A Probability Model for Verifying Deterministic Forecasts of Extreme Events

CHRISTOPHER A. T. FERRO

School of Engineering, Computing and Mathematics, University of Exeter, Exeter, United Kingdom

(Manuscript received 9 August 2006, in final form 16 January 2007)

ABSTRACT

This article proposes a method for verifying deterministic forecasts of rare, extreme events defined by exceedance above a high threshold. A probability model for the joint distribution of forecasts and observations, and based on extreme-value theory, characterizes the quality of forecasting systems with two key parameters. This enables verification measures to be estimated for any event rarity and helps to reduce the uncertainty associated with direct estimation. Confidence regions are obtained and the method is used to compare daily precipitation forecasts from two operational numerical weather prediction models.

1. Introduction

Forecasting extreme weather events is a vital task and a focus for international research activities such as The Observing-System Research and Predictability Experiment (THORPEX) of the World Meteorological Organization. This article addresses a key component of this research: how should we assess the quality of forecasts of extreme events?

We focus on events that are both extreme and rare. Such events pose at least three difficulties for forecast verification. First, only a small number of events may be observed. This produces large variation in verification measures between datasets and, therefore, large uncertainty about forecast quality. Second, most verification measures necessarily degenerate to trivial values as event rarity increases (D. B. Stephenson et al. 2006, unpublished manuscript). This projects misleading impressions of forecast quality and complicates the discrimination between forecasting systems. Third, events may be observed inaccurately because of short space–time scales and nonevents may even pass unrecorded. The former penalizes good forecasts and the latter leaves some verification measures indeterminate.

We shall address the first two problems, of large uncertainty and degeneracy, with a probability model for the joint distribution of observations and forecasts of extreme events; we disregard the possibilities of inaccurate observations and unrecorded nonevents. We consider events that are observed to occur when a continuous, scalar *observation variable* Y, such as daily precipitation, exceeds an *observation threshold* v, such as 50 mm. This precludes important examples of extreme events such as tornadoes, but includes rare, scalar combinations of multiple variables such as sea surge and wave height. We consider only deterministic forecasts for which the event is forecasted to occur when a continuous, scalar *forecast variable* X exceeds a *forecast threshold* u. We shall not, therefore, verify probabilistic forecasts, which are desirable for rare events (Murphy 1991), but X itself may be a probabilistic forecast (e.g., Toth et al. 2003), an index derived from a probabilistic forecast (Lalaurette 2003), or simply a forecast for the value of Y.

Suppose that we have n pairs of forecast and observation variables, and have chosen thresholds u and v. The resulting n binary forecasts and observations are usually summarized in a contingency table such as Table 1. Summaries of Table 1 that describe the correspondence between the forecasts and observations are measures of forecast performance. Three such measures are the hit rate,

$$h = \frac{a}{a + c}; \tag{1}$$

the critical success index,

$$\mathrm{CSI} = \frac{a}{a + b + c}; \tag{2}$$

Corresponding author address: C. Ferro, School of Engineering, Computing and Mathematics, University of Exeter, Harrison Bldg., North Park Rd., Exeter EX4 4QF, United Kingdom.
E-mail: c.a.t.ferro@exeter.ac.uk

DOI: 10.1175/WAF1036.1

TABLE 1. A contingency table representing the number of n forecast–observation pairs in each cell.

	Observed	Not observed	
Forecasted	a	b	$a + b$
Not forecasted	c	d	$c + d$
	$a + c$	$b + d$	n

and the odds ratio,

$$\mathrm{OR} = \frac{ad}{bc}. \tag{3}$$

These and many other measures can be found in Marzban (1998) and Mason (2003).

We are particularly interested in how performance changes as the observation threshold v increases and events become rarer. If a forecast threshold is specified for each v, then we could construct Table 1 and evaluate performance measures for each v, but this direct approach suffers from the two disadvantages mentioned before of uncertainty and degeneracy, as we now illustrate with some precipitation forecasts.

We compare radar observations of daily precipitation totals with two sets of operational forecasts for a single grid point (52.5°N, 3.5°W) in mid-Wales in the United Kingdom. The forecasting models are the mesoscale Unified Model (UM) of the Met Office and the Aire Limitée Adaptation Dynamique Développement International (ALADIN) model of Météo-France. Forecasts are initialized at 0000 UTC for the subsequent 24 h, with the first on 1 January 2005 and the last on 11 November 2006. There are 72 days with missing data, leaving 607 pairs of forecast and observation variables for both the UM and ALADIN datasets. The data are plotted in Figs. 1 and 2. Both models tend to overestimate small precipitation totals (below 2 mm) and slightly underestimate large totals (above 4 mm). The dependence between forecasts and observations appears strong for both models.

To demonstrate the problems with verifying extreme events, we split the UM data randomly into 10 subsets of approximately 60 pairs each. For each subset, we vary the forecast and observation thresholds so that the *base rate* $(a + c)/n$ is always equal to $(a + b)/n$ and ranges from 0.5 to 0. Matching the two rates yields *calibrated* forecasts, which we shall focus on later; low base rates correspond to rare events. We plot the hit rate [(1)] against the base rate for each subset in Fig. 3. The hit rate becomes more variable as events become rarer, and always converges to either 0 or 1 before becoming indeterminate when the thresholds are so high that $a = c = 0$. This behavior is typical of most verification measures.

