METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 41–50 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.53

RMetS
ROYAL METEOROLOGICAL SOCIETY

The extreme dependency score: a non-vanishing measure for forecasts of rare events

D. B. Stephenson,[a]* B. Casati,[b] C. A. T. Ferro[a] and C. A. Wilson[c†]
[a] *School of Engineering, Computing, and Mathematics, University of Exeter, UK*
[b] *Meteorological Service of Canada, Montreal, Canada*
[c] *Meteorological Office, Exeter, UK*

ABSTRACT: Accurate prediction of rare high-impact events represents a major challenge for weather and climate forecasting. Assessment of the skill at forecasting such events is problematic because of the rarity of such events. Skill scores traditionally used to verify deterministic forecasts of rare binary events, such as the equitable threat score (ETS), have the disadvantage that they tend to zero for vanishingly rare events. This creates the misleading impression that rare events cannot be skilfully forecast no matter which forecasting system is used.

This study presents a simple model for rare binary-event forecasts and uses it to demonstrate the trivial non-informative limit behaviour of several often-used scores such as ETS. The extreme dependency score (EDS) is proposed as a more informative alternative for the assessment of skill in deterministic forecasts of rare events. The EDS has the advantage that it can converge to different values for different forecasting systems and furthermore it does not explicitly depend upon the bias of the forecasting system.

The concepts and scores are demonstrated using an example of 6-hourly precipitation total Met Office forecasts for Eskdalemuir in Scotland over the period 1998–2003.

KEY WORDS verification; extreme; skill

Received 17 September 2007; Revised 10 December 2007; Accepted 2 January 2008

1. Introduction

One of the important roles of weather forecasting is to help forewarn society about rare extreme events such as tornadoes that can lead to severe losses. However, to be able to do this it is necessary that the forecasting systems have skill at forecasting such rare events.

Estimating the skill for such events is problematic for several reasons such as the trivial non-informative limit of many scores for rarer events, and the large amount of sampling uncertainty on scores estimated on past rare events. This article will address the first of these issues in the simplest context of deterministic forecasts of binary events.

There has been a long history of development (and reinvention!) of scores for deterministic forecasts of binary events (Mason, 2003). In an intelligent critique of Finley's use of proportion correct (PC) for the forecasting of rare events (US tornadoes), Gilbert (1884) suggested two new scores which he referred to as *ratio of verification* and *ratio of success in forecasting*. The ratio of verification is the ratio of the number of hits 'a' to the number of events that are not correct rejections ($a + b + c$) and is now known as the *threat score* (Palmer and Allen, 1949) or the *critical success index* (Donaldson *et al*., 1975; Mason, 1989; Schaefer, 1990). Refer to Table I for mathematical definitions of this and some other commonly used scores. By comparing the threat score to what one would obtain for random forecasts, it is possible to construct a skill score that is now known as the equitable threat score (ETS) (Doswell *et al*., 1990; Gandin and Murphy, 1992). This skill score is the ratio of success first proposed by Gilbert (1884). Both the threat score and the ETS are widely used operationally to assess the performance at forecasting events over a range of thresholds. Other skill scores for rare-event forecasts such as the Peirce skill score (PSS) and the Heidke skill score (HSS) are reviewed and compared in various articles (Doswell *et al*., 1990; Schaefer, 1990; Marzban, 1998; Mason, 2003).

The limit of these and other deterministic scores for increasing rarity is not easy to understand. Göber *et al*. (2004) noted that while the PSS vanished for rarer rainfall events, the odds ratio (OR) increased. They argued that the OR, a measure of association, was perhaps a more reliable measure of skill for extreme events. This study goes one step further and shows that an even better approach is to use a measure of association for

* Correspondence to: D. B. Stephenson, School of Engineering, Computing, and Mathematics, University of Exeter, Exeter EX4 4QF, UK. E-mail: d.b.stephenson@exeter.ac.uk
† This article was co-written by C.A. Wilson of the Met Office, Exeter. It is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

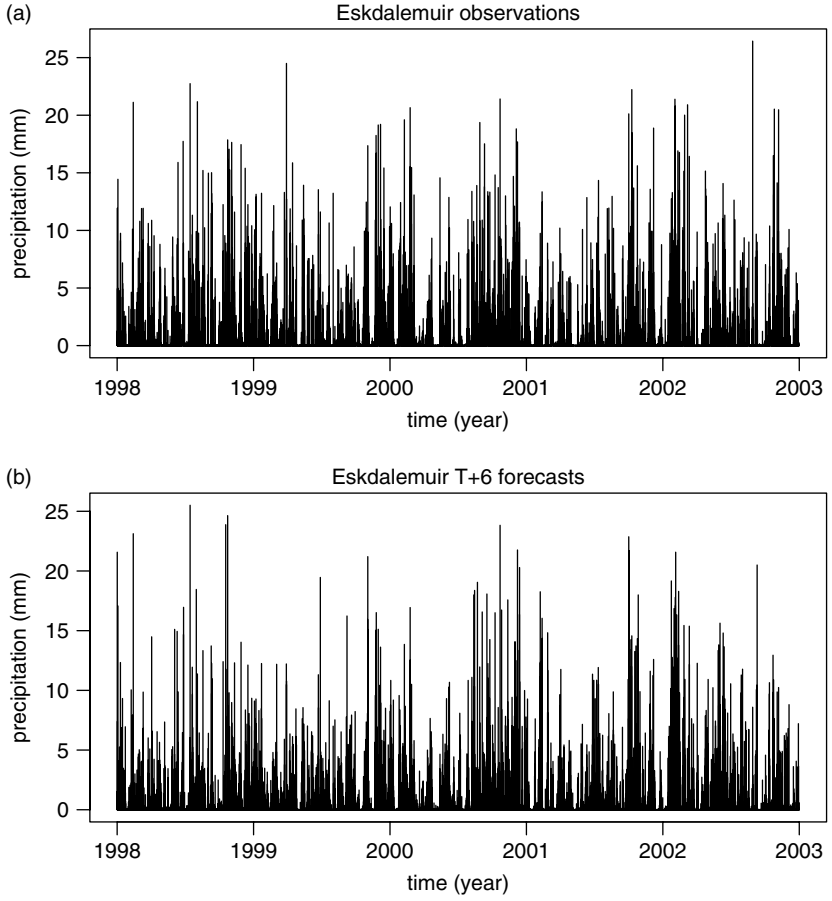

Figure 1. Time series of 6-h precipitation totals: (a) the gauge observations at Eskdalemuir, and (b) the nearest-neighbour grid point Met Office 6-h lead forecasts.

bivariate extreme events (Coles, 2001). This measure of association is simple to calculate from the number of hits and misses and has several important advantages over the scores traditionally used to assess the skill of forecasting rare binary events.

