if we also have a fixed upper limit corresponding to a perfect forecast. We note that S_0 should not be a function of n, since often forecasters do have control over how large a sample they are assessed with and, therefore, would have the opportunity to artificially inflate their score.

Gandin and Murphy's (1992) second requirement given above is rather less easy to justify as an essential companion to their first. Linearity in general was deemed desirable by Hogan et al. (2009) since it ensures that a measure is equally sensitive to changes in skill throughout its range; that is, "near perfect" and "near random" forecasting systems will be rewarded by the same increase in score if 1 is added to a and d and 1 is subtracted from b and c. This was demonstrated when they estimated the "half life" of numerical weather forecasts, that is, a time scale for the decay of the score toward S_0 as a function of the lead time into the forecast; in particular, the highly nonlinear "odds ratio skill score" yielded a misleadingly high half life. Conversely, linearity appears to be incompatible with another desirable property, that of being a reliable indicator of skill for rare events. It was shown by Stephenson et al. (2008) that almost all measures (and certainly all linear ones) are "degenerate" in that they tend to zero or some other constant as the base rate tends to zero. Their extreme dependency score and its symmetric counterpart proposed by Hogan et al. (2009) overcome this problem via a nonlinear (specifically a logarithmic) dependence on the elements of the contingency table, preventing these measures from satisfying requirement 2. So what was the reason for inclusion of requirement 2 by Gandin and Murphy (1992)? It is possible that it was included simply to restrict the class of measures considered in order that the mathematical condition for a measure to satisfy requirement 1 could be derived, rather than having any overriding merit of its own. Indeed, another interpretation of Gandin and Murphy (1992) is not that requirement 2 is fundamental to the concept of equitability, but simply that they chose to consider only the equitability (embodied in requirement 1) of a limited class of measures (those that also satisfy requirement 2). As will be demonstrated later in this paper, requirement 2 is certainly not a necessary condition in order for requirement 1 to be satisfied.

There have been several examples in the literature where the term equitability has been used in a different sense than that of Gandin and Murphy, although in each case Gandin and Murphy (1992) was provided as the reference. Marzban (1998) equated "inequitability" with being able to hedge a measure, specifically that an inequitable measure is one that is optimized by forecasting with a frequency bias other than $B = 1$ and therefore induces under- or overforecasting. By this definition, *all* measures for verifying binary forecasts that he considered were found to be inequitable, including the Peirce skill score (also known as the true skill score), which Gandin and Murphy (1992) deemed to be equitable. Baldwin and Kain (2006) also adopted Marzban's concept of equitability. Marzban and Lakshmanan (1999) considered measures that satisfied requirement 2 but argued that requirement 1 could be relaxed in the definition of equitability, on the basis that it can be difficult to reconcile with the requirement that a measure is optimized by an unbiased forecast. The findings of Marzban (1998) and Marzban and Lakshmanan (1999) are simply a consequence of them taking the fundamental concept of equitability to be something different to Gandin and Murphy (1992). This is not to claim that the work of these authors is invalid, merely to suggest that the property they were considering to be paramount (that a measure is optimized by an unbiased forecast) should be referred to as something other than equitability.

We argue that requirement 1 is fundamental to the concept of equitability as envisioned by Gandin and Murphy (1992), and its desirability is amply justified by the arguments we have provided above. By contrast, we argue that requirement 2 can be safely discarded in the definition of equitability, and in this paper we explore the very interesting class of measures that are now classified as equitable. In section 2, we derive the necessary

conditions for requirement 1 to be satisfied and explore the nature of the class of measures that we deem to be equitable but that do not satisfy Gandin and Murphy's requirement 2. In section 3 we demonstrate the nonequitability of measures such as the so-called equitable threat score using examples with a sample size small enough that all the possible contingency tables can be written out explicitly. In section 4 we introduce the concept of ''asymptotic equitability,'' whereby a number of measures are shown to tend to equitability only in the limit of an infinite sample size. Then, in section 5, we demonstrate how an interesting class of nonlinear but nonetheless truly equitable measures can be constructed, for example, by rescaling inequitable measures.

2. How to determine whether a measure is equitable

a. The expectation over all possible contingency tables

In this section we show how, from requirement 1 in the previous section, we may derive an expression that all truly equitable measures must satisfy, and how by omitting requirement 2 a new class of equitable measures is admitted. We define a random forecasting system as one that issues forecasts randomly with a fixed probability, q_p, irrespective of any other information; q_p is the ''population'' forecast rate of occurrence (to contrast with the ''sample'' forecast rate of occurrence), and may be any value in the range 0 to 1, inclusive. To calculate whether this forecasting system would yield an expected score of zero, or some other constant value S_0, we need to consider all possible sequences of n forecasts that could be made, the score that they would each be awarded by the verification measure being tested, and the probability of each sequence being issued by chance. We consider both n and the base rate $p = (a + c)n$ to be held fixed in this exercise, and are therefore calculating the expectation given a particular sequence of events in reality. At the end of this section we demonstrate that if we wish to calculate the expectation over all possible base rates (e.g., if p is treated as a *sample* base rate that is just a realization of the *population* base rate), we arrive at the same conclusions as to which measures are equitable.

It is illuminating to consider specific examples of a small sample size n in which all the possible contingency tables can be written out explicitly. Figure 1 (see also Fig. 3) shows two such examples, one with $n = 4$ and $p = ½$, and the other with $n = 3$ and $p = ⅓$. A particular random forecasting system may predict the occurrence of the event between 0 and n times, with the sample forecast rate of occurrence given by $q_s = (a + b)/n$. This variable is used as the abscissa of the figures, and for each value of q_s there may be several possible contingency tables, which are shown by the vertical columns with the contingency table that corresponds to the best performance at the top. It is obvious from Figs. 1 and 3 that the verification problem for binary forecasts is inherently two-dimensional, with two numbers needed to characterize the performance of a particular set of forecasts uniquely. As plotted here, the abscissa is directly related to the *bias* (e.g., the frequency bias $B = q_s/p$), while the ordinate is directly related to the *skill*. Although this paper is primarily concerned with the best measures for characterizing the skill, it should be borne in mind that whatever measure is recommended for skill does not on its own give a complete picture of the performance of the forecast system, but should be reported alongside a measure characterizing the bias. For reasons that will become obvious in section 2c, the measure of the apparent skill against which we choose to plot the contingency tables is the Peirce skill score, defined as

$$\mathrm{PSS} = \frac{a}{a+c} - \frac{b}{b+d}. \tag{2}$$

(Note that this measure has also previously been referred to as the true skill statistic, the true skill score, and the Hanssen–Kuipers performance index.)

If a verification measure $S(a, b, c, d)$ is defined in terms of the elements of the contingency table, then its expected value given a particular base rate, $E(S|p)$, is calculated by summing over the m possible contingency tables, each weighted by their probability of occurring, $P(a, b, c, d|p)$:

$$E(S|p) = \sum_{i=1}^{m} S(a_i, b_i, c_i, d_i)P(a_i, b_i, c_i, d_i|p). \tag{3}$$

A measure is equitable by requirement 1 in the introduction if, for a random forecasting system, $E(S|p) = S_0$ for all $p \in \{0, 1/n, \ldots, 1\}$ and all $q_s \in [0, 1]$, and where we take S_0 to be constant for a particular score (it cannot vary with p or q_p, for instance). Since n and p are fixed, the four degrees of freedom in the contingency table reduce to two (as seen by the fact that the possible contingency tables in Fig. 1 occupy a plane), for example, described purely by a and b. For random forecasts, these may be treated as random, binomially distributed variables: $a|p \sim \mathrm{Bin}(np, q_p)$ and $b|p \sim \mathrm{Bin}(n - np, q_p)$. Since these variables are also independent, the probability of a contingency table occurring randomly may be written as $P(a, b, c, d|p) = P(a|p)P(b|p)$, where $c = np - a$, $d = n(1 - p) - b$, $0 \le a \le np$, and $0 \le b \le n(1 - p)$.