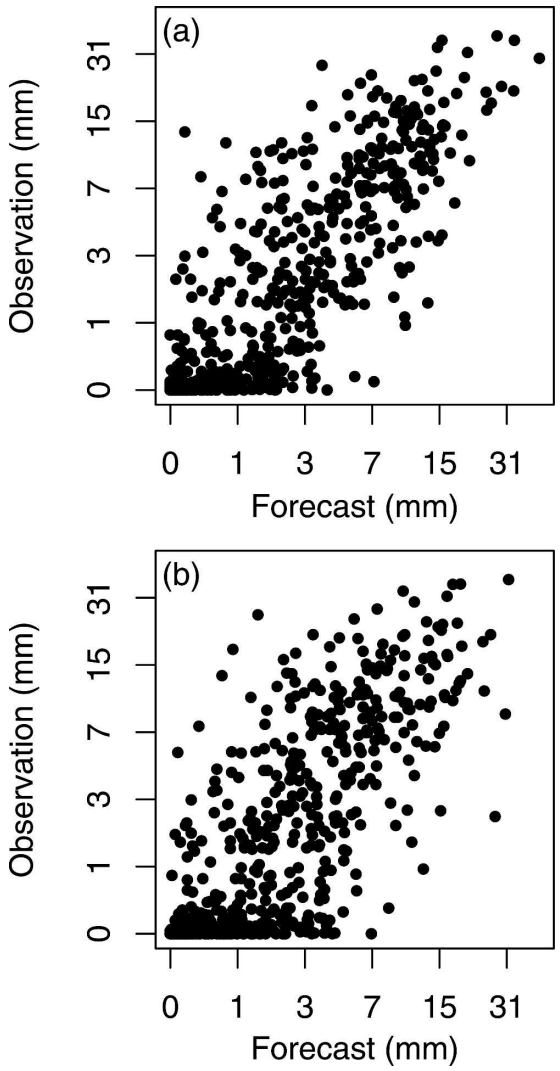

FIG. 1. Observations compared with (a) UM and (b) ALADIN forecasts of 24-h precipitation totals (mm) on a $\log_2$ scale.

We describe our probability model for reducing these problems in section 2 and discuss its interpretation in section 3. Parameter estimation and other practical issues are addressed in sections 4 and 5. We apply our method to the precipitation forecasts in section 6.

2. A model for rare-event verification

a. Approach

If we divide each entry in Table 1 by n, then we obtain the proportion of pairs that falls into each cell. We consider this table of proportions to be an approximation to the ideal table that would be obtained for an infinite number of pairs. The entries in this ideal table are the probabilities that a forecast–observation pair belongs to the different cells and are shown in Table 2.

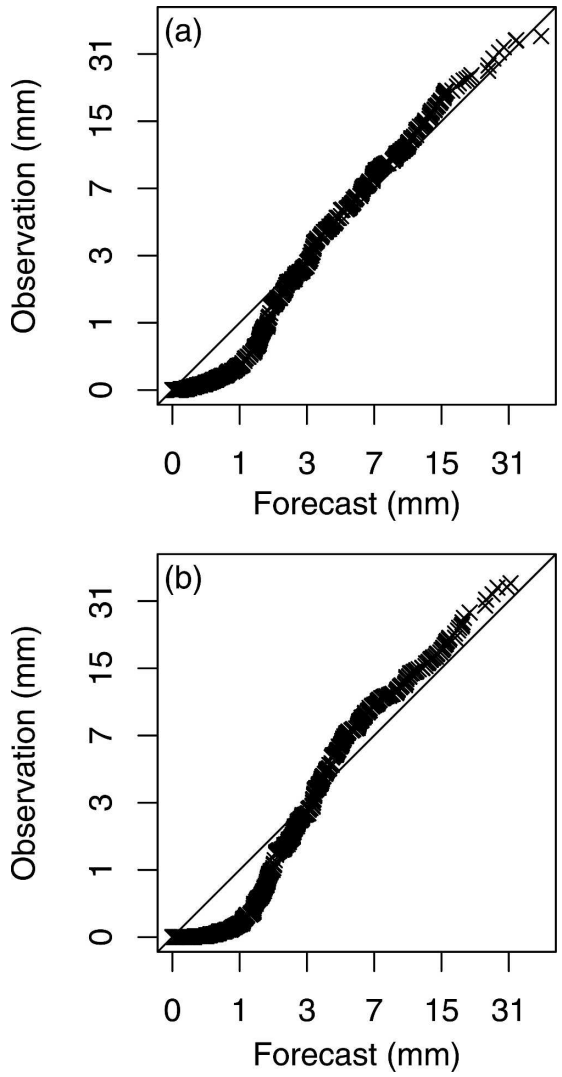

FIG. 2. Ordered observations compared with ordered (a) UM and (b) ALADIN forecasts of 24-h precipitation totals (mm) on a $\log_2$ scale.

The ideal table embodies the true forecast performance that we wish to estimate; for example, the hit rate [(1)] is considered an estimate of the equivalent quantity in Table 2, the conditional probability

$$\Pr(X > u | Y > v) = \frac{\Pr(X > u, Y > v)}{\Pr(Y > v)}.$$

Table 2 must be estimated to determine forecast performance. The table of proportions is one possible estimate, but this leads to the disadvantages illustrated in section 1. We attempt to overcome these disadvantages by applying extreme-value theory (e.g., Beirlant et al. 2004; Coles 2001). Despite relying on only weak assumptions, this theory shows that the probabilities in Table 2 can be modeled as simple functions of u, v, and

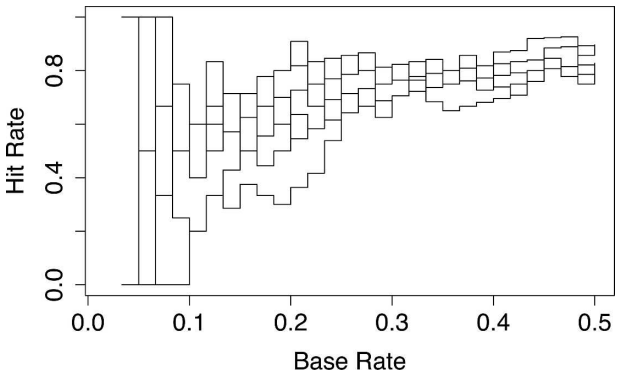

FIG. 3. Hit rate compared with base rate for each of 10 subsets of forecast–observation pairs. The 10 graphs in the plot take discrete values and, thus, undergo step changes and often overlap.

two other parameters. These parameters are independent of u and v when both thresholds are large, so they do not degenerate, but instead measure the fundamental quality of the forecasting system for extreme events by characterizing the relationship between the forecast and observation variables. Furthermore, the parameters do not need to be reestimated for each choice of u and v: they are estimated once only, by fitting the model to a subset of the data. This yields a model version of Table 2 in which u and v can then be varied to obtain estimates of verification measures for different thresholds. Compared with the direct, model-free approach of section 1 in which an ever-diminishing number of threshold exceedances is used to reestimate Table 2 for each pair of thresholds, restricting attention to a two-parameter model and estimating the parameters only once helps to reduce the sampling uncertainty.