2. Forecast example: Eskdalemuir precipitation totals

The concepts in this study are illustrated using an example consisting of 6266 forecasts of 6-hourly rainfall totals in the period 1 January 1998 to 31 December 2003 (Figure 1). The observations (Figure 1(a)) are the synoptic reports of 6-h totals measured by gauge at the long-running Eskdalemuir observatory in Scotland (55°19′N, 3°12′W, 242 m elevation above mean sea-level). The precipitation forecasts (Figure 1(b)) are the direct forecast output of the UK-mesoscale model at the nearest grid box to Eskdalemuir accumulated over 6-h range. Although not apparent in Figure 1, 3084 out of the 6266 observed totals are zero at Eskdalemuir (i.e. 49% of the totals), whereas the forecasts have fewer dry totals (2316).

The UK-mesoscale model was until September 2006 the main model guidance for short-range weather forecasts over the United Kingdom at the Met Office (Cullen *et al*., 1997; Davies *et al*., 1999; Webster *et al*., 2003). The model was a limited area non-hydrostatic version of the Unified Model having a 12 km horizontal grid and 38 vertical layers and mixed-phase microphysics (Wilson and Ballard, 1999). The mesoscale-model had its own data assimilation cycle and it was coupled to the global forecast model only through the lateral boundaries. (The UK-mesoscale model has now been replaced by a larger domain 12 km grid version covering the North Atlantic and Europe.)

For comparison, we shall also show results for persistence and random forecasts. Results for 6-h-ahead persistence forecasts were obtained by calculating contingency counts between the observed precipitation totals and those in the preceeding 6-h period (6113 events were available). There are fewer persistence forecasts than the original number of observations because of missing values in the observational record. Expected scores for random forecasts were also derived by recalculating new contingency table relative frequencies as products of the marginal frequencies of the original Met Office forecasts (Mason, 2003). This procedure ensures no association between the forecasts and observations while maintaining the same bias as in the original forecasts (Stephenson, 2000).

Figure 2(a) shows a scatter plot of the Met Office mesoscale-model forecasts against the observations. There appears to be some positive association but it is difficult to observe this clearly due to skewness in the

DOI: 10.1002/met

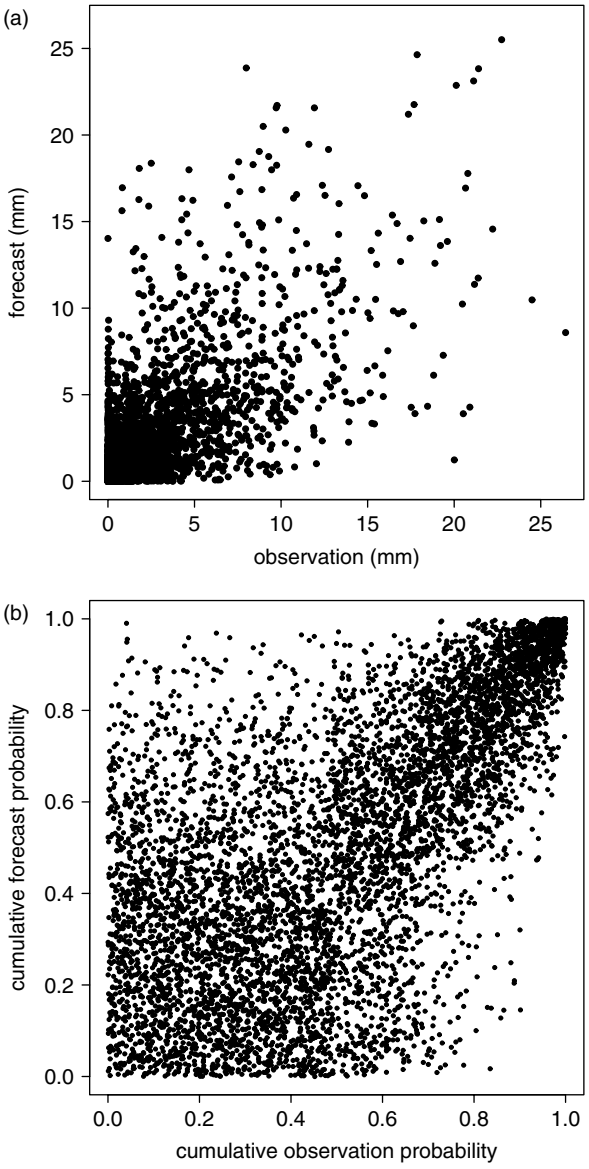

Figure 2. Scatter plot of (a) 6-h total precipitation forecasts *versus* observations, (b) empirical probabilities of forecasts *versus* those of observations. Note the strong association in the upper right hand corner for extreme precipitation events.

rainfall totals. The skewness in the marginal distributions can be transformed out by calculating rank-based empirical probabilities $q_t = (rank(x_t) - 1)/(n - 1)$ where x_t is the rainfall total (observed or forecast) at time t and n is the total number of rainfall values (e.g. 6266 for the mesoscale-model forecasts). The empirical probability is an estimate of the probability $\Pr(X \leq x_t)$ of having a rainfall amount less than the observed value x_t and so is largest for more extreme rainfall amounts. Figure 2(b) shows a scatter plot of the empirical probabilities of the mesoscale-model forecasts *versus* those of the corresponding observations. A strong association is now remarkably apparent between the large-rainfall values having large empirical probabilities (Figure 2(b); top right hand corner). In other words, there is evidence of association at extreme-rainfall amounts. The empirical cumulative distribution functions (q_t *vs* x_t) for the observations and mesoscale-model forecasts were found to be very similar for this set of forecasts (not shown).