An alternative way of splitting up this probability is $P(a, b, c, d|p) = P(a, b, c, d|p, q_s)P(q_s|p)$, where $P(q_s|p)$ is the probability of a particular sample forecast rate of occurrence (i.e., the probability of being in a particular

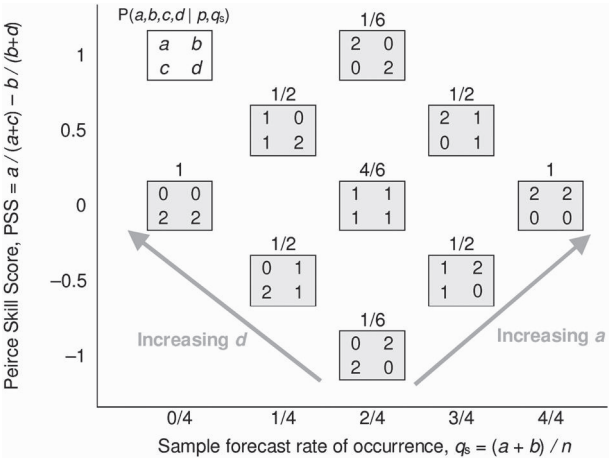

FIG. 1. The possible contingency tables for the number of forecasts $n = 4$ and base rate $p = ½$, as a function of the sample forecast rate of occurrence q_s and the Peirce skill score. The elements of the contingency tables are a–d as shown in the white table to the top left. The numbers above each box give $P(a, b, c, d|p, q_s)$, the probability of that table occurring randomly given one has a particular base rate and sample forecast rate of occurrence q_s. Therefore, these conditional probabilities sum to one in each column.

column in Fig. 1) and $P(a, b, c, d|p, q_s)$ is the probability of a particular table occurring given that we are in a certain column of Fig. 1. The advantage of this approach is that we may write the expected score over all possible tables as the sum over the expected scores given each particular value of q_s:

$$E(S|p) = \sum_{q_s} E(S|p, q_s)P(q_s|p), \tag{4}$$

where $E(S|p, q_s) = \sum_i S(a_i, b_i, c_i, d_i)P(a_i, b_i, c_i, d_i|p, q_s)$.

The probabilities may be calculated as follows. Again, assuming the forecast system to have a population forecast rate of occurrence of q_p, the number of events that are actually forecast in a particular sample, nq_s, follows the binomial distribution: $nq_s|p \sim \text{Bin}(n, q_p)$. Therefore, the probability of a particular q_s is

$$P(q_s|p) = C(n, nq_s)q_p^{nq_s}(1 - q_p)^{n(1-q_s)}, \tag{5}$$

where $C(n, k) = n!/[k!(n - k)!]$ is the binomial coefficient, expressing the number of ways k events can occur in n trials. The random conditional probability $P(a, b, c, d|p, q_s)$ is now a function of just one random variable (e.g., a) and may be written as

$$P(a, b, c, d|p, q_s) \equiv P(a|p, q_s) = \frac{C(np, a)C(n - np, nq_s - a)}{C(n, nq_s)}, \tag{6}$$

where $b = nq_s - a$, $c = np - a$, $d = n(1 - p - q_s) + a$, and $\max\{0, n(p + q_s - 1)\} \le a \le \min\{np, nq_s\}$. The denominator on the right-hand side of (6) is the total number of ways of selecting nq_s cases from n, and the numerator is the number of those ways in which we have a hits, since then we must select a from the np cases in which the event occurs and $nq_s - a$ from the $n - np$ cases in which the event does not occur.

To illustrate these probabilities being applied to a small sample of forecasts, the numbers above each contingency table in Figs. 1 and 3 indicate the probability of it occurring randomly given that one is in a particular column, $P(a, b, c, d|p, q_s)$. Figures 2 and 4 depict the scores that would be awarded for each contingency table in Figs. 1 and 3 for a range of different verification measures, and beneath each column is shown the random expected score for that column, $E(S|p, q_s)$. A measure is deemed equitable if $E(S|p) = S_0$, and from (4) it can be seen that the easiest way for this to occur is if $E(S|p, q_s) = S_0$ for all q_s. In principle, one could have $\text{E}(S|p, q_s) \neq S_0$, but achieve cancellation between the differences from S_0 in each column such that $\text{E}(S|p) = S_0$ (as in Fig. 2e). In practice, this is not possible for all values of q_p, since each possible q_p leads to a different weighting between the columns. Take the example in Fig. 1: for $q_p = ½$, the probabilities of obtaining each possible value of q_s are $P(q_s = \{0/4, 1/4, 2/4, 3/4, 4/4\}) = \{1/16, 4/16, 6/16, 4/16, 1/16\}$, but for $q_p = 1/4$, the probabilities become $\{0.316, 0.422, 0.211, 0.047, 0.004\}$. Therefore, our definition of equitability requires simply that

$$E(S|p, q_s) = S_0 \quad \text{for all } p \text{ and } q_s. \tag{7}$$

We prove this result formally in the appendix, and make use of it later in the paper. It also ensures equitability for the possible (although improbable) forecaster who fixes q_s by declaring before the first forecast "in the 10 forecasts that will be verified I will predict occurrence exactly 5 times."

Finally in this section, we consider the consequence of requiring that the "expected score" in the definition of equitability involves calculating the expectation for the fixed "population base rate" p_p, over all the possible sample base rates, which we temporarily denote p_s. Now the expected score that we had previously written as $E(S|p)$ is considered to be for a given sample base rate and so is denoted $E(S|p_s)$. Following the reasoning that led to (4), the expected score considering all possible base rates is given by

$$E(S) = \sum_{p_s} E(S|p_s)P(p_s), \tag{8}$$

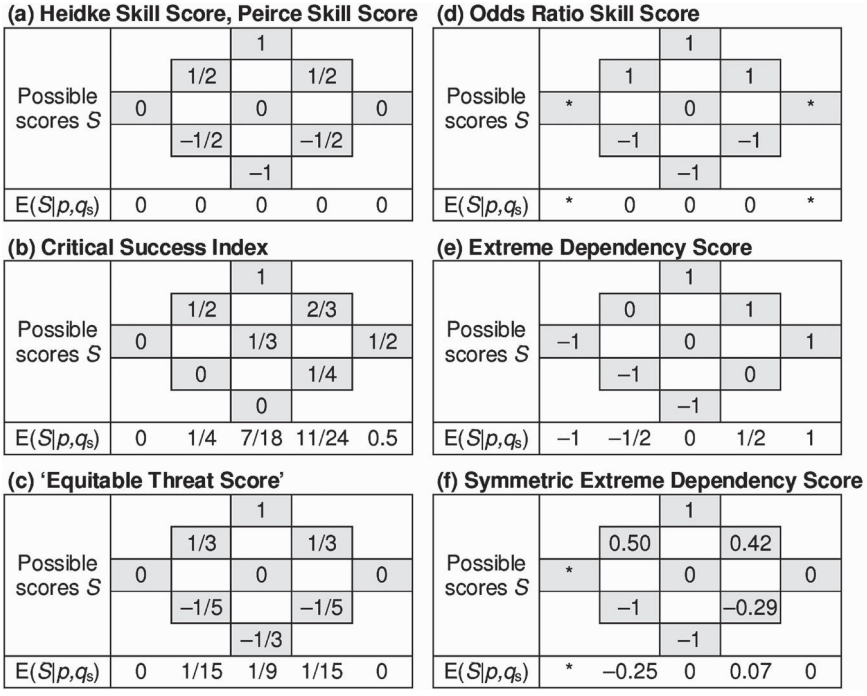

FIG. 2. Values of the scores for each of the contingency tables given in Fig. 1, together with the expected score, $E(S|p, q_s)$, for each value of the sample forecast rate of occurrence q_s at the bottom of each column. The asterisks in (d) and (f) indicate that a value is undefined due to zero divided by zero or infinity divided by infinity.

where $P(p_s)$ is the probability of a particular sample base rate. We may if we choose assume that the number of events that occurred follows a binomial distribution $np_s \sim \text{Bin}(n, p_p)$, as for nq_s, but the essential result is that if $\text{E}(S|p_s) = S_0$ for all p_s, then $E(S) = S_0$. Hence, if a measure is equitable for one particular sequence of occurrences in reality, then it will also be equitable when calculating the expectation over all possible ''realizations'' of reality.