Instead of relying on a forecaster to specify the forecast threshold for each observation threshold, we shall consider idealized thresholds that satisfy

$$\Pr(X > u) = p = \Pr(Y > v) \tag{4}$$

for specified base rates p. In other words, u and v will be theoretical upper p quantiles of the distributions of X and Y, and we shall investigate the forecast performance as p decreases to zero over a range of small values. By considering such thresholds, we implicitly calibrate the forecasts by ensuring that the proportion of forecasted events always equals the proportion of

TABLE 2. The ideal contingency table for an infinite sample size.

	Observed	Not observed	
Forecasted	$\Pr(X > u, Y > v)$	$\Pr(X > u, Y \le v)$	$\Pr(X > u)$
Not forecasted	$\Pr(X \le u, Y > v)$	$\Pr(X \le u, Y \le v)$	$\Pr(X \le u)$
	$\Pr(Y > v)$	$\Pr(Y \le v)$	1

observed events. Our analysis will therefore identify the performance that would be achieved by a set of forecasts were they to be perfectly calibrated. This may be considered to be a measure of potential performance that depends on the properties of the forecasting system through the joint distribution of X and Y but not on a forecaster's particular choice of forecast threshold. Our method says nothing about any bias or other discrepancies, such as those revealed by Fig. 2, between the forecast and observation variables. Extensions of our approach to uncalibrated forecasts are discussed briefly in section 7.

The marginal probabilities $\Pr(X > u)$ and $\Pr(Y > v)$, and the joint probability $\Pr(X > u, Y > v)$ determine, by additivity, all of the other cells in Table 2 and thus any derived verification measure. Our choice of thresholds [(4)] means that we know $\Pr(X > u)$ and $\Pr(Y > v)$ for any base rate p, so that the only unknown quantity in Table 2 is $\Pr(X > u, Y > v)$, for which we now develop a model and estimation procedure.

b. Joint probability

Let $F(x) = \Pr(X \le x)$ and $G(y) = \Pr(Y \le y)$ be the distribution functions of X and Y. Then, writing

$$\tilde{X} = -\log[1 - F(X)],$$

$$\tilde{Y} = -\log[1 - G(Y)], \tag{5}$$

and

$$Z = \min\{\tilde{X}, \tilde{Y}\},$$

we can reduce the joint probability to a one-dimensional quantity:

$$\begin{aligned}\Pr(X > u, Y > v) &= \Pr\{\tilde{X} > -\log[1 - F(u)], \tilde{Y} > -\log[1 - G(v)]\} \\ &= \Pr(\tilde{X} > -\log p, \tilde{Y} > -\log p) \\ &= \Pr(Z > -\log p).\end{aligned}$$

Thus, if we can specify a model for the distribution function of Z, which we denote by $H(z) = \Pr(Z \le z)$, then we shall have a model for the joint probability. In fact, we need a model for only the upper tail of H because we are interested in small values of p.

Our data are n pairs of forecast and observation variables denoted by the time sequence $\{(X_t, Y_t): t = 1, \ldots, n\}$, which we shall transform later into realizations $Z_t = \min\{\tilde{X}_t, \tilde{Y}_t\}$ of Z. Results from extreme-value theory (e.g., Coles 2001, chapter 7) indicate that, in many situations, the pattern of points formed by exceedances of the Z_t above a high level w_0 (choice of which is discussed later) is well approximated by a Poisson process. Furthermore, Ledford and Tawn (1996) show that the expected number of points (out of n) exceeding any high level $z \ge w_0$ has the form

$$\exp\left[-\left(\frac{z - \alpha}{\eta}\right)\right], \tag{6}$$

where α is a location parameter and η is a scale parameter, $0 < \eta \le 1$, known as the coefficient of tail dependence.[1] The distribution function of Z is then modeled by the expected proportions:

$$H(z) = 1 - \frac{1}{n}\exp\left[-\left(\frac{z - \alpha}{\eta}\right)\right] \quad \text{for all} \quad z \ge w_0. \tag{7}$$

Writing $\kappa = n^{-1}\exp(\alpha/\eta)$, we have

$$\Pr(Z > -\log p) = 1 - H(-\log p) = \kappa p^{1/\eta} \quad \text{for all } p \le \exp(-w_0), \tag{8}$$

which defines our model for $\Pr(X > u, Y > v)$, and the complete modeled version of Table 2 is given by Table 3.

3. Interpretation of the model

The model parameters η and κ are independent of the base rate and also of the marginal distributions of the forecast and observation variables because Z is formed only after standardizing X and Y by transformation [(5)]. Rather, η and κ describe the strength of the dependence between X and Y at high levels and so measure the quality of forecasts of extreme events. Moreover, η and κ determine Table 3, and therefore forecast performance, for all small base rates, p. These two parameters are thus key measures of the quality of forecasting systems for extreme events.

Consider two forecasting systems: the first with parameters (η_1, κ_1) and the second with parameters (η_2, κ_2). If $\eta_2 > \eta_1$ and $\kappa_2 > \kappa_1$, then

$$\kappa_2 p^{1/\eta_2} > \kappa_1 p^{1/\eta_1} \tag{9}$$

[1] Ledford and Tawn (1996) consider the quantity $-1/\log[1 - \exp(-Z)]$ but our expression [(6)] is obtained after transforming their results to Z. See also Ledford and Tawn (1997), Heffernan (2000), and Segers and Vandewalle (2004).

TABLE 3. The modeled version of Table 2.

	Observed	Not observed	
Forecasted	$\kappa p^{1/\eta}$	$p - \kappa p^{1/\eta}$	p
Not forecasted	$p - \kappa p^{1/\eta}$	$1 - 2p + \kappa p^{1/\eta}$	$1 - p$
	p	$1 - p$	1

for all p. This implies that, for any base rate, the entries on the main diagonal of Table 3 are larger for the second system than for the first, and the off-diagonal entries are smaller. Performance is therefore always superior for the second system. If, on the other hand, $\eta_2 > \eta_1$ but $\kappa_2 < \kappa_1$, then the inequality [(9)] holds if and only if p is less than

$$p^* = \left(\frac{\kappa_2}{\kappa_1}\right)^{\eta_1\eta_2/(\eta_2-\eta_1)}.$$

This implies that the second system is superior only for base rates below p^*.

These relationships can be represented graphically by plotting κ against η. If (η_2, κ_2) lies in the quadrant to the top right of (η_1, κ_1), then the second system is superior for all base rates; if (η_2, κ_2) lies to the top left of (η_1, κ_1), then the second system is superior only for base rates above p^*; if (η_2, κ_2) lies to the bottom right of (η_1, κ_1), then the second system is superior only for base rates below p^*; and if (η_2, κ_2) lies to the bottom left of (η_1, κ_1), then the second system is inferior for all base rates. These quadrants are shown in Fig. 4 for $(\eta_1, \kappa_1) = (0.5, 1)$, the parameter values that correspond to random forecasts.