As shown in Casati *et al.* (2004), empirical distribution functions can be used to non-linearly recalibrate precipitation forecasts. This approach will be used here to define binary rainfall events as those that exceed a pre-defined probability threshold rather than a given rainfall amount. In other words, the extremeness of rainfall events in forecasts and observations is defined by their corresponding rarity as estimated by the empirical *exceedance probability* $p_t = 1 - q_t$. For example, rainfall totals exceeding around 5 mm were found to have an exceedance probability of 0.1 (the solid black lines in Figure 2(b)). This simple recalibration approach has the virtue of eliminating frequency bias in the probability of forecasting the event – by definition, the probability of occurrence of the forecast event is identical to the probability of occurrence (the *base rate*) of the observed event. In other words, the threshold used to define the binary event, and the decision threshold used to determine whether to issue a binary warning of the event are varied together in such a way that the frequency bias stays equal to one. To aid interpretation, Figures 3–5 in this article have the x-axis labelled both with the empirical probability threshold q and the corresponding empirical quantile of the observations.

3. The contingency table for rare events

The performance of binary-event forecasts at a specific threshold is a summary of the 4 counts (a, b, c, d) in the 2×2 contingency table that contains the number of forecast hits, false alarms, misses, and correct rejections (Mason, 2003). For example, for a base rate of 0.1, the counts can be obtained by counting the number of dots in the top right, top left, bottom right, and bottom left quadrants, respectively shown in Figure 2.

Table I shows how the relative frequencies in such a table can be written in terms of the *base rate* of the observed event $p = (a + c)/n$, the *hit rate* $H = a/(a + c)$, and the *frequency bias* $B = (a + b)/(a + c)$. Note that $B = 1$ by definition in our example because of the use of empirical probabilities to recalibrate the forecasts and observations. For example, the *false-alarm* rate defined as $F = b/(b + d)$ is simply equal to $F = (B - H)p/(1 - p)$: the product of $B - H$ and the odds of the observed event.

In the limit of increasingly rare events and finite bias, $pB \to 0$, the cell counts a, b, c tend to zero but at different rates. One needs to be particularly careful when calculating the rare-event limit for various scores in order to obtain the correct result. For example, Doswell *et al.* (1990) and Marzban (1998) have both noted that there is potentially ambiguity in how to take the rare-event limit of certain scores so as to avoid singularities. To avoid such ambiguities, it is necessary to make clear assumptions about how the hit rate H and the bias B change as a function of base rate as the base rate $p \to 0$. Here, we will make the following assumptions:

DOI: 10.1002/met

1. The hit rate tends to zero as $H \sim \kappa p^{1/\eta-1}$ where $0 < \eta < 1$ and $0 \leq \kappa \leq 1$
2. The frequency bias tends to a constant $B \sim \beta$ where $0 < \beta < \infty$

with the parameters β and κ constrained so that all four relative frequencies in the contingency table lie in the interval (0,1).

The exponent η is a key parameter in determining how fast the hit rate converges to zero for rarer events, yet it is unimportant for the false-alarm rate, $F = (B - H)p/(1 - p)$, which behaves as βp in the limit $p \to 0$ when $\eta < 1$. Hence, under these assumptions, the locus of points $(F, H) \sim (\beta p, \kappa p^{1/\eta-1}) \to (0, 0)$ as $p \to 0$ when $\eta < 1$. Such behaviour is typically observed to occur for operational forecasting systems. The behaviour is analogous to the regularity property observed to occur for receiver-operating characteristic (ROC) curves (Swets, 1986; Mason, 2003). However, it should be noted that the locus of points (F, H) here is not an ROC because both the threshold defining the binary event and the forecast threshold are being varied simultaneously whereas for ROC curves the event threshold is held fixed. In order to vanish as $p \to 0$, the hit rate must have a power-law limiting behaviour with positive exponent. The hit rate could, in principle, tend to a non-zero constant as the base rate tends to zero but this has never been documented to occur for meteorological forecasting systems. However, for the sake of mathematical completeness, we shall also consider this special case of $\eta = 1$ where the hit rate tends to the constant $H \sim \kappa$ as $p \to 0$. Extreme-value theory arguments show that such strong asymptotic dependence can, in principle, occur and so should therefore also be considered (Coles *et al.*, 1999; Ferro, 2007).

The frequency bias is assumed to stay finite and non-zero as the base rate tends to zero. It is easy to imagine forecasting systems where this would not be the case, for example, a system where $B \to \infty$ because the system forecasts too many events as the base rate tends to zero. However, one could argue that such forecasting systems are unrealistic and so should either be recalibrated or redesigned so as to have a finite non-zero frequency bias. It is reasonable to expect that forecast users should receive warnings that have bias