b. Gandin and Murphy's condition for equitability

The previous subsection established the important result that a binary verification measure is equitable if and only if $E(S|p, q_s) = S_0$ for all p and q_s. So how can we apply this rule to test the equitability of particular measures without needing to consider a specific example such as in Fig. 1? We start this discussion by considering the approach of Gandin and Murphy (1992). They considered only measures S that may be written as the linear weighted sum of the terms a–d,

$$S = \frac{(S_a a + S_b b + S_c c + S_d d)}{n}, \tag{9}$$

and then proceeded to calculate the relationships between the weights S_i that are necessary for the measure to be equitable. Requirement 2 in the introduction includes the possibility of adding an offset S_0 to (9), but following Gandin and Murphy (1992), if we can prove a measure with $S_0 = 0$ to be equitable, then adding a nonzero offset S_0 to it will also yield an equitable measure. Gandin and Murphy made one further assumption (also part of requirement 2), which was that the weights can depend only on the base rate p. To impose the condition that constant forecasts of occurrence ($q_s = 1$) yield a 0 expected score requires that $E(S|p, q_s) = 0$ when $a = np$, $b = n(1 - p)$, $c = 0$, and $d = 0$. Thus,

$$E(S|p,\, q_s = 1) = pS_a + (1 - p)S_b = 0. \tag{10}$$

Likewise, to require that constant forecasts of nonoccurrence ($q_s = 0$) also yield zero leads to

$$E(S|p,\, q_s = 0) = pS_c + (1 - p)S_d = 0. \tag{11}$$

The third condition sets the *scale* of the measure; for example, to make all perfect forecasts score unity, we write

$$S(np,\, 0,\, 0,\, n - np) = pS_a + (1 - p)S_d = 1. \tag{12}$$

Gandin and Murphy (1992) did not take this last step, but chose instead to impose the scale by explicitly setting

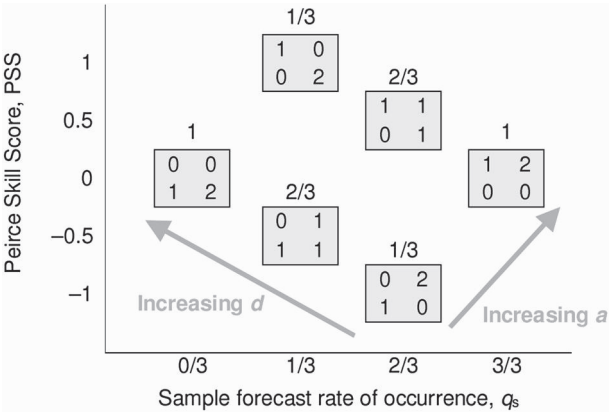

FIG. 3. As in Fig. 1, but for the number of forecasts $n = 3$ and base rate $p = 1/3$.

$S_b = S_c = -1$, thereby also making the explicit assumption that misses and false alarms are weighted equally. As shown by Manzato (2005), the assumption of equal weighting between misses and false alarms has no effect on the final result, and so it is clearer not to make it at all. We prefer to set the scale of the measure via (12). Thus, we have three equations [(10)–(12)] and four unknowns $(S_a, \ldots, S_d)$, seemingly an underconstrained system. However, the four unknowns are not independent of one another because we know that $a + c = np$ and $b + d = n(1 - p)$; so, therefore, we can relate the elements of the contingency table via $b + d = (a + c) \times (1 - p)/p$. This means that we can shuffle "mass" between the four weights while keeping the resulting measure unchanged. For example, suppose all four weights were nonzero and we wanted to eliminate S_d. From this relationship between the elements of the contingency table we may replace $S_d d$ in (9) with $S_d[-b + (a + c) \times (1 - p)/p]$, which is equivalent to defining a new set of weights given by the following primed values:

$$S'_a = S_a + \frac{S_d(1-p)}{p}; \quad S'_b = S_b - S_d;$$
$$S'_c = S_c + \frac{S_d(1-p)}{p}; \quad S'_d = 0. \tag{13}$$

Gandin and Murphy (1992) used (10) and (11), together with their requirement on S_b and S_c, to show that one particular equitable measure is defined by the weights $S_a = (1 - p)/p$, $S_b = S_c = -1$, and $S_d = p/(1 - p)$. This also satisfies our scale condition given by (12). This may

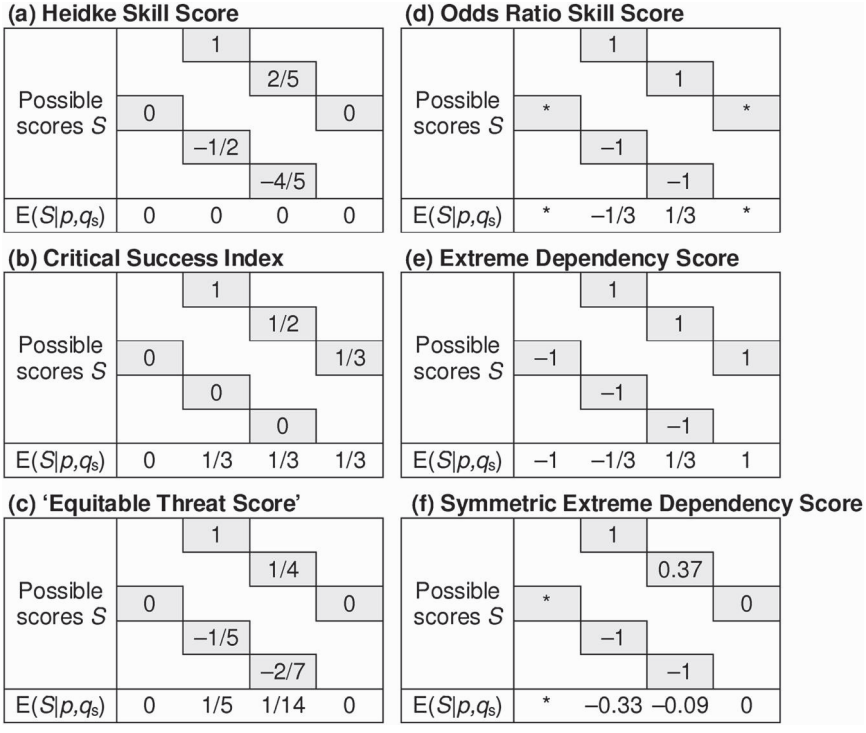

FIG. 4. Values of the scores for each of the contingency tables given in Fig. 3, together with the expected score for each value of the sample forecast rate of occurrence at the bottom of each column, $E(S|p, q_s)$. Measures are only truly equitable if they have $E(S|p, q_s) = 0$ in each column. Note that unlike in Fig. 2a, (a) corresponds only to the HSS and not the PSS, which is different in the case of $q_s = 2/3$. The asterisks in (d) and (f) indicate that a value is undefined.

be simplified by eliminating S_d using the operations given in (13) to yield $S_a = 1/p$, $S_b = -1/(1 - p)$, and $S_c = S_d = 0$, which is the same as the Peirce skill score (PSS) defined by (2). [It should also be noted that p in Gandin and Murphy's derivation is the sample base rate, but subsequent practice, at least for 3×3 tables, has been to use a longer-term "population" base rate in defining equitability, e.g., Livezey (2003).] Thus, the only degree of freedom in this derivation is the scale of the measure shown by the right-hand side of (12).