If a specific base rate p is of interest, then all systems with parameters (η, κ) satisfying the equality

$$\kappa p^{1/\eta} = \kappa_1 p^{1/\eta_1}$$

have the same performance as the system with parameters (η_1, κ_1). The locus of points satisfying this equality corresponds to a curve in the κ–η diagram, and is shown in Fig. 4 for $p = 0.1$. All points to the right of the curve correspond to systems with superior performance for this base rate; all points to the left correspond to systems with inferior performance for this base rate.

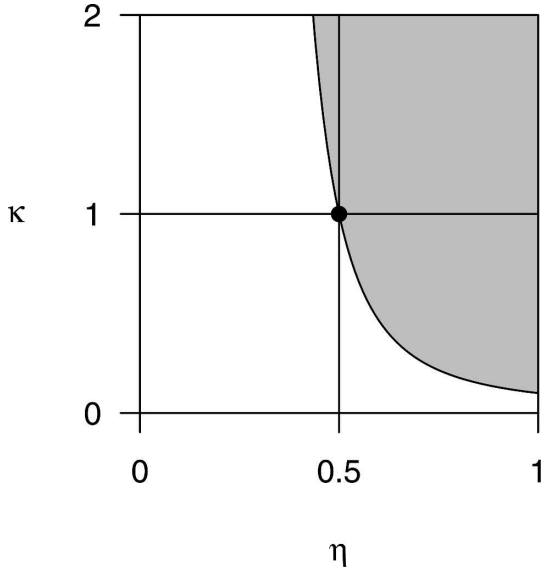

FIG. 4. A κ–η diagram; point (0.5, 1) corresponds to random forecasts. The relative performance of systems with points in each of the four quadrants is as follows: top right, superior for all base rates; top left, superior for base rates above p^*; bottom right, superior for base rates below p^*; and bottom left, inferior for all base rates. The curve locates points with performance equal to random forecasts when the base rate is $p = 0.1$. Systems with points in the shaded region to the right of the curve have superior performance for this base rate; systems with points to the left have inferior performance for this base rate.

4. Parameter estimation

To estimate parameters η and κ, we need to construct realizations $Z_t = \min\{\tilde{X}_t, \tilde{Y}_t\}$ of Z from our data $\{(X_t, Y_t): t = 1, \ldots, n\}$. This involves the transformation [(5)] from (X_t, Y_t) to $(\tilde{X}_t, \tilde{Y}_t)$ via the unknown F and G, which we must therefore estimate. Simple estimators for F and G are the empirical distribution functions:

$$\hat{F}(x) = \frac{1}{n+1}\sum_{t=1}^{n} I(X_t \le x) \quad \text{and}$$
$$\hat{G}(y) = \frac{1}{n+1}\sum_{t=1}^{n} I(Y_t \le y),$$

where $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. The denominator $n + 1$ is used to ensure that the transformed data,

$$\tilde{X}_t = -\log[1 - \hat{F}(X_t)] \quad \text{and}$$
$$\tilde{Y}_t = -\log[1 - \hat{G}(Y_t)], \tag{10}$$

are all finite: denominator n would transform the largest X_t and Y_t to $-\log 0$.

Given our Z_t and a high level w_0, above which our model [(7)] for the distribution function of Z holds, the parameters α and η can be estimated by maximizing the likelihood,

$$\exp\left\{-\exp\left[-\left(\frac{w_0 - \alpha}{\eta}\right)\right]\right\}\prod_{t:Z_t>w_0}\frac{1}{\eta}\exp\left[-\left(\frac{Z_t - \alpha}{\eta}\right)\right], \tag{11}$$

subject to the constraint $0 < \eta \leq 1$. This maximization can be performed analytically, ignoring the singularity at $\eta = 0$, to obtain the estimators

$$\hat{\eta} = \min\left\{1, \frac{1}{m}\sum_{t:Z_t > w_0}(Z_t - w_0)\right\}$$

and $\hat{\alpha} = w_0 + \hat{\eta}\log m$, where m is the number of Z_t exceeding w_0. The estimator for κ is

$$\hat{\kappa} = \frac{1}{n}\exp\left(\frac{\hat{\alpha}}{\hat{\eta}}\right) = \frac{m}{n}\exp\left(\frac{w_0}{\hat{\eta}}\right).$$

Estimates $\hat{\eta}$ and $\hat{\kappa}$ can then be substituted into Table 3 and estimates of verification measures can be obtained for any base rate $p \leq \exp(-w_0)$. For example, the hit rate [(1)] would be $\hat{\kappa}p^{1/\hat{\eta}-1}$.

5. Implementation issues

a. Model assumptions

Our model is inappropriate if either the forecast or observation variables takes only discrete values. We also made three implicit assumptions about the data in section 2: stationarity, serial independence, and some further, extreme-value conditions. We comment on these assumptions and their implications for our method in the remainder of this section.

b. Stationarity

We assumed in section 2 that the probability distribution of the pairs (X_t, Y_t) and, therefore, the entries in Table 3 do not change with time. If the data are instead nonstationary, then naïve application of our model [(8)] may yield a poor fit to the data. In this case, the model is an unreliable source of information about forecast quality. If the model fits well despite the presence of nonstationarity, then the results still provide a gross assessment of the forecast quality, although part of the forecast performance may be due solely to the successful representation of cycles or trends in the forecasting model. This caveat also applies to the direct method described in section 1. One way to overcome the problems with nonstationarity is to focus attention on time periods over which both the climate and the forecasting system are approximately stable. Restricting to a particular season, for example, can reduce the effects of an annual cycle. This remedy is applicable for both our model and the direct approach. An alternative remedy is to extend the model of section 2 to accommodate nonstationarity. This might involve time-varying thresholds and marginal distributions, and parametric models for temporal variations in η and κ. In practice, this requires careful statistical modeling for each new dataset to ensure that nonstationarity is represented accurately and to avoid erroneous identification of changes in forecast quality. Despite the presence of a weak annual cycle in our precipitation data, we shall ignore nonstationarity in this paper in order to demonstrate the basic model of section 2.