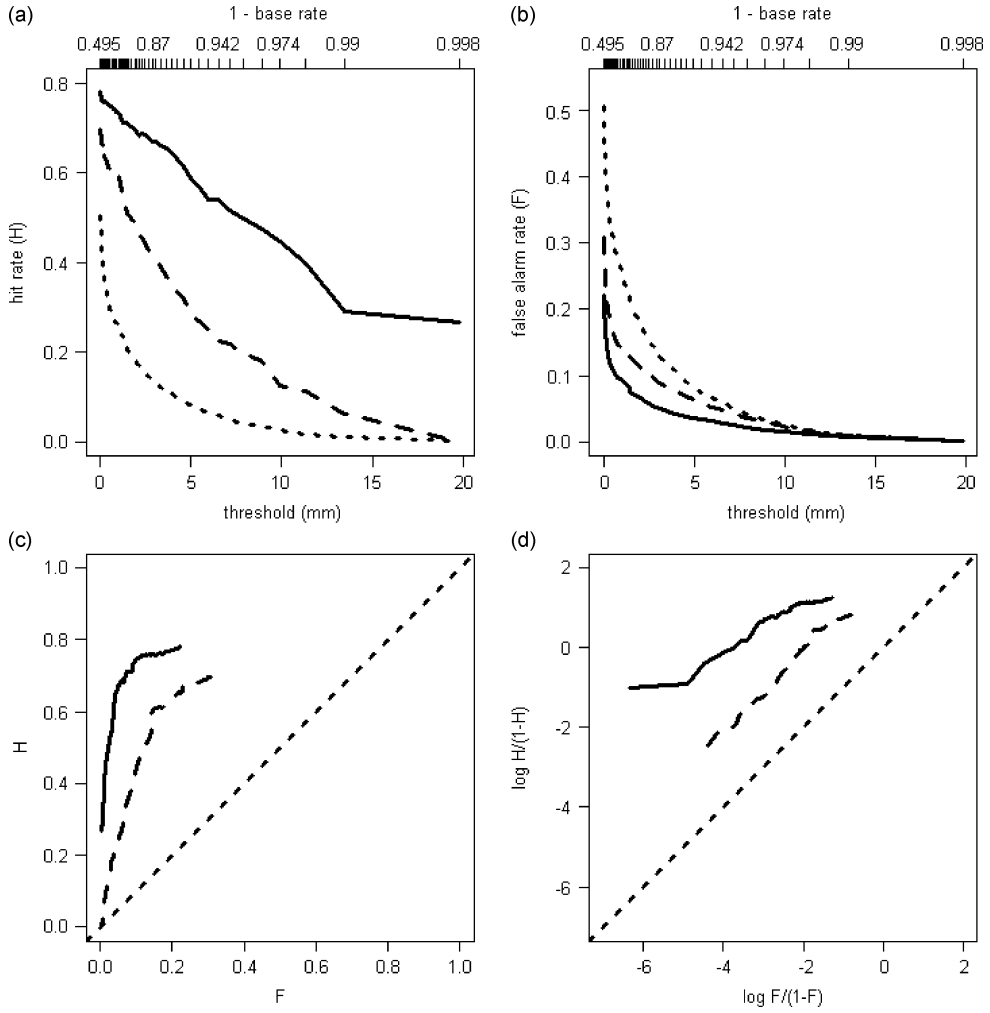

Figure 3. (a) Hit rate for different thresholds, (b) false-alarm rate for different thresholds, (c) hit rate *versus* false-alarm rate, (d) same as (c) but on logistic axes. Solid line denotes Met Office forecasts, dashed line denotes 6-h-lead persistence forecasts, and the dotted line denotes random forecasts.

DOI: 10.1002/met

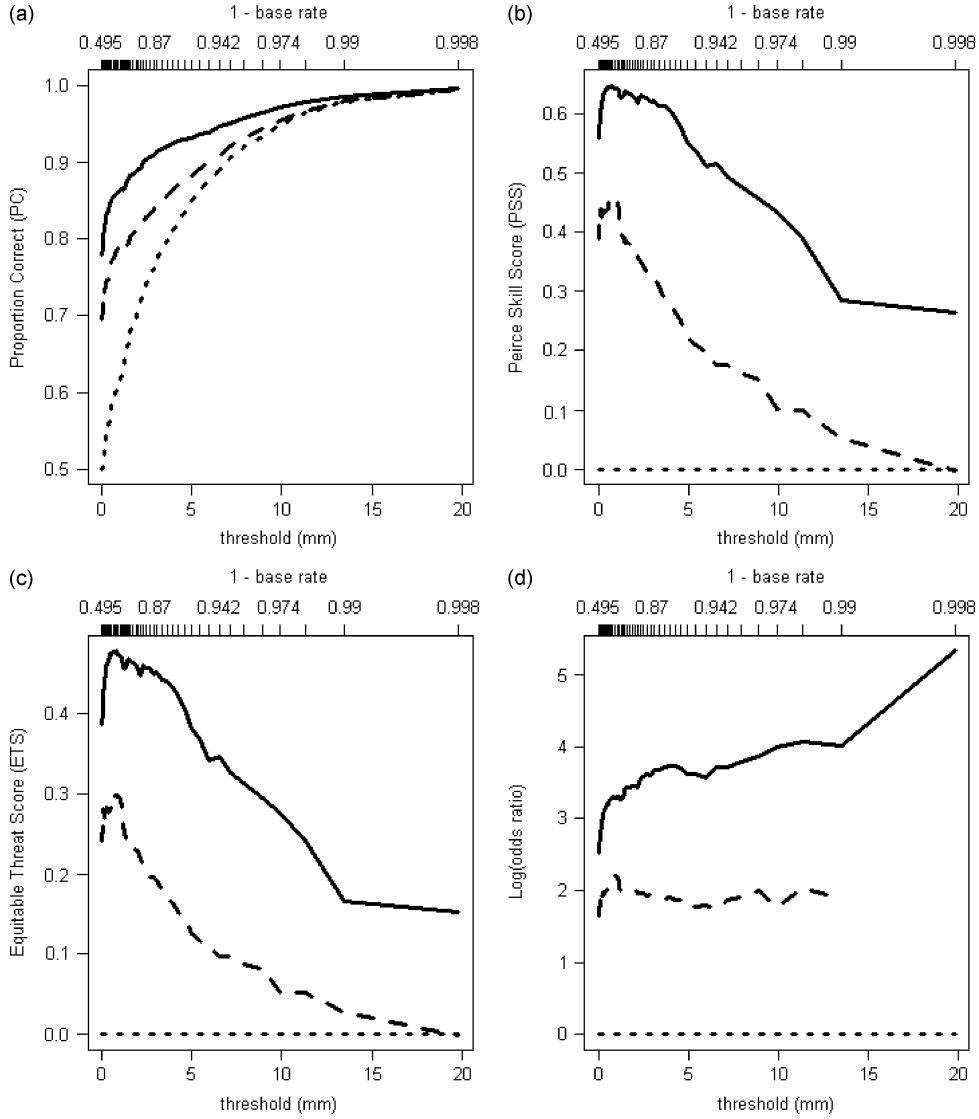

Figure 4. Scores *versus* threshold: (a) proportion correct PC, (b) the Peirce skill score (PSS), (c) the equitable threat score (ETS), and (d) logarithm of the OR. Solid line denotes Met Office forecasts, dashed line denotes 6-h-lead persistence forecasts, and the dotted line denotes random forecasts.

in an acceptable range that is bounded from above and does not include zero. Although the forecasting example in this article is recalibrated so as to be unbiased with $B = 1$ for all values of base rate, we have included the possibility of different finite non-zero β in the equations that follow.

It is instructive to consider lower and upper bounds for forecasting systems: random forecasts and perfect forecasts. Random forecasts are *independent* of the observed events and so have relative frequencies that are products of the marginal frequencies, for example, $a/n = (a+b)(a+c)/n^2$. For such forecasts, $H = Bp$ and hence, the asymptotic limit of random forecasts is given by $\eta = 1/2$ and $\kappa = \beta$. Perfect forecasts are *identical* to the observed events and so produce no misses or false alarms ($b = c = 0$) and therefore are unbiased with $B = (a+b)/(a+c) = 1$. Hence, for perfect forecasts $H = 1$ and $F = 0$ and so in the asymptotic limit are described by $\eta = 1$, $\kappa = 1$, and $\beta = 1$. Note that perfect forecasts provide one example of a forecasting system where the hit rate does not tend to zero as the base rate goes to zero.