We may conclude that *the Peirce skill score is the only measure with scale 1 and $S_0 = 0$ that satisfies Gandin and Murphy's definition of equitability.* Recognizing that the scale and the offset are not fixed by their requirements of equitability, "Gandin–Murphy equitable measures" must have the form

$$S_{\mathrm{GM}} = f(p)\mathrm{PSS} + S_0, \tag{14}$$

where $f(p)$ may be any positive function. Positivity is necessary to ensure that forecasts better than random are awarded a score greater than the expected random score S_0. In principle, a measure could have f depending also on sample size n and still strictly satisfy the definition of equitability given in the introduction, but it is difficult to see when this would ever be an advantage as it would mean that multiplying all elements of the contingency table by a constant factor would change the score. Gandin and Murphy (1992) actually stated that only PSS and *monotonic transformations* of it satisfy the definition of equitability set forth in their paper. One presumes that by "monotonic" they actually meant "linear," since nonlinear transformations would violate the linearity expressed in (9) and requirement 2. Moreover, it is important to note that the scaling factor in the linear transformation may depend on p, as shown in (14), but this was not clear from Gandin and Murphy (1992). A simple example of a measure other than PSS that satisfies (14) is the one of Gringorten (1967), which has $f = 1$ and $S_0 = 1$.

c. *A general condition for linear, equitable measures*

So where does this leave measures that cannot be written in the form of (14), yet that appear to be equitable by requirement 1 in the introduction? An example is the Heidke skill score (HSS; Heidke 1926), which is written in two alternative ways (e.g., Hogan et al. 2009):

$$\begin{aligned}\mathrm{HSS} &= \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)} \\ &= \frac{a + d - E(a|p, q_s) - E(d|p, q_s)}{n - E(a|p, q_s) - E(d|p, q_s)},\end{aligned} \tag{15}$$

where $E(a|p, q_s) = npq_s$ and $E(d|p, q_s) = n(1 - p)(1 - q_s)$ are the expected values of a and d for a random forecast with a particular sample forecast rate of occurrence q_s. For the two examples in Figs. 2a and 4a, HSS does appear to be equitable in that the expected score for each column, $E(S|p, q_s)$, is 0. By rearrangement, it is possible to express HSS in a form similar to (14), with $S_0 = 0$ and $f = f_{\mathrm{HSS}}$, where

$$f_{\mathrm{HSS}} = \frac{2p(1 - p)}{(p + q_s - 2pq_s)}, \tag{16}$$

that is, with f a function of both p and q_s. It turns out that $f_{\mathrm{HSS}} = 1$ (and hence HSS = PSS) both when $p = ½$ and when $p = q_s$. This is confirmed in Fig. 2a: $p = ½$ leads to HSS = PSS for all contingency tables. In Fig. 4a, where $p = ⅓$, it can be seen that HSS = PSS for the column corresponding to $q_s = ⅓$, but in the column corresponding to $q_s = ⅔$, we find HSS = 0.8 PSS.

The Heidke skill score may also be expressed in the form of (9), but only if the weights are allowed to vary with both p and q_s as follows:

$$S_a = S_d = 1 \quad \text{and}$$
$$S_b = S_c = -\frac{[pq_s + (1 - p)(1 - q_s)]}{[1 - pq_s - (1 - p)(1 - q_s)]}.$$

Allowing these weights, and in turn the value of f, to vary with q_s violates Gandin and Murphy's requirement 2 in the introduction, but HSS does satisfy the more important requirement 1. To prove this, consider a single column of contingency tables in Fig. 1, for which we know that the expected value of PSS for a random forecast with a given value of q_s is zero; that is, $E(\mathrm{PSS}|p, q_s) = 0$. Scaling PSS by an arbitrary factor f still yields the same result, $\mathrm{E}(f\ \mathrm{PSS}|p, q_s) = 0$, and the resulting measure is still equitable by the definition in (7). The only requirement on f is that it is constant within a column, which means that it may vary with both p and q_s, but may not depend on any individual elements of the contingency table. Thus, by our more permissive definition of equitability, the condition for a verification measure to be equitable *and linear* is that it can be written in the form

$$S_{\mathrm{equitable,\,linear}} = f(p, q_s)\mathrm{PSS} + S_0, \tag{17}$$

where $f(p, q_s)$ is any positive function of p and q_s. By linear we mean simply that for given p and q_s, the measure varies linearly with every individual member of the contingency table; that is, the measure can be written in the form of (9) with the weights S_i being functions of only p and q_s. Requirement 2 of Gandin and Murphy (1992) was a little more restrictive, requiring

linear variation with the elements of the contingency table just for a given p (and therefore excluding HSS). The merits or otherwise of allowing f to depend on q_s deserve a little discussion. On one hand, it allows measures such as HSS to be *transpose symmetric* (Stephenson 2000), that is, to be invariant if the observations and the forecasts are swapped. On the other hand, if f increased strongly with q_s, say, then a forecaster with some skill may be able to increase his or her expected score by randomly changing some forecasts of nonoccurrence to occurrence (although this is not true of HSS). However, these considerations are independent of the property of equitability encapsulated in requirement 1. In the next two sections we demonstrate that it is the nonlinearity of a number of measures that prevents them from being equitable, and in section 5 it is shown how they may be transformed into equitable measures while retaining their nonlinearity.

3. Measures that are inequitable for small sample sizes

There are a large number of verification measures in the literature that cannot be written in the form of (17), and in this section we demonstrate that they are inequitable by writing out the scores explicitly for the examples in Figs. 1 and 3.

a. *The critical success index and the equitable threat score*

Motivated by Finley's paper, Gilbert (1884) proposed two verification measures, the first of which is now most commonly referred to as the critical success index (Donaldson et al. 1975), given by

$$\mathrm{CSI} = \frac{a}{(a+b+c)}, \tag{18}$$

although it is also sometimes called the threat score. As recognized by Gilbert (1884) and reiterated by Mason (1989) and Schaefer (1990), this measure has the disadvantage that equally unskillful forecasting systems, for example, those that always predict occurrence and those that always predict nonoccurrence, yield a different score, and therefore in modern terminology it is definitely inequitable. This is revealed clearly in Fig. 2b, where we can see also that the expected scores for random forecasts $E(S|p, q_s)$ increase steadily with q_s.

Gilbert proposed an alternative measure in which the expected number of hits $[E(a|p, q_s) = npq_s = (a + b)(a + c)/n]$ obtained by a random forecasting system with the same forecast rate of occurrence (q_s) as the actual forecasting system is subtracted from both the numerator and denominator. This measure is now most commonly referred to as the equitable threat score, given by

$$\mathrm{ETS} = \frac{[a - E(a|p, q_s)]}{[a - E(a|p, q_s) + b + c]}. \tag{19}$$

The presence of a, b, and c on the denominator means that this measure cannot be rewritten in the form of (13) and therefore is inequitable by Gandin and Murphy's requirement 2 in the introduction. Figure 2c shows that ETS is also inequitable by requirement 1; although it is 0 for constant forecasts of occurrence or nonoccurrence, its expected value is positive for random forecasting systems with $0 < q_s < 1$. This pattern of behavior appears to originate from its nonlinear dependence on the elements of the contingency table; in a given column of Fig. 1, as one progresses up through the contingency tables, the elements *a–d* each change by only one from one table to the next. Yet, ETS changes by an increasing amount toward the more positive scores. Hence, there is incomplete cancellation between the positive and negative scores, leading to a positive expected value. It turns out that ETS is monotonically but nonlinearly related to the truly equitable Heidke skill score via ETS = HSS/(2 – HSS) (Doswell et al. 1990).

So what is the origin of the term equitable threat score? Schaefer (1990) analyzed it and its relationship to CSI, but called it the Gilbert score. Mason (2003) attributed the name ETS to both Schaefer (1990) and Doswell et al. (1990), but neither actually used the term. The term appears to have been first used in the literature by Mesinger and Black (1992), who cited Gandin and Murphy's (1992) statement that ''many skill scores used to evaluate forecasts of discrete variables are inequitable, in the sense that constant forecasts of some events lead to better scores than constant forecasts of other events.'' Since the modified version of the threat score considered by Schaefer (1990) did not have this deficiency, Mesinger and Black felt justified in referring to it as the ''equitable'' threat score, without noting that it did not satisfy the other aspect of Gandin and Murphy's requirement 1 for a measure to be equitable, that all random forecasts must to receive the same expected score as all constant forecasts. The name has stuck and ETS is now one of the most widely used verification measures in meteorology. One of the anonymous reviewers informed the authors that the term equitable threat score was introduced at the then U.S. National Meteorological Center by Mesinger in 1991 following a seminar by Gandin, but before the Gandin and Murphy (1992) paper had been published and their full definition was available.