c. Serial independence

The modeling approach outlined in section 2 is appropriate if the exceedances of w_0 are serially independent, and so tend to occur singly rather than in clusters. Precisely the same approach may be used, however, even if exceedances cluster. Although such exceedances will deviate from a simple Poisson process, the expected number of exceedances still has the same form [(6)] and the same likelihood [(11)] may be used to estimate the parameters (e.g., Ledford and Tawn 2003). Serial dependence should be accounted for when estimating standard errors and constructing confidence intervals, however, and this is discussed in the appendix. Two other approaches for clustered exceedances, which are not pursued here, are to base estimation on cluster maxima (e.g., Ferro and Segers 2003, 2004) or to model the dependence inside clusters (Smith et al. 1997).

d. Level choice and model fit

To arrive at Table 3, we must choose a level, w_0, above which our model [(7)] is assumed to hold. Three factors should influence this choice. First, if we wish to use the model to estimate forecast performance at a base rate p, then we must choose $w_0 \leq -\log p$. Second, the model parameters are estimated using exceedances of w_0, so lower levels admit more data, increasing the precision of the estimates. Third, the extreme-value theory that motivates the model is more accurate at higher levels, so using higher w_0 yields a more accurate description of forecast performance. This trade-off suggests that we should select the lowest level at which the extreme-value approximations are acceptable.

The extreme-value theory behind our model [(7)] assumes a particular form of decay for the upper tail of the joint distribution function of X and Y (Ledford and Tawn 1996). Distinguishing whether or not this condition is satisfied from knowledge of the physical processes generating the data is probably infeasible because the condition admits a wide class of behaviors. Most probability distributions commonly fitted to meteorological variables satisfy the condition, for example. Checking that our model fits the data is nevertheless an important part of the analysis, and model fit

should guide the selection of w_0. We recommend assessing two features: quality of fit and model stability (Davison and Smith 1990).

1) QUALITY OF FIT

If the model [(7)] holds above w_0, then any excesses $Z_t - w_0$ have the following distribution function:

$$\Pr(Z_t - w_0 \le z | Z_t > w_0) = \frac{H(w_0 + z) - H(w_0)}{1 - H(w_0)} = 1 - \exp(-z/\eta), \quad (12)$$

where $z \ge 0$. Consequently, the transformed excesses $(Z_t - w_0)/\eta$ have an exponential distribution with mean 1, and the model fit can be assessed with probability–probability and quantile–quantile plots. Formal hypothesis tests of fit, such as the Kolmogorov–Smirnov, Anderson–Darling, and Cramer–von Mises tests, are also available (e.g., Coelho et al. 2006, manuscript submitted to *J. Climate*).

2) MODEL STABILITY

If our model [(7)] holds above w_0, then it should also hold above all $w > w_0$. This stability can be checked by plotting against w the parameter estimates obtained by fitting the model to exceedances of w. Any level above which these graphs are approximately constant is a candidate for w_0. Adding confidence intervals to these graphs helps to visualize the sampling uncertainty. Another useful tool is a plot of the sample mean of excesses $Z_t - w$ against w. The mean excess from the exponential distribution [(12)] is η, so the graph should be approximately horizontal for all sufficiently high w.

In addition to these empirical checks, knowledge of the physical processes governing the observations and of the forecasting system should be used to challenge the model assumptions. Although arguing for particular distributional forms is hard, we may still address the more basic assumption that the extreme values to which the model is fitted represent a single population of forecast and observation variables with a common joint distribution function. Tabony (1983) gives several examples for which a set of extreme values represents more than one population, each associated with a different dominant mechanism: hurricane and nonhurricane winds, for example. Multiple populations can also derive from critical values of the measured variable above which either the physical or measurement processes or the forecasting system exhibits qualitatively different behavior: surface temperatures above freezing, for example. Such inhomogeneities may be revealed during model checking by systematic departures from the theoretical patterns of fit and stability. If we aim to extrapolate to very rare events, however, then we must be confident that no new factors will become influential at those higher levels.

e. Summary

Our verification procedure is as follows:

1) Convert the data (X_t, Y_t) to $Z_t = \min\{\tilde{X}_t, \tilde{Y}_t\}$ by transformation [(10)].
2) Estimate the parameters of the model [(8)] by maximizing the likelihood [(11)] over a range of candidate values for w_0.
3) Examine the quality of fit and model stability over this range and select the lowest acceptable value for w_0.
4) Using the $\hat{\eta}$ and $\hat{\kappa}$ obtained for the selected w_0, construct Table 3 for each base rate $p \le \exp(-w_0)$ of interest and compute verification measures.

In the following section we apply this procedure to the precipitation forecasts introduced in section 1. There, we also report confidence regions for the model parameters and verification measures, which are constructed using the bootstrap methodology documented in the appendix.

6. Application to precipitation forecasts

We apply our verification procedure to both the UM and ALADIN precipitation forecasts introduced in section 1. Having applied the transformations [(10)], we need to select a level, w_0, for each dataset above which our model [(7)] is used for the distribution of the Z_t. Plots of parameter estimates and of mean excesses against candidate levels are shown in Figs. 5 and 6. These indicate that levels around the upper 12% quantiles of the Z_t, equal to 1.72 and 1.69 for the UM and ALADIN forecasts, respectively, are feasible. The probability and quantile plots in Figs. 7 and 8 show that the transformed excesses above these levels do approximate exponential distributions with mean 1, with the possible exception of a single outlying value for the ALADIN forecasts. The goodness-of-fit tests mentioned in section 5d all have bootstrap p values exceeding 0.7 for both datasets, indicating no evidence that the exceedances above these levels fail to follow the model [(7)], so we select $w_0 = 1.72$ for the UM forecasts and $w_0 = 1.69$ for the ALADIN forecasts.