Figure 3 shows the hit rate and false alarm-rate behaviour as a function of threshold. The hit rates for the Met Office and persistence forecasts decay more slowly than the base rate with increasing threshold (Figure 3(a)) whereas the false-alarm rates tend towards the base rate at high thresholds as to be expected for unbiased forecasts (Figure 3(b)). The loci of (F, H) points shown in Figure 3(c) all converge towards the point $(F, H) = (0, 0)$. The Met Office (F, H) locus is above the other loci, which suggests that the Met Office forecasts have the best forecast discrimination. The (F, H) loci are close to being piecewise straight lines when plotted on logistic axes in Figure 3(d) – in other words, when $H/(1-H)$ is plotted against $F/(1-F)$ on logarithmic axes. The reason for this can be understood from our assumptions which imply that in the limit $p \to 0$ (the left hand side of Figure 3(d)) then $\log H/(1-H) \sim \log\kappa + (1/\eta - 1)\log p$ and $\log F/(1-F) \sim \log\beta + \log p$ if $\eta < 1$.

DOI: 10.1002/met

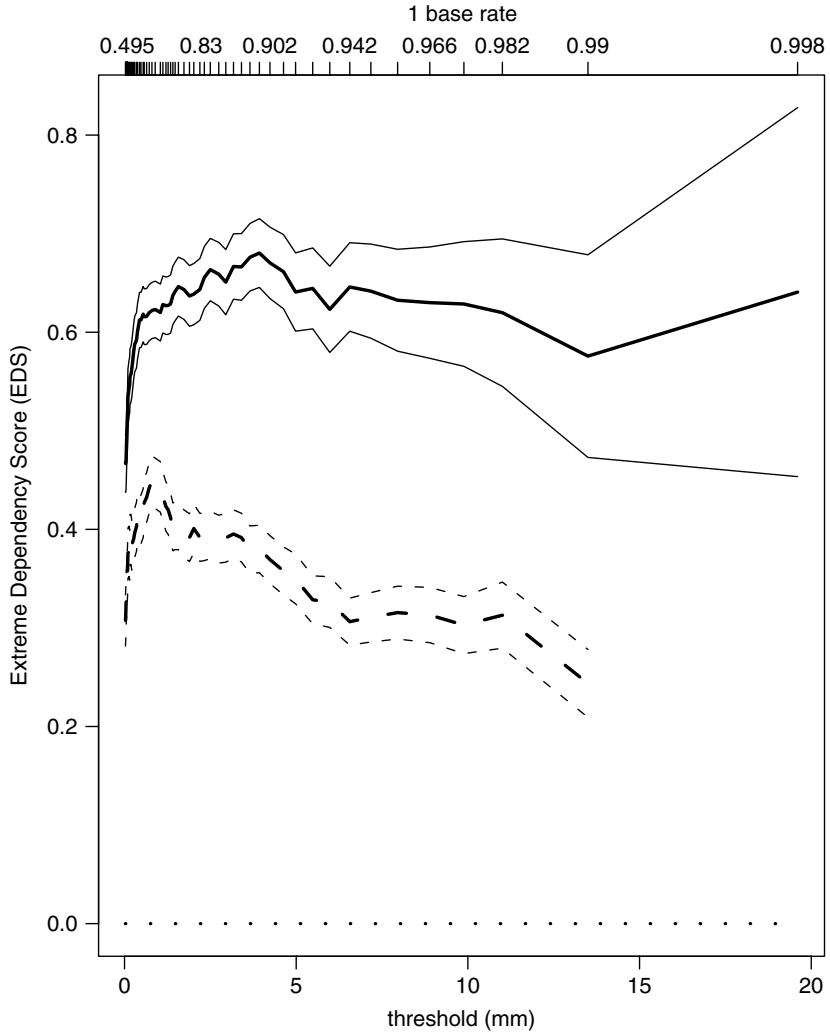

Figure 5. The extreme dependency score *versus* threshold for the Met Office forecasts (solid line), 6-h-lead persistence forecasts (dashed line), and random forecasts (dotted line).

Table I. Limits of some standard scores as the base rate tends to zero: proportion correct (PC), threat score (TS)/critical success index (CSI), equitable threat score (ETS)/Gilbert skill score (GSS), Heidke skill score (HSS), Peirce skill score (PSS), and OR. We write $\delta = 1/\eta - 1$ and denote convergence from below, above (grey shaded), and unspecified by $\uparrow$, $\downarrow$, and $\updownarrow$ respectively. The counts $a_r = (a+b)(a+c)/n$ and $d_r = (c+d)(b+d)/n$ are those expected for random forecasts having the same marginal counts as the original forecasts.

Score	Definition	Score as $f(\kappa, \beta, \delta, p)$	Faster rate of decrease than random $0 < \eta < 1/2$	Same rate of decrease as random $\eta = 1/2$	Slower rate of decrease than random $1 > \eta > 1/2$	No decrease in hit rate $\eta = 1$
PC	$\frac{a+d}{n}$	$1-(1+\beta-2\kappa p^{\delta})p$	$1-(1+\beta)p \uparrow 1$	$1-(1+\beta)p \uparrow 1$	$1-(1+\beta)p \uparrow 1$	$1-(1+\beta-2\kappa)p \uparrow 1$
TS/CSI	$\frac{a}{a+b+c}$	$\frac{\kappa p^{\delta}}{1+\beta-\kappa p^{\delta}}$	$\frac{\kappa p^{\delta}}{1+\beta} \downarrow 0$	$\frac{\kappa p}{1+\beta} \downarrow 0$	$\frac{\kappa p^{\delta}}{1+\beta} \downarrow 0$	$\frac{\kappa}{1+\beta-\kappa}$
ETS/GSS	$\frac{a-a_r}{a+b+c-a_r}$	$\frac{\kappa p^{\delta}-\beta p}{1+\beta(1-p)-\kappa p^{\delta}}$	$\frac{-\beta p}{1+\beta} \uparrow 0$	$\frac{(\kappa-\beta)p}{1+\beta} \updownarrow 0$	$\frac{\kappa p^{\delta}}{1+\beta} \downarrow 0$	$\frac{\kappa-\beta p}{1+\beta(1-p)-\kappa} \updownarrow \frac{\kappa}{1+\beta-\kappa}$
HSS	$\frac{a+d-a_r-d_r}{n-a_r-d_r}$	$\frac{2(\kappa p^{\delta}-\beta p)}{1+\beta-2\beta p}$	$\frac{-2\beta p}{1+\beta} \uparrow 0$	$\frac{2(\kappa-\beta)p}{1+\beta} \updownarrow 0$	$\frac{2\kappa p^{\delta}}{1+\beta} \downarrow 0$	$\frac{2(\kappa-\beta p)}{1+\beta-2\beta p} \updownarrow \frac{2\kappa}{1+\beta}$
PSS	$\frac{ad-bc}{(a+c)(b+d)}$	$\kappa p^{\delta} - \frac{p}{1-p}(\beta-\kappa p^{\delta})$	$-\beta p \uparrow 0$	$(\kappa-\beta)p \updownarrow 0$	$\kappa p^{\delta} \downarrow 0$	$\kappa - p(\beta-\kappa) \uparrow \kappa$
OR	$\frac{ad}{bc}$	$\frac{\kappa p^{\delta-1}(1-p(1+\beta-\kappa p^{\delta}))}{(1-\kappa p^{\delta})(\beta-\kappa p^{\delta})}$	$\frac{\kappa p^{\delta-1}}{\beta} \downarrow 0$	$\frac{\kappa(1-p(1+\beta))}{\beta-(\beta+1)\kappa p} \updownarrow \frac{\kappa}{\beta}$	$\frac{\kappa p^{\delta-1}}{\beta} \uparrow \infty$	$\frac{\kappa p^{-1}}{(1-\kappa)(\beta-\kappa)} \uparrow \infty$