Although strictly inequitable, ETS does have the property that if we calculate the expected values of each

of the elements of the contingency table over all possible random forecasts with the same p and q_s, and apply the measure to them, then we obtain

$$\mathrm{ETS}[E(a|p, q_s), E(b|p, q_s), E(c|p, q_s), E(d|p, q_s)] = 0.$$

This is indicated by the zeros in the middle row of Figs. 2c and 4c (where PSS is also zero). It appears to be a widespread misconception that if a measure yields 0 when applied to the expected values of the elements of the contingency table, then it will also have an expected value of zero over all realizations of a random forecast. This is only valid if a measure is linear [i.e., satisfies (17)]. Indeed, Gandin and Murphy (1992) took this shortcut for the linear measures they considered. The extent to which ETS is inequitable for larger samples will be examined in section 4, along with several other measures.

b. Other inequitable measures

A number of verification measures have been advocated more recently that have desirable properties yet also cannot be written in the form given by (17). Stephenson (2000) proposed the odds ratio skill score (sometimes referred to as Yule's Q; Yule 1900), defined as

$$\mathrm{ORSS} = \frac{(\mathrm{OR} - 1)}{(\mathrm{OR} + 1)},$$

where the odds ratio is defined as $\mathrm{OR} = ad/(bc)$. It can be seen in Figs. 2d and 4d that ORSS is inequitable except for the special case of $p = ½$, although Stephenson (2000) and Mason (2003) stated it to be equitable. (Note that even for $p = ½$ we have to neglect the case of q_s equal to 0 or 1, as this leads to ORSS being undefined, although it would be a simple matter to *define* ORSS to be zero for these values of q_s.) As with ETS, ORSS does award the expected random contingency table a score of zero, but its nonlinearity means that this does not lead to its expected value over all possible random forecasts being zero. Stephenson also proposed the log of odds ratio, defined as $\mathrm{LOR} = \ln(\mathrm{OR})$, which is rather more difficult to judge for small sample sizes due to it taking the value $\pm\infty$ when any of the elements of the contingency table are 0. In practice, LOR is a more reliable indicator of skill for rare events than, for example, the Heidke skill score, which tends to the meaningless limit of 0 (e.g., Hogan et al. 2009).

Stephenson et al. (2008) proposed the extreme dependency score,

$$\mathrm{EDS} = \ln(p^2)/\ln(a/n) - 1, \tag{20}$$

specifically for verification of rare events; they showed that for unbiased samples it is a reliable indicator of skill as $p \to 0$. However, Figs. 2e and 4e show that this measure is inequitable and, moreover, random forecasts that overpredict occurrence (i.e., those with higher q_s) are rewarded by a higher expected score (see also Primo and Ghelli 2009). This property is shared by CSI. Because of this, Stephenson et al. (2008) specified that EDS should always be used with calibrated (zero bias) forecasts. Hogan et al. (2009) proposed the symmetric extreme dependency score, defined as

$$\mathrm{SEDS} = \ln(pq_s)/\ln(a/n) - 1, \tag{21}$$

which can be applied to uncalibrated forecasts while retaining the same desirable properties for verifying rare events. Figures 2f and 4f demonstrate that it is still inequitable for these small samples, although this time the expected random contingency table for a given q_s (i.e., the middle rows of Figs. 1 and 3, corresponding to PSS = 0) always receives a score of 0. The exception is the case of $q_s = 0$, which results in SEDS being undefined due to the logarithm of 0 appearing in the numerator and denominator.

None of the example cases in Figs. 1–4 are equitable for all values of base rate without satisfying the criterion for linear equitability embodied in (17). The *apparent* need for linearity may be illustrated by considering what happens when we take a nonlinear transformation of the equitable Heidke skill score that is symmetric about a score of zero, for example, HSS cubed (HSS^3). In the case of $p = ½$, it can be seen on inspection of Fig. 2a that the numbers awarded for each contingency table would be changed, but there would still be cancellation between positive and negative scores when calculating the expected score. However, in the case of $p < ½$, it can be seen from Fig. 4a that there would no longer be direct cancellation, and the expected score would be nonzero in general. This yields an apparent conflict, since it is the very nonlinearity of some measures that imparts particular desirable properties. For example, the presence of logarithms in the definition of the log of odds ratio and the extreme dependency score (and its symmetric equivalent) is what makes them still convey information on the intrinsic skill of a forecast for low base rates, when the truly equitable HSS and PSS tend to a meaningless value of 0 (Stephenson 2000; Stephenson et al. 2008; Hogan et al. 2009). A solution to this dilemma will be presented in section 5.

4. The concept of asymptotic equitability

a. Numerical examples for increasing sample size

The probability that a random forecasting system produces a contingency table very different from the expected

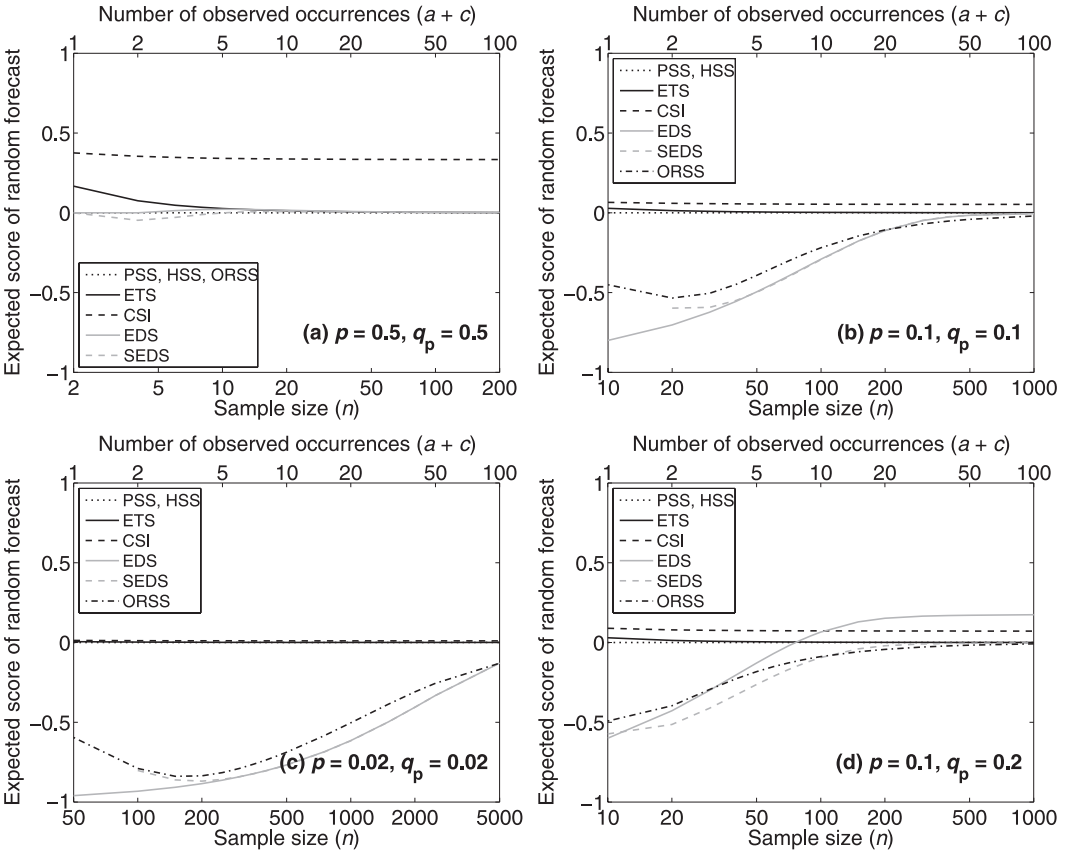

FIG. 5. The expected values of a number of verification measures $E(S|p)$ vs sample size, for random forecasting systems with base rate p and population forecast probability of occurrence q_p of (a) $p = q_p = 0.5$, (b) $p = q_p = 0.1$, (c) $p = q_p = 0.02$, and (d) $p = 0.1$ and $q_p = 0.2$. The undefined values are removed from consideration when calculating SEDS, but when these occupy more than 25% of the probability space, SEDS is not plotted; this only affects (b) and (c).

contingency table decreases as n increases. Furthermore, the expected value of a function of the contingency table will then tend toward the same function applied to the expected contingency table as n increases. These arguments can be made precise by the weak law of large numbers and the continuous mapping theorem (e.g., Severini 2005, p. 336) as long as the function is suitably continuous. Within the context of this paper,

$$\lim_{n\to\infty} E[S(a, b, c, d)|p, q_s] = S[E(a|p, q_s), E(b|p, q_s), E(c|p, q_s), E(d|p, q_s)]. \quad (22)$$