The resulting parameter estimates are $\hat{\eta} = 0.75$ and $\hat{\kappa} = 1.18$ for UM, and $\hat{\eta} = 0.72$ and $\hat{\kappa} = 1.25$ for ALADIN. Bootstrapped 90% confidence regions for the parameters are shown in Fig. 9, and we find no significant difference between the parameters for the

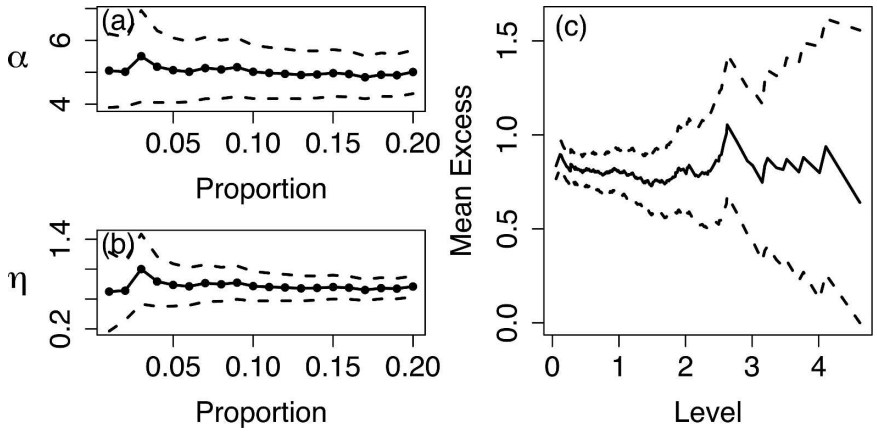

FIG. 5. The UM precipitation forecasts. Estimates (solid) of parameters (a) α and (b) η for the model [(7)] of Z compared with the proportion of data exceeding the candidate values for w_0, and (c) the mean excesses (solid) of Z compared with the candidate levels. Approximate 95% confidence intervals (dashed) are shown.

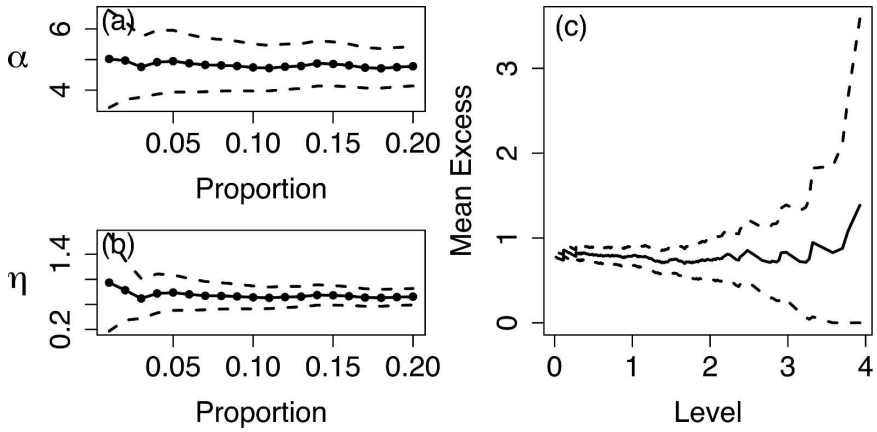

FIG. 6. Same as Fig. 5 but for the ALADIN forecasts.

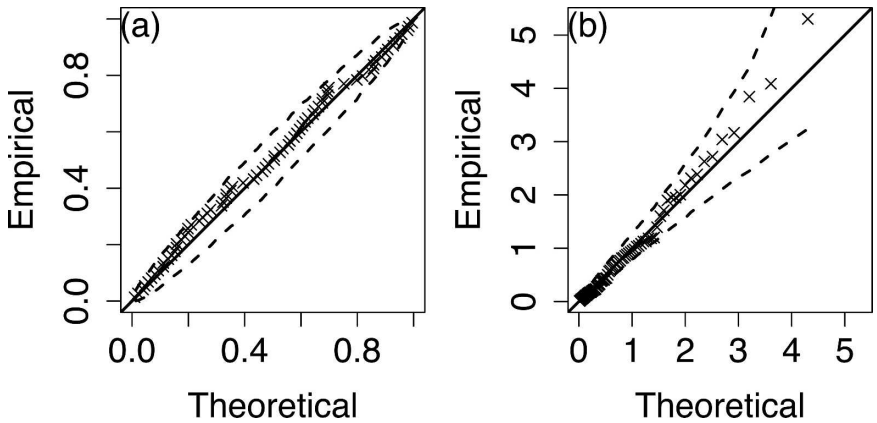

FIG. 7. The UM precipitation forecasts. (a) Probability–probability and (b) quantile–quantile plots comparing the transformed excesses $(Z_t - w_0)/\eta$ to a standard exponential distribution. Approximate 90% confidence intervals (dashed) are shown.

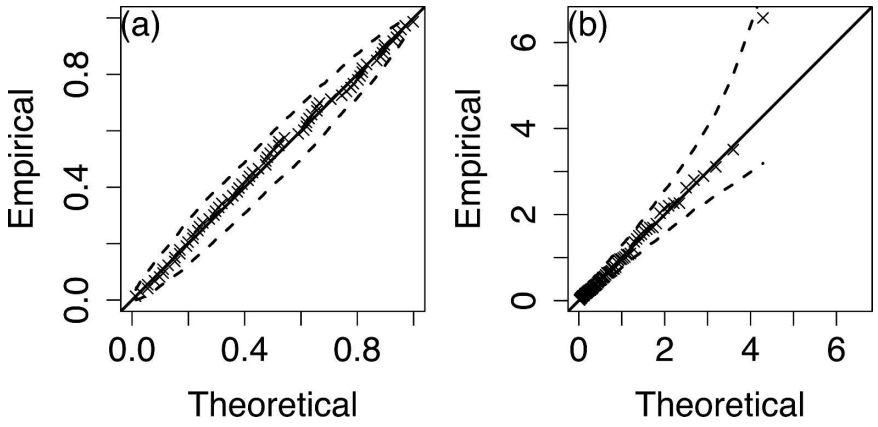

FIG. 8. Same as in Fig. 7 but for the ALADIN forecasts.

two forecasting models. We can also say that the forecast quality of both models is significantly different from those of the random forecasts at the 10% level of significance because neither confidence region contains the point (0.5, 1). In fact, because both sets of parameters lie in the quadrant to the upper right of (0.5, 1), we conclude that both models outperform random forecasts for all base rates below $\exp(-w_0) \approx 0.18$.

We compare estimates of the hit rate [(1)], critical success index [(2)], and odds ratio [(3)] obtained by the direct method discussed in section 1 and by our model for both the UM and ALADIN forecasts. The estimates are plotted in Fig. 10 against the return period (reciprocal base rate) to emphasize the rare events. For return periods up to about 100 days, the model-based estimates follow the direct estimates closely, lying within the latter's 90% confidence intervals, demonstrating that our model is able to capture the changes in the measures with event rarity. For longer return periods, the direct estimates of H, CSI, and OR degenerate to 0 or $-\infty$, while the model-based estimates show a smooth continuation. The confidence intervals for the model-based estimates are also much narrower, revealing the extent to which the model helps to reduce uncertainty. All three measures are slightly lower for ALADIN than for UM at all return periods considered here, but the differences are small relative to the 90% confidence intervals. Several other verification measures (not shown) were also examined and yielded qualitatively similar results.