DOI: 10.1002/met

Table II. Contingency table of relative frequencies expressed in terms of base rate p, hit rate H, frequency bias B, and total number of trials $n = a + b + c + d$.

		Observed		
		Event	No event	Total
Forecast	Event	$\frac{a}{n} = pH$	$\frac{b}{n} = p(B - H)$	$\frac{a+b}{n} = pB$
	No event	$\frac{c}{n} = p(1 - H)$	$\frac{d}{n} = 1 - p(1 + B - H)$	$\frac{c+d}{n} = 1 - pB$
	Total	$\frac{a+c}{n} = p$	$\frac{b+d}{n} = 1 - p$	1

Therefore, the gradient of the lines in Figure 3(d) is one way that could be used to estimate the exponent $\delta = \eta^{-1} - 1$. As will be shown later, there is a more direct method for finding this exponent.

4. Limiting behaviour of standard verification scores

Table II shows how various verification scores behave in the limit of vanishing base rate. The limiting expressions in columns 4–7 of Table II were obtained by first using Table I to express the scores in terms of bias, hit rate and base rate (leading to column 3 of Table II) and then substituting the asymptotic expressions $(F, H) \sim (\beta p, \kappa p^{1/\eta - 1})$, and $B \sim \beta$ as $p \to 0$.

First it is important to realize that qualitatively different behaviour arises for many of the scores depending on the rate at which the hit rate converges to zero as $p \to 0$:

- Constant hit rate with no convergence to zero ($\eta = 1$), for example, perfect forecasts;
- Slower than linear convergence ($1 > \eta > 1/2$), for example, forecasts that have greater hit rates than random forecasts;
- Linear convergence ($\eta = 1/2$), for example, random forecasts;
- Faster than linear convergence ($1/2 > \eta > 0$), for example, forecasts that have smaller hit rates than random forecasts.

Second, it can be noted from Table II that with the exception of the OR, all the scores tend to the trivial limits of either 0 or 1 as $p \to 0$ when $\eta < 1$. The scores degenerate to trivial limits, which could give misleading interpretations of skill at forecasting rare events. Unless $\eta = 1/2$, the OR tends to either 0 (when $\eta < 1/2$) or ∞ (when $\eta > 1/2$) and so is able to distinguish between worse than random and better than random forecasts for rarer events. Therefore, the OR skill score (Stephenson, 2000) will also therefore tend to either 0 or 1, respectively, depending upon whether the hit rates converge to zero faster or slower than random forecasts.

As first noted by Peirce (1884) and Gilbert (1884), PC is not a good score to use for rare events since it tends to unity for all forecasts (even random ones!). Furthermore, PC only depends on base rate and bias in the asymptotic limit when $\eta < 1$ and so ignores the useful information contained in the hit rate. The HSS, a skill score based on the PC, avoids these two problems but has the disadvantage that it tends to zero for all forecasting systems when $\eta < 1$ as does the threat score (TS) (also known as the critical success index CSI), the equitable threat score (ETS), and the PSS. Furthermore, many of these scores have the non-intuitive property that they actually converge to zero from below (i.e., they go negative and then back up to zero) as the base rate is decreased (e.g., HSS, ETS, and PSS when $\eta \leq 1/2$). Although it has been repeatedly noted that the threat score depends on bias (e.g., Gilbert, 1884; Mason, 1989; Schaefer, 1990), it can be noted from Table II that all the other skill scores depend explicitly on the bias β as $p \to 0$ with the exception of the PSS when $1 > \eta > 1/2$. This means that, in principle, it is possible to modify these scores by hedging. For example, for a fixed hit rate and base rate, forecasting more of the rare events would increase the frequency bias β and thereby reduce the ETS when $1 > \eta > 1/2$.

Several of the scores are related to each other and so provide no new information for rare events. For example, the HSS and the ETS are related to each other by $HSS = 2ETS/(1 + ETS)$ (Schaefer, 1990). By expressing both scores in terms of the hit rate, base rate, and bias, it is easy to demonstrate that this identity holds under all conditions. For vanishing ETS, the relationship becomes $HSS = 2ETS$ when $\eta < 1$ which can be seen in the limits given in Table II. As suggested by Schaefer (1990), ETS also behaves like TS as $p \to 0$, however, this identity only holds when $1 > \eta > 1/2$. That PSS behaves like H, and HSS like $2(1 + TS^{-1})^{-1}$ as $p \to 0$ was noted previously by Doswell *et al.* (1990, Equations (1) and (2)) but these relationships are also only valid when $1 > \eta > 1/2$.

The PSS is identical to the HSS for unbiased forecasts for all base rates.