Since many nonlinear measures are designed such that the right-hand-side of (22) is zero, this leads to them approaching equitability as the sample size is increased. To demonstrate this, Fig. 5a shows the expected value of a random forecast, with a base rate and population forecast rate of occurrence q_p both equal to 0.5, as a function of the number of samples n. This was calculated by numerically computing the score for every possible contingency table and applying (4). In order for the number of occurrences np to be an integer, n must be divisible by $1/p = 2$. The value of np is shown in the scale at the top of the figure. It can be seen that for small n, the expected value of the ETS for a random forecast is considerably greater than 0, but it decreases rapidly as n increases, to less than 0.01 for $n > 30$, for any base rate. We therefore describe ETS as being asymptotically equitable, that is, tending to equitability (by the definition in this paper) as n tends to infinity. SEDS also falls into this category, although CSI does not as it can be seen not to be equitable even in the limit of large n [indeed, CSI converges to $pq_s/(p + q_s - pq_s)$ as n increases with p and q_s held fixed]. As found in the previous section, for $p = 0.5$, HSS, PSS, and ORSS are all equitable for all n and, therefore, fall along the dotted line shown in Fig. 5a. Note that in calculating the expected value of ORSS, we are assuming that it is defined to be 0 whenever q_s is equal to 0 or 1, and similarly for SEDS when $q_s = 0$.

Figures 5b and 5c show the same plots but for p and q_p both reduced by a factor of 5 and 25 from that shown in Fig. 5a. The abscissa is changed since again it is only meaningful to consider integer values of the total number of occurrences np. This time only HSS and PSS are truly equitable, as shown by the horizontal dotted line. ETS exhibits the same dependence on n as before, becoming essentially equitable for n greater than around 30, while ORSS, EDS, and SEDS tend toward equitability only for considerably larger sample sizes; for $p = q_p = 0.1$, we find that one needs a sample size greater than around 1000 before the magnitude of the expected score for random forecasts falls below around 0.01, while for $p = q_p = 0.02$, this number is many times greater. Figure 5d shows the case for a biased forecast with $p = 0.1$ and $q_p = 0.2$. In this case the unconditional inequitability of EDS is evident from its non zero asymptote in the limit of large n.

In the introduction we noted that ETS and ORSS have previously been described as equitable (e.g., Mason 2003; Stephenson 2000); this is presumably because it has been implicitly assumed that the equality in (22) holds even without taking the limit of large n, which is only true for linear measures. Another way of expressing this is that the word ''expected'' in the definition of requirement 1 in the introduction has been treated as if it had been applied to ''random forecasting system'' rather than ''score,'' such that if the expected contingency table for a random forecasting system scored 0, then the measure has been treated as equitable.

The extent of the inequitability of ORSS and SEDS for $p = q_p = 0.02$ is quite startling; for sample sizes smaller than around 1000, a random forecasting system has an expected score of less than -0.5. An original justification for the requirement 1 in the introduction was that it made hedging impossible for random forecasting systems, but either of these measures could be hedged by forecasting occurrence all the time (or almost all the time in the case of ORSS, since it is undefined for $q_s = 1$), although admittedly this would only increase the score awarded to 0.

To determine whether asymptotically equitable measures approach zero from above or below, we may use Jensen's inequality, which states that the expectation of a convex function of a random variable is bounded below by the function applied to the expectation of the random variable, while for a concave function it is bounded above. For given p and q_s, the only random variable in the contingency table is a, and it is apparent from their definitions that ETS is a convex function of a while SEDS and ORSS are concave functions of a. Jensen's inequality therefore predicts that $E[\mathrm{ETS}(a)|p, q_s)] \geq \mathrm{ETS}[E(a|p, q_s)]$. Since $\mathrm{ETS}[E(a|p, q_s)] = 0$, the expectation of ETS is positive or 0, while the expectations of SEDS and ORSS are negative or 0, which is indeed what is observed in Fig. 5.

Table 1 presents a list of all the measures used in this paper, but placed into the appropriate categories of truly equitable, asymptotically equitable, and not equitable. Here, a measure is judged to be truly equitable by the criterion of this paper if it satisfies requirement 1 given in the introduction, or equivalently satisfies Eq. (7). The Peirce skill score is the only measure in Table 1 that also satisfies Gandin and Murphy's (1992) criteria for equitability. Note that there are many other not equitable measures in the literature that we have not considered (e.g., hit rate, false alarm rate, and proportion correct).

b. *Application to Finley's tornado data*

A classic example of a set of forecasts of rare events is the set of tornado forecasts of Finley (1884). He considered $n = 2803$ forecasts for which only $np = 51$ tornados occurred, so the base rate was very low at $p = 0.018$, similar to Fig. 5c. His contingency table had the following elements: $a = 28$, $b = 72$, $c = 23$, and $d = 2680$. Table 1 shows the values of the various measures for this contingency table, together with the expected scores for a random forecasting system $E(S|p)$, calculated by summing over all possible contingency tables using a population forecast rate of occurrence q_p of 0.0357, which is the proportion of forecasts in which a tornado was actually forecast in Finley's dataset. Of the asymptotically equitable measures, it can be seen that the expected values of ETS and HSS[3] for a random forecast are close enough to 0 that they can be considered to be effectively equitable for this value of n.

The other asymptotically equitable measures (SEDS and ORSS) cannot be considered equitable for the Finley data, as expected scores for a random forecasting system are notably different from zero. For these measures, the key parameter in determining the number of samples required before equitability can be assumed appears to be the number of hits that would be expected by chance for a random forecasting system, $E(a|p) = npq_p$. In considering a range of values of p and q_p, we find empirically that when $E(a|p)$ is less than around 10 (corresponding to $n = 1000$ for $p = qp = 0.1$ and $n = 25{,}000$ for $p = q_p = 0.02$), the magnitude of the expected score for a random forecasting system can exceed 0.01 and therefore these measures cannot be treated as equitable. In the case of the Finley data, the number of hits expected by chance (now for a given q_s) is only $E(a|p, q_s) = 1.82$ (see also Table 4 of Stephenson 2000).

c. *When does the difference between asymptotically equitable and truly equitable matter?*

In the case of Finley's tornado data, the forecasts are actually much better than random, so it might be argued that in this case the theoretical concerns about using an

TABLE 1. The first column classifies various verification measures into those that are truly equitable (i.e., satisfy requirement 1 in the introduction for any sample size n, base rate p, and population forecast rate q_p), those that are asymptotically equitable (equitable only in the limit $n \to \infty$, for all p and q_p), and those that are not equitable. The second column gives the value of the measure when applied to Finley's (1884) tornado forecasts, together with its standard error.* The third column gives the expected value of the measure for an equivalent random forecasting system with the same p as in Finley's data, and a population forecast rate of occurrence q_p set equal to the value of q_s for Finley's data; hence, the value shown is $E(S|p)$, as defined in (3) and (4). The fourth column gives the score when applied to the expected values of a–d for an equivalent random forecasting system, i.e., $S[E(a|p, q_s), E(b|p, q_s), E(c|p, q_s), E(d|p, q_s)]$. The truly equitable measures are the only ones to have a score of 0 in the third column.

Name of measure	Results for Finley's tornado forecasts		
	Score*	Expected random score	Score for expected random table
Truly equitable (linear)			
Peirce skill score (PSS)	0.523 ± 0.069	0	0
Heidke skill score (HSS)	0.355 ± 0.058	0	0
Truly equitable (nonlinear)			
Equitably transformed ETS	0.216 ± 0.043	0	−0.0001
Equitably transformed ORSS	0.963 ± 0.011	0	0.13
Equitably transformed SEDS	0.646 ± 0.038	0	0.13
Eq. (25)	0.296 ± 0.077	0	−0.0007
Asymptotically equitable			
Equitable threat score (ETS)	0.216 ± 0.043	0.0001	0
Heidke skill score cubed (HSS^3)	0.045 ± 0.022	0.000004	0
Odds ratio (OR)	45 ± 14	+∞/+1.03***	1
Log of odds ratio (LOR)	3.81 ± 0.31	−∞/−0.04***	0
Odds ratio skill score (ORSS)**	0.957 ± 0.013	−0.14	0
Symmetric extreme dependency score (SEDS)	0.593 ± 0.044	−0.15	0
Not equitable			
Critical success index (CSI)	0.228 ± 0.038	0.012	0.012
Extreme dependency score (EDS)	0.740 ± 0.048	−0.07	0.091

*Standard errors have been calculated for the various scores according to the following papers or methods: PSS, Stephenson (2000); HSS, Hogan et al. (2009); ETS (as a monotonic function of HSS); OR, LOR, and ORSS, Stephenson (2000); SEDS, Hogan et al. (2009); CSI, Hilliker (2004); EDS, Stephenson et al. (2008); the equitably transformed measures [by performing an error analysis on Eq. (23)]; and Eq. (25), as outlined in the text.