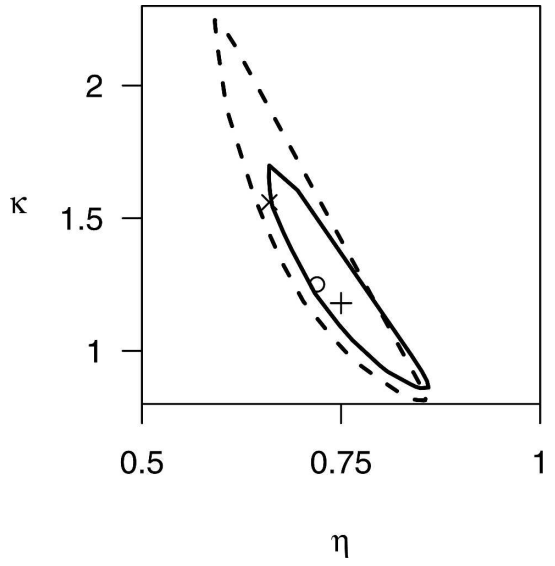

FIG. 9. Estimates of and approximate 90% confidence regions for η and κ for the UM (plus sign, solid) and ALADIN (circle, dashed) precipitation forecasts. The point estimate obtained after removing an outlier from the ALADIN data is also shown (multiplication sign).

The ALADIN dataset contains an outlying Z value caused by a particularly good forecast of the largest observation (Fig. 1b). Removing this outlier changes the estimates of η and κ slightly, but they remain within the 90% confidence region shown in Fig. 9. The model-based estimates of h, CSI, and OR are accordingly changed only slightly, while the direct estimates all become zero for return periods exceeding 70 days (not shown).

7. Discussion

We used a probability model from extreme-value theory to verify forecasts of extreme events. The model helps to reduce the uncertainty in verification measures and identifies two key parameters for describing the quality of forecasting systems for extreme events. Bootstrap confidence regions were proposed for quantifying sampling uncertainty and for comparing the quality of different forecasting systems. The model can accommo-

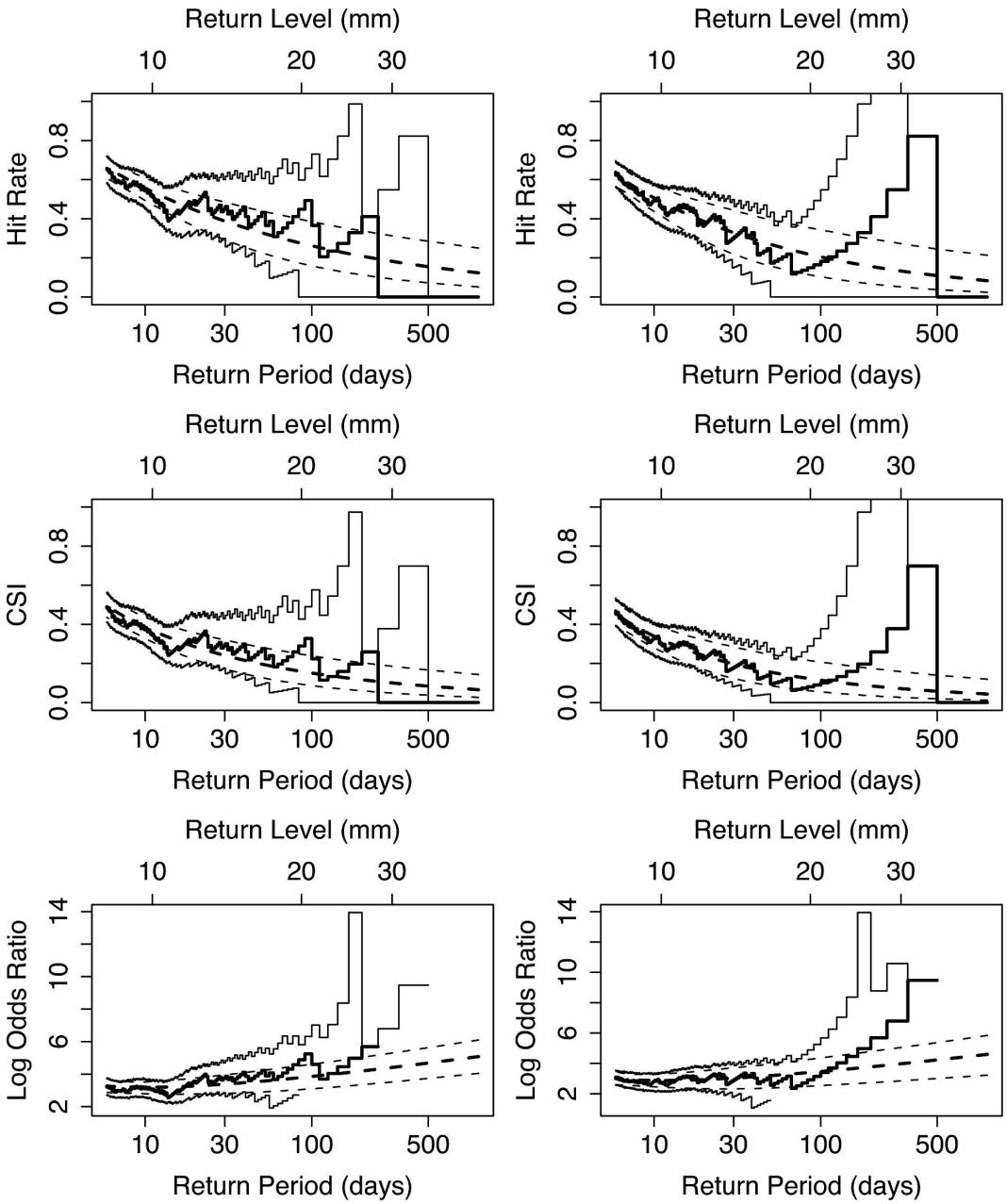

FIG. 10. Direct (solid) and model-based (dashed) estimates (thick) for (left) UM and (right) ALADIN precipitation forecasts of (top) h, (middle) CSI, and (bottom) log OR compared with the return period (reciprocal base rate) and return level (corresponding quantile of the observations) with bootstrap 90% confidence intervals (thin).

date serial dependence and was found to describe accurately the quality of two sets of precipitation forecasts. The model can also be extended to handle nonstationarity, but this is left for future work. Although we defined events by exceedances over high thresholds, a similar approach can be used for deficits below low thresholds as long as events remain rare.