The scores can also give apparently contradictory behaviour. Note first that the OR can be viewed as a skill score, measuring forecast accuracy relative to random forecasts, for which OR equals 1 (Stephenson, 2000). Now, when $1/2 < \eta < 1$ for example, the OR increases as events become rarer, suggesting more skill, while other skill scores except PC decrease, suggesting less skill! Göber *et al.* (2004) found such opposing behaviour while verifying forecasts of rainfall threshold exceedances. They disregarded scores that showed lower skill at higher

DOI: 10.1002/met

thresholds, explaining away their behaviour as an artefact of their formulation that forces them to decrease with the base rate (Mason, 2003), and attributed the increasing skill suggested by the OR to clearer precursors of heavy rainfall events. Column 6 of Table II shows, however, that the PSS should be expected to decrease while the OR increases when $\eta > 1/2$. How can such contradictions be reconciled? The explanation is two-fold. First, it is important to realize that even scores without explicit dependence on base rate can easily change when the threshold used to define the event is changed. This is because the scores also depend on the conditional probabilities in the contingency table and these will change as the threshold is altered. Second, scores measure different aspects of the joint distribution and so there is nothing, in principle, to prevent two scores from moving in opposite directions as the forecast threshold changes.

The possibility of opposing changes in performance measures with forecast threshold is widely recognized. We should compute a range of scores to obtain a detailed view of how performance changes, or focus on skill scores such as the OR or the area under the ROC curve (e.g. Swets, 1986) that are less sensitive to forecast threshold. We should proceed similarly when considering changes in event threshold: either compute several scores, or consider those that are invariant to event threshold. For example, one could use invariant measures of association such as correlation or the area under the relative operating levels curve (Mason and Graham, 1999). A simple invariant score that is able to measure the association at extreme levels will be presented in the following section.

Figure 4 shows the base-rate dependence of some of the scores defined in Table II for the Met Office forecasts, 6-h-ahead persistence forecasts, and random forecasts for Eskdalemuir. As expected from the previous discussion, PC (Figure 4(a)) tends to one (perfect skill!) at high thresholds for all three forecasting procedures. At thresholds above 10 mm, it becomes progressively more difficult to discern differences in skill using PC. The PSS (Figure 4(b)) can be seen to be about twice the ETS (Figure 4(c)) in agreement with the limit results in Table II. Although these skill scores tend to zero at high threshold, it is still possible, yet difficult, to discern differences in skill between the three different forecasting systems at high thresholds. The Met Office forecasts outperform the persistence forecasts which, in turn, outperform the random forecasts. The difference in skill between forecasting systems is more clearly visible in the logarithm of the OR shown in Figure 4(d). The OR increases for the Met Office forecasts with increasing threshold whereas the persistence forecasts flatten out.

5. A simple non-vanishing measure of association for extremes

The previous section has shown the importance of the exponent η that characterizes how fast the hit rate converges to zero for increasing precipitation thresholds. However, none of the limits of any of the scores provide a simple measure of η. With the exception of PC and OR, the limit values of the scores in Table II could be used to distinguish between $\eta = 1$ (asymptotic dependence) and $\eta < 1$ (asymptotic independence). Only the limit value of OR could then be used to distinguish between the three different asymptotically independent classes: $\eta < 1/2$, $\eta = 1/2$, and $1 > \eta > 1/2$.

Fortunately, there is a simple score that can be used to find η. Recent statistical work has led to the development of an improved measure of extreme dependence for bivariate extreme events. Coles *et al*. (1999) proposed the limit of the following statistic

$$\frac{2\log((a+c)/n)}{\log(a/n)} - 1 \tag{1}$$

as a measure of extremal dependence for bivariate extremes. We will refer to this sample statistic as the EDS and illustrate its benefits for use in forecast verification. From Table I, EDS can be written in terms of base rate and hit rate as follows:

$$\frac{2\log p}{\log\kappa + \eta^{-1}\log p} - 1 \tag{2}$$

Therefore, EDS tends to $2\eta - 1$ rather than 0 in the limit as $p \to 0$ and does not explicitly depend on the bias of the forecasting system. EDS provides a skill score in the range $[-1, 1]$ that can be used to find the hit-rate exponent. EDS takes the value of 1 for perfect forecasts and 0 for random forecasts, and is greater than zero for forecasts that have hit rates that converge slower than those of random forecasts.

Figure 5 shows the EDS calculated for different thresholds for the unbiased Met Office, persistence, and random forecasts. For random unbiased forecasts, $\eta = 1/2$ and $\kappa = 1$ and so EDS is identically zero for all values of base rate. EDS for the Met Office and persistence forecasts can be seen to converge quickly (at thresholds less than 5 mm) to nearly constant values of around 0.65 and 0.35 respectively. These correspond to hit-rate exponents of 0.21 for the Met Office and 0.48 for the persistence forecasts – the hit rates for the Met Office forecasts decrease more slowly than do those for the persistence forecasts as the threshold is increased. The fast convergence and near-constant behaviour of the EDS is of great value for its use in the assessment of different forecast systems. Note that unlike the other skill scores, the EDS can converge to different constants for different forecasting systems. It is impressive how quickly EDS converges even at low thresholds (personal communication, Dr, M. Mittermaier). Approximately 95% confidence intervals on the EDS were estimated as in Coles *et al*. (1999) by adding and subtracting 1.96 times the approximate standard error of the EDS given by

$$S_{EDS} = \sqrt{\frac{H(1-H)}{np}} \times \left(\frac{2\log p}{H(\log pH)^2}\right)$$

DOI: 10.1002/met

This is based on the delta approximation $S_{EDS} \approx S_H \frac{d(EDS)}{dH}$ where EDS is considered to be a non-linear function of the hit rate for any fixed base rate (personal communication, Dr J. Heffernan). More accurate intervals can be estimated by fitting a bivariate extreme-value model to the forecasts and observations (Ferro, 2007).

Association for extreme-rainfall events may at first appear unphysical if one regards rare events as unpredictable outliers. This is not the case, however, since extreme precipitation events are often embedded in storms that, in principle, can be predicted, whereas low-intensity precipitation events often occur almost at random without requiring any mesoscale features as precursors. EDS has demonstrated here that there is dependency between the Met Office forecasts and the observations for more rare events, which is masked by the traditional skill scores that converge to zero as the base rate vanishes.