**Note that ORSS is truly equitable for the special case of $p = 0.5$.

***Strictly the expected values of OR and LOR are infinity, but if the one or two occurrences of infinity are removed from the mean, the resulting expected values are 1.03 and −0.04, respectively.

asymptotically equitable score can be dismissed since the actual number of hits, $a = 28$, is so much larger than the expected number by chance $E(a|p, q_s) = 1.82$. It is true that the detrimental consequences of using asymptotically equitable measures on small samples are more easily demonstrated for poor forecasts. Consider a forecaster whose forecasts of rare events are routinely evaluated using either SEDS or ORSS. Figures 5b–d show that random forecasts may yield an expected value of these two measures that is very much lower than zero. If the forecaster had low but better-than-random skill, on average, and was evaluated using samples small enough that sometimes the forecast sample performed worse than would be expected by chance, then it is possible that the mean score received by the forecaster could be less than 0. He or she would rightly want to be assessed on samples large enough that the verification measure could be reasonably regarded as equitable, for example, based on the criterion given earlier of $npq_p \geq 10$. Therefore, considerable care should be taken while using such measures when this criterion is not satisfied.

A potential solution is always to report confidence intervals on verification measures, as these will indicate whether a set of forecasts can be considered to be significantly better than random. In the case of the forecaster with low skill, his or her low value of the number of hits will result in a large error in the calculated performance measure, likely indicating that the forecast is indistinguishable from random, and therefore a larger sample size is required. This can be illustrated with Finley's tornado data. Suppose the tornado predictions were only slightly better than random, such that (from the final two columns of Table 1) the expected value of ORSS awarded lay between −0.14 and 0. The standard error of ORSS calculated on a forecast that actually happened to predict close to the expected random elements of the contingency table (e.g., Table 4 of Stephenson 2000) is around 0.68. Therefore, it would be very clear that the sample was insufficiently large to distinguish the forecast from a random one. The footnote to Table 1 provides references for the calculation of the standard error of each of the measures shown.

An alternative way to distinguish an actual forecast from a random one is to calculate the "*p* value," which is the probability that a score equal to or greater than that awarded could have been obtained by chance. If we condition this probability on p and q_s, then the probability of a random forecasting system obtaining ORSS ≥ 0 for Finley's tornados is $P(\text{ORSS} \geq 0|p, q_s) = 0.55$, thus indistinguishable from random. Conversely, the probability of randomly forecasting as well as or better than Finley's actual forecasts is $P(\text{ORSS} \geq 0.957|p, q_s) = 6 \times 10^{-29}$. Conveniently, this second p value is the same no matter what verification measure we choose.

5. Nonlinear equitable measures

a. How to make an inequitable measure equitable

It turns out that we may transform any inequitable or asymptotically equitable measure S (even a nonlinear one) into a truly equitable measure S' via the simple linear transformation:

$$S' = \frac{S - E(S|p, q_s)}{\max(S) - E(S|p, q_s)}, \tag{23}$$

where the only requirement on S is that the expected score for a random forecasting system, $E(S|p, q_s)$, and the score for a perfect forecasting system, $\max(S)$, are finite. Unfortunately, this excludes the odds ratio and the log of odds ratio. Clearly, (23) is very similar to the classic definition of a skill score, in which some function of the contingency table S is compared to the value for a baseline forecast, which may be climatology, persistence, or a random forecast. In (23) we require that this baseline value is specifically set to be $E(S|p, q_s)$.

To illustrate this transformation, consider the values of equitable threat scores for $n = 4$ and $p = q_s = ½$, shown by the middle column of the contingency tables in Fig. 2c. It can be seen that a perfect forecast scores 1, the expected contingency table for a random forecast scores 0, and the worst possible forecast scores $-⅓$. This results in $E(S|p, q_s) = ⅑$. Performing the transformation in (23) yields corresponding values for the "equitably transformed ETSs" of 1, $-⅛$, and $-½$, respectively. It is then easily confirmed that $\mathrm{E}(S'|p, q_s) = 0$. We do have the curious property that the expected contingency table for a random forecasting system no longer scores 0, even though the expected value of the score overall is 0. This is a consequence of the fact that the nonlinearity has been retained, so it is not possible, in general, for both of these quantities to be 0 as it is for linear measures such as PSS and HSS.

Table 1 shows the equitably transformed versions of ETS, ORSS, and SEDS applied to Finley's forecasts. Of course, this is specifically a linear transform, and an alternative would be to apply a nonlinear transform, such as that transforming ETS into HSS. However, this case removes the nonlinearity of the original measure, which may not be desirable. In the case of the linear equitable transform of SEDS, it appears that we are able to have the best of both worlds: a truly equitable measure that is also nondegenerate for rare events via its nonlinearity. Note that for many nonlinear measures it is difficult or impossible to express $E(S|p, q_s)$ analytically, so in practice it would need to be calculated numerically in the application of an equitable transform.

b. A method to generate nonlinear equitable measures

Rather than transforming existing nonlinear measures to make them equitable, another approach to developing equitable nonlinear measures is to generalize (17) to

$$S = f(p, q_s)\left\{\frac{g(a)}{\mathrm{E}[g(a)|p, q_s]} - \frac{h(b)}{\mathrm{E}[h(b)|p, q_s]}\right\} + S_0, \tag{24}$$

where $g(a)$ and $h(b)$ can be any positive, monotonically increasing function of their arguments, and $f(p, q_s)$ must be positive as before. The resulting measure is equitable, since for a given p and q_s, a random forecast will have expected values of 1 for both of the two terms inside the braces; these will therefore cancel to yield an expected value for the measure of S_0.

The Peirce skill score represents the simplest case, with $f = q_s$, $S_0 = 0$, and $g(x) = h(x) = x$. An example of a nonlinear measure created from (24) uses $f = q_s^2$, $S_0 = 0$, and $g(x) = h(x) = x(x - 1)$. Since the expected values of $g(a)$ and $h(b)$ can be written out analytically in this case, we obtain

$$\begin{aligned} S &= \frac{a(a-1)}{np(np-1)} - \frac{b(b-1)}{(n-np)(n-np-1)} \\ &= \frac{a(a-1)}{(a+c)(a+c-1)} - \frac{b(b-1)}{(b+d)(b+d-1)} \end{aligned}. \tag{25}$$

The value of f has been chosen such that a perfect forecast always scores 1. Application to Finley's forecasts in Table 1 confirms that this measure is indeed equitable. Error bounds can be calculated for this measure by modeling a and b as independent binomially distributed variables: $a|p \sim \mathrm{Bin}(np,)$ and $b|p \sim \mathrm{Bin}[n(1 - p), F]$, where the hit rate is $H = a/(a + c)$ and the false alarm rate is $F = b/(b + d)$, enabling the errors in a and b to be estimated.

Note that we are not arguing that the measure shown in (25) has any desirable properties apart from equitability; it simply serves to demonstrate how nonlinear equitable measures may be generated. It would be interesting to see if a measure of this form could be developed

that is reliable for rare events, but such work is beyond the scope of this paper.