The verification model proposed here is explored further by D. B. Stephenson et al. (2006, unpublished manuscript). They evaluate several verification measures using Table 3, investigate how η and κ determine their limiting behaviors as the base rate decreases, and discuss the implications for verification studies.

Our analysis estimated the performance of only calibrated forecasts [(4)] so that the joint probability in Table 2 could be reduced to a one-dimensional quantity [(8)] that can be modeled with just two parameters. If no restriction is placed on the relationship between

$\Pr(X > u)$ and $\Pr(Y > v)$, then the joint probability cannot be so reduced and extreme-value theory does not provide a characterization in terms of a finite number of parameters, making the model considerably more complex. Models do exist, however; see Segers and Vandewalle (2004, section 9.5) and references therein for further details. Alternatively, the performance of uncalibrated forecasts could be estimated via the relative operating characteristics curve along which the odds ratio is constant and takes the value obtained for the calibrated forecasts (e.g., Swets 1986).

Computer code for the procedures presented in this article and written in the statistical programming language R is available from the author.

Acknowledgments. Conversations with Professor I. T. Jolliffe and Drs. F. J. Doblas-Reyes, M. Göber, M. Mittermaier, and D. B. Stephenson, plus comments from the referees, helped to stimulate and clarify this work. I am grateful to the Met Office and Météo-France for releasing the forecast data. This work was conducted as part of the NCAS–Climate Programme.

APPENDIX

Bootstrap Methodology

We obtain standard errors and confidence intervals for the parameters, η and κ, of our model [(8)] and for the estimated verification measures by employing the following, nonparametric bootstrap resampling scheme. The same approach is used in section 6 to obtain confidence intervals for direct estimates of verification measures. See Davison and Hinkley (1997) for a general introduction to the bootstrap.

1) Resample n pairs $\{(X_t^*, Y_t^*): t = 1, \ldots, n\}$ with replacement from $\{(X_t, Y_t): t = 1, \ldots, n\}$.
2) Reestimate the verification model and recompute any verification measures of interest for these new data.

These steps are repeated N times to yield estimates $\{\hat{\phi}_i^*: i = 1, \ldots, N\}$ of quantities of interest, ϕ, such as model parameters or verification measures for particular base rates. If $\hat{\phi}$ is the estimate from the original data, then an estimate of the standard error of $\hat{\phi}$ is the standard deviation of the $\hat{\phi}_i^*$. A P% confidence interval for ϕ is given by the lower and upper $(50 - P/2)$% quantiles of the $\hat{\phi}_i^*$. Usually, $N = 100$ is preferred for standard errors, and $N = 1000$ for confidence intervals (Davison and Hinkley 1997, pp. 25 and 202). Block resampling (e.g., Wilks 1997) can be used to account for any serial dependence in the data, but this made little difference for our precipitation forecasts.

We also construct two-dimensional confidence regions for the vector (η, κ). A P% confidence region contains P% of the points in the bootstrap sample $\{(\hat{\eta}_i^*, \hat{\kappa}_i^*): i = 1, \ldots, N\}$. The following algorithm (Hall 1987) uses the notion of a convex hull, the smallest convex region containing a set of points, to define unique confidence regions.

1) Find the convex hull of the points $\{(\hat{\eta}_i^*, \hat{\kappa}_i^*): i = 1, \ldots, N\}$.
2) If the proportion of the original N points lying strictly inside the convex hull is greater than P%, then remove from the bootstrap sample those points lying on the boundary of the convex hull.
3) Find the convex hull of the reduced bootstrap sample.

Steps 2 and 3 are repeated until approximately P% of the original points lie inside the convex hull, which is then taken to be the P% confidence region. Green and Silverman (1979) describe an algorithm for finding convex hulls.

REFERENCES

Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels, 2004: *Statistics of Extremes: Theory and Applications*. Wiley and Sons, 490 pp.

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, 208 pp.

Davison, A. C., and R. L. Smith, 1990: Models for exceedances over high thresholds. *J. Roy. Stat. Soc., Ser. B,* **52,** 393–442.

——, and D. V. Hinkley, 1997: *Bootstrap Methods and Their Application*. Cambridge University Press, 592 pp.

Ferro, C. A. T., and J. Segers, 2003: Inference for clusters of extreme values. *J. Roy. Stat. Soc., Ser. B,* **65,** 545–556.

——, and ——, 2004: Extremes of stationary time series. *Statistics of Extremes,* J. Beirlant et al., Eds., Wiley and Sons, 369–428.

Green, P. J., and B. W. Silverman, 1979: Constructing the convex hull of a set of points in the plane. *Comput. J.,* **22,** 262–266.

Hall, P., 1987: On the bootstrap and likelihood-based confidence regions. *Biometrika,* **74,** 481–493.

Heffernan, J. E., 2000: A directory of coefficients of tail dependence. *Extremes,* **3,** 279–290.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. Roy. Meteor. Soc.,* **129,** 3037–3057.

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

——, and ——, 1997: Modelling dependence within joint tail regions. *J. Roy. Stat. Soc., Ser. B,* **59,** 475–499.

——, and ——, 2003: Diagnostics for dependence within time series extremes. *J. Roy. Stat. Soc., Ser. B,* **59,** 521–543.

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

Mason, I. B., 2003: Binary events. *Forecast Verification: A Practi-*

tioner's Guide in Atmospheric Science, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley and Sons, 37–76.

Murphy, A. H., 1991: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting,* **6,** 302–307.

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

Smith, R. L., J. A. Tawn, and S. G. Coles, 1997: Markov chain models for threshold exceedances. *Biometrika,* **84,** 249–268.

Swets, J. A., 1986: Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychol. Bull.,* **99,** 100–117.

Tabony, R. C., 1983: Extreme value analysis in meteorology. *Meteor. Mag.,* **112,** 77–98.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley and Sons, 137–163.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate,* **10,** 65–82.