6. Conclusions

On the basis of some simple assumptions, this study has proposed a simple three-parameter model for how hit rate and bias depend on base rate for vanishingly rare events. The model has then been used to calculate how standard scores will behave for such events. Limit behaviour depends strongly on whether the hit rate exponent, η, is below 1/2, equal to 1/2 (e.g. random forecasts), between 1/2 and 1, or equal to 1 (asymptotically dependent forecasts e.g., perfect forecasts). For $\eta < 1$, with the exception of the OR, all of the standard scores such as the ETS degenerate to the non-informative limit of 0 (or 1 for PC) no matter how good the forecasting system may be. Therefore, we have proposed the use of an alternative measure, referred to as the EDS, that can be used to find η and so can tend to different finite values for different forecasting systems. EDS is easy to calculate from the number of hits and misses. Unlike many of the other scores, EDS does not explicitly depend on the bias in the system for vanishing base rate and so is less prone to improvement by hedging the forecasts.

In this study, we have focussed on the importance of η as measured by EDS. For vanishing base rate, EDS is a measure of association for bivariate extremes (Coles *et al*., 1999; Coles, 2001). As pointed out by Ferro (2007), a forecasting system with larger η will always have a larger hit rate than a forecasting system with a smaller η for sufficiently rare events. However, a forecast user may be interested in either interpolating or extrapolating scores to different base rates in which case κ is also necessary. Ferro (2007) showed how to estimate the two parameters (η, κ) using a bivariate extreme-value model and then demonstrated how these can be used to compare the performance of various forecasting systems. Such an approach based on a rigorous probability model has several advantages. The model can be used to (1) interpolate smoothly between different base rates, (2) extrapolate to even smaller base rates than have been observed and hence make inference about skill for even more extreme events, and (3) provide uncertainty estimates on the skill measures. As demonstrated by Ferro (2007), the model can also be used to optimally recalibrate the forecasts and the parameters of the model can be used to compare different forecasting systems. Such a distribution-oriented approach has many advantages over the simple measure-oriented approach presented in this article.

EDS has the disadvantage that it is based only on the numbers of hits and misses, and so ignores information about false alarms and correct rejections. Therefore, EDS is non-informative about forecast bias, and a forecasting system with a good EDS could be very biased. Therefore, one should present EDS together with the frequency bias as a function of threshold in order to provide a complete summary of forecast performance. This would also allow one to check the assumption made here that the bias of the forecasting system tends to a constant for vanishing base rates. Ferro (2007) suggests how one might extend the extreme-value model to be able to handle biased forecasts. This is an important issue because sometimes it is neither feasible nor desirable to recalibrate the forecasts, as has been done for convenience in the example here and in Ferro (2007). For example, tornado warning systems are assessed on face value yet are well known to forecast too many tornadoes compared to how many tornadoes are actually observed. For such uncalibrated forecasts, the mathematical approach to bias presented here could be of use but further work is required.

This study has shown the difficulties that can occur in verifying even the simplest deterministic approach for forecasting extreme events. Ideally, one should issue probabilities for forecasts of highly uncertain rare events (Murphy, 1991). This raises the difficult and as yet unaddressed issue of which approaches should be used for probability forecasts of extremes. Standard approaches can become non-informative, for example, the Brier score converges to a non-informative zero if one issues reliable probability forecasts for an event with vanishing base rate. Development of verification methods for probability forecasts of extreme events is an important area that clearly requires more attention.

Acknowledgements

We are very grateful to Brian Golding at the UK Met Office for providing the funding that supported Barbara Casati's PhD work that helped lead to this publication. We also wish to thank the following colleagues for useful and stimulating discussions: Andrew Colman, Martin Göber, Jan Heffernan, Ian Jolliffe and Marion Mittermaier.

References

Casati B, Ross G, Stephenson DB. 2004. A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications* **11**: 141–154.

DOI: 10.1002/met

Coles S. 2001. *An Introduction to Statistical Modelling of Extreme Values*. Springer-Verlag: London; 208.

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

Cullen MJP, Davies T, Mawson MH, James JA, Coulter S. 1997. An overview of numerical methods for the next generation of NWP and climate models. In *Numerical Methods in Atmosphere and Ocean Modelling*, The Andre Robert memorial volume, Lin C, Laprise R, Ritchie H (eds). Canadian Meteorological and Oceanographic Society: Ottawa; 425–444.

Davies T, Cullen MJP, Mawson MH, Malcolm AJ. 1999. A new dynamical formulation for the UK meteorological office unified model. In *Proceedings of ECMWF Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling, 7–11 September 1998. European Centre for Medium-Range Weather Forecasts*, Shinfield Park, Reading, Berkshire RG2 9 AX. UK, 202–225.

Donaldson RJ, Dyer RM, Kraus MJ. 1975. An objective evaluator of techniques for predicting severe weather events. *Preprints, Ninth Conference on Severe Local Storms*. American Meteorological Society: Norman, OK; 321–326.

Doswell CA III, Davies-Jones R, Keller DL. 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting* **5**: 576–585.

Ferro CAT. 2007. A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting* **22**: 1089–1100.

Gandin LS, Murphy AH. 1992. Equitable scores for categorical forecasts. *Monthly Weather Review* **120**: 361–370.

Gilbert GK. 1884. Finley's tornado predictions. *American Meteorological Journal* **1**: 166–172.

Göber M, Wilson CA, Milton SF, Stephenson DB. 2004. Fairplay in the verification of operational quantitative precipitation forecasts. *Journal of Hydrology* **288**: 225–236.

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

Mason IB. 1989. Dependence of the Critical Success Index on sample climate and threshold probability. *Australian Meteorological Magazine* **37**: 75–81.

Mason IB. 2003. Binary events. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds). John Wiley and Sons: Chichester, UK; 37–76.

Mason SJ, Graham NE. 1999. Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting* **14**: 713–725.

Murphy AH. 1991. Probabilities, odds, and forecasts of rare events. *Weather and Forecasting* **6**: 302–307.

Palmer WC, Allen RA. 1949. Note on the Accuracy of Forecasts Concerning the Rain Problem. U.S. Weather Bureau manuscript, Washington, DC.

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

Schaefer JT. 1990. The critical success index as an indicator of warning skill. *Weather and Forecasting* **5**: 570–575.

Stephenson DB. 2000. Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting* **15**: 221–232.

Swets JA. 1986. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin* **99**: 100–117.

Webster S, Brown AR, Cameron DR, Jones CP. 2003. Improvements to the representation of orography in the Met Office Unified Model. *Quarterly Journal of the Royal Meteorological Society* **129**: 1989–2010.

Wilson DR, Ballard SP. 1999. A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quarterly Journal of the Royal Meteorological Society* **125**: 1607–1636.

DOI: 10.1002/met