6. Discussion and conclusions

Gandin and Murphy (1992) pioneered the concept of equitability and proposed the definition expressed by the two requirements stated in the introduction. In this paper we have argued that only the first is necessary: that an equitable verification measure awards all random forecasting systems, including those that always forecast the same value, the same expected score. A detailed discussion of why this is desirable was given in the introduction. By removing Gandin and Murphy's linearity requirement, equitability is no longer incompatible with some other desirable properties, such as being a reliable indicator of skill for rare events. We have highlighted that there appears to be some confusion in the literature about the meaning of equitability, by demonstrating that the widely used verification measure ''equitable threat score'' (ETS) is not in fact equitable. Given this fact, we recommend that in future the measure should be known by one of its other names, most obviously the Gilbert skill score (Mason 2003), and that the terminology ETS should be avoided. We suggest the term ''asymptotically equitable'' to describe measures, like ETS, that are equitable only in the limit of a large sample of forecasts, leading to the hierarchy of equitability shown in Table 1.

This has implications for the selection of verification measures to use in a particular application. Many newcomers to the field of forecast verification are bewildered by the number of different measures available to measure the skill of a set of binary forecasts, and ask why there is not one that is the best to use. Murphy (1991) pointed out that the verification problem is inherently multidimensional, and indeed Figs. 1 and 3 plot the possible sets of forecasts in a two-dimensional space, indicating that at least two numbers must be reported to fully characterize performance (e.g., a measure of *bias* and a measure of *skill*). As has been stated by previous authors, the further problem is that different measures of skill have desirable properties (e.g., equitability, difficulty to hedge, and usefulness for rare events) in different amounts, and none is strong in all. Nonetheless, if equitability is regarded as high on the list of desirable properties for a measure of skill, then Table 1 provides some guidance on preferred measures to use.

In general, the case for advocating asymptotically equitable measures over inequitable ones is easy to make since the other desirable properties can always be retained. If we adhere to Gandin and Murphy's requirements for equitability, then the same does not hold for advocating true equitability over asymptotic equitability. A clear example is the property of tending to a useful value in the limit of very rare events ($p \to 0$), which is possible for EDS and SEDS by the use of logarithms in the definition of these measures, but the resulting nonlinear dependence on the elements of the contingency table is what makes them violate Gandin and Murphy's second requirement, as given in the introduction, and the two desirable properties of true equitability and being nondegenerate for rare events appear to be incompatible. If we have a large enough sample, then an asymptotically equitable measure will be close enough to equitable that this dilemma goes away; Stephenson et al. (2008) and Hogan et al. (2009) clearly showed that for large datasets, the truly equitable HSS and PSS measures were degenerate for rare events, a problem overcome by the use of EDS or SEDS (provided the former is calibrated first).

For smaller sample sizes when the inequitability of these measures is likely to be more of a problem, there appears to be a simple solution: by rejecting Gandin and Murphy's second requirement of equitability, we are permitted to rescale nonlinear measures such as SEDS so that they are truly equitable while retaining their desirable properties for verifying rare events. This opens up the possibility of new equitable measures to be designed that have many more desirable properties than has previously been possible to encapsulate within a single measure.

Acknowledgments. We thank Harold Brooks, Barbara Brown, Charles Doswell, Ian Mason, Fedor Mesinger, Dan Wilks, and two of the anonymous reviewers for very useful correspondence and information regarding the history of equitability and the equitable threat score.

APPENDIX

Proof of Eq. (7)

In this appendix we prove that $E(S|p) = S_0$ for all p and q_p if and only if $E(S|p, q_s) = S_0$ for all p and q_s, leading to Eq. (7). The backward implication is straightforward: $E(S|p) = E[E(S|p, q_s)] = E(S_0) = S_0$ for all p and q_p. For the forward implication, we let $f = nq_s$. Then,

$$E(S|p) = E[E(S|p, f)] = \sum_{f=0}^{n} P(f|p) \sum_{a} P(a|p, f) S(a), \tag{A1}$$

where we have written $P(a|p, f)$ for $P(a|p, q_s)$ and $S(a)$ for $S(a, f - a, pn - a, n - pn - f + a)$. The two summations can be thought of as summing over the columns and rows of the possible contingency tables in Figs. 1 and 3. Substitution of (5) into (A1) leads to

$$E(S|p) = \sum_{f=0}^{n} C(n, f) q_p^f (1 - q_p)^{n-f} \sum_a P(a|p, f) S(a). \quad \text{(A2)}$$

Now $P(a|p, f)$ and $S(a)$ are independent of q_p; so, from (A2) we see that $E(S|p)$ is a polynomial in q_p. But $E(S|p)$ is constant in q_p by assumption and equal to S_0. Therefore, the coefficients of must be 0 for all $f > 0$ and must equal S_0 when $f = 0$. We now show by induction that these restrictions imply that

$$E(S|p, f) = \sum_a P(a|p, f) S(a) = S_0, \quad \text{(A3)}$$

for all $f = 0, 1, \ldots, n$. To find the coefficient of q_p^f in the polynomial, we first note that the binomial theorem gives

$$(1 - q_p)^{n-f} = \sum_{r=0}^{n-f} C(n - f, r)(-1)^r q_p^r. \quad \text{(A4)}$$

Therefore,

$$E(S|p) = \sum_{f=0}^{n} \sum_{r=0}^{n-f} C(n, f) C(n - f, r)(-1)^r q_p^{r+f} \sum_a P(a|p, f) S(a), \quad \text{(A5)}$$

and rewriting the summation indices yields

$$E(S|p) = \sum_{t=0}^{n} \sum_{r=t}^{n} C(n, t) C(n - t, r - t)(-1)^{r-t} q_p^r \sum_a P(a|p, t) S(a). \quad \text{(A6)}$$

The coefficient of q_p^f is therefore

$$\sum_{t=0}^{f} C(n, t) C(n - t, f - t)(-1)^{f-t} \sum_a P(a|p, t) S(a). \quad \text{(A7)}$$

When $f = 0$, we obtain

$$S_0 = \sum_a P(a|p, 0) S(a). \quad \text{(A8)}$$

We may assume that $\Sigma_a P(a|p, r) S(a) = S_0$ for all $r = 0, \ldots, f - 1$. The coefficient of q_p^f is then

$$S_0 \sum_{t=0}^{f-1} C(n, t) C(n - t, f - t)(-1)^{f-t} + C(n, f) \sum_a P(a|p, f) S(a). \quad \text{(A9)}$$

By expanding the binomial coefficient $C(\cdot,\cdot)$ into its component factorials, we find that

$$\sum_{t=0}^{f-1} C(n, t) C(n - t, n - f)(-1)^{f-t} = (-1)^f C(n, f) \sum_{t=0}^{f-1} C(f, t)(-1)^t = (-1)^f C(n, f) \left[-(-1)^f + C \sum_{t=0}^{f} C(f, t)(-1)^t \right] = -C(n, f). \quad \text{(A10)}$$

The last step of (A10) is a consequence of the binomial theorem [Eq. (A4)], which shows that the summation is zero. Thus,

$$-C(n, f) S_0 + C(n, f) \sum_a P(a|p, f) S(a) = 0,$$

and the result [Eq. (A3)] follows.

REFERENCES

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

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

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

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–586.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.,* **1,** 85–88.

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

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

Gringorten, I. I., 1967: Verification to determine and measure forecasting skill. *J. Appl. Meteor.,* **6,** 742–747.

Heidke, P., 1926: Calculation of the success and goodness of strong wind forecasts in the storm warning service. *Geogr. Ann. Stockholm,* **8,** 301–349.

Hilliker, J. L., 2004: The sensitivity of the number of correctly forecasted events to the threat score: A practical application. *Wea. Forecasting,* **19,** 646–650.

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.

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

Livezey, R. E., 2003: Categorical events. *Forecast Verification—A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 77–96.

Manzato, A., 2005: An odds ratio parameterization for ROC diagram and skill score indices. *Wea. Forecasting,* **20,** 918–930.

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

——, and V. Lakshmanan, 1999: On the uniqueness of Gandin and Murphy's equitable performance measures. *Mon. Wea. Rev.,* **127,** 1134–1136.

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

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

Mesinger, F., and T. L. Black, 1992: On the impact on forecast accuracy of the step-mountain (eta) vs. sigma coordinate. *Meteor. Atmos. Phys.,* **50,** 47–60.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

——, 1996: The Finley affair: a signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

——, 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.

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

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

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

Severini, T. A., 2005: *Elements of Distribution Theory*. Cambridge University Press, 515 pp.

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.

Yule, G. U., 1900: On the association of attributes in statistics. *Philos. Trans. Roy. Soc. London,* **194A,** 257–319.