THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY OF NORTH CAROLINA

MARKOV CHAIN MODELS FOR THRESHOLD EXCEEDANCES

by

R.L. Smith

J.A. Tawn

&

S.G. Coles

June 1993

Mimeo Series #2303

DEPARTMENT OF STATISTICS

Chapel Hill, North Carolina

Markov Chain Models for Threshold Exceedances

by

Richard L. Smith
Dept of Statistics
University of North Carolina
Chapel Hill
NC 27599–3260
U.S.A.

Jonathan A. Tawn
Dept of Mathematics
University of Lancaster
Lancaster LA1 4YW
U.K.

Stuart G. Coles
Dept of Mathematics
The University
University Park
Nottingham NG7 2RD
U.K.

Date *20 June 1993*

Abstract

In recent research on extreme value statistics, there has been an extensive development of threshold methods, first in the univariate case but subsequently in the multivariate case as well. In this paper, we develop an alternative methodology for extreme values of univariate time series, by assuming that the time series is Markovian and using bivariate extreme value theory to suggest appropriate models for the transition distributions. We develop an alternative form of the likelihood representation for threshold methods, and then show how this can be applied to a Markovian time series. A major motivation for developing this kind of theory, in comparison with existing methods based on cluster maxima, is the possibility of calculating probability distributions for extremal functionals more complicated than the maxima or extreme quantiles of the series. In the latter part of the paper, we develop this theme, showing how a theory of compound Poisson limits for additive functionals can be combined with simulation to obtain numerical solutions for problems of practical interest.

1 Introduction

Davison and Smith (1990) proposed a method for extreme value analysis of time series which depends on three essential features. The first is *stationarity* of the time series. The second is the imposition of a *threshold*, u, for distributional modelling. The third feature is the focus on the *peak* value within each cluster of exceedances over the threshold, as a means of handling dependence in the time series. The broad idea is that peak values usually are of prime concern and that these are approximately independent from cluster to cluster, thus avoiding the need for explicit modelling of the temporal structure within clusters.

Within this framework, the process of clusters of exceedances of u is modelled by a Poisson process of rate λ, and the peak excesses are modelled by a *generalized Pareto* distribution, GPD(σ,ξ), with density

$$\frac{1}{\sigma}(1+\xi x/\sigma)^{-1/\xi-1} \tag{1.1}$$

$(x>0, \sigma+\xi x>0)$, in which ξ is a real–valued shape parameter and $\sigma>0$ is a scale parameter. With λ as the mean number of clusters per year, the process of clusters which exceed the level

$u + x$ ($x > 0$) is then a Poisson process with rate

$$\lambda(1 + \xi x/\sigma)_+^{-1/\xi},$$

where $y_+ = \max(y, 0)$. Consequently, the probability that the level $u + x$ is not exceeded during a year is given by

$$\exp\{-\lambda(1 + \xi x/\sigma)_+^{-1/\xi}\}. \tag{1.2}$$

Thus, equation (1.2) is the basis for the modelling and prediction of quantiles of the distribution of the annual maximum. This approach, together with covariate model extensions to handle aspects of non–stationarity (Smith, 1989), is now widely applied in the analysis of extreme values of environmental time series.

This whole procedure is justified by probabilistic theory (Pickands, 1971; Hsing, 1987), which shows that the two dimensional point process, consisting of the times and magnitudes of cluster peaks, converges to a non–homogeneous Poisson process as both the threshold and the length of the series tend to the distributional upper endpoint and infinity respectively.

This paper is concerned with removing the third of the above mentioned features, namely the restriction to peaks as a way of avoiding having to model temporal structure. The existing method has one great advantage – simplicity – but three substantial disadvantages. First there is the problem of identifying independent clusters: misclassified observations lead to biased parameter and standard error estimates. Second, efficiency is lost by discarding non–peak values from the cluster, since the distribution of arbitrary exceedances is also GPD with exactly the same parameters as for the cluster maxima, (Hsing, 1987; Anderson, 1990). Finally, no temporal characteristics of exceedances within a cluster are identified. The latter feature has received some independent attention in the form of estimation of the extremal index (Leadbetter, 1983; O'Brien, 1987), which can be regarded as the reciprocal of the mean number of threshold exceedances per cluster. However, all current estimation procedures for the extremal index also depend on *ad hoc* methods for identifying clusters (Leadbetter *et al.*, 1989; Hsing, 1991; Smith and Weissman, 1994).

Our approach uses a Markov chain to model the temporal dependence of exceedances. This has some similarities with the time series threshold methods of Tong (1990), in particular the feature that different models are assumed for different portions of the sample space, but in constrast to Tong we only attempt to model the behaviour of the process above a high threshold. Much simplicity of the dependence structure then results. For example, for a stationary first order Markov series, $\{Y_n\}$, if $\Pr\{Y_{n+1} > u \mid Y_n > u\} \to 0$, as u tends to the distributional upper endpoint, then exceedances of u will occur independently and so the approach would be identical to the cluster maxima method in this case. As this condition, and higher order extensions of it, apply for all Gaussian series (Leadbetter *et al.*, 1983), the use of Gaussian time series models is ruled out if clustering of extreme levels is observed.

To model the dependence of successive extreme observations, we use multivariate extreme value theory. Threshold methods for multivariate extremes have previously been developed by Coles and Tawn (1991) and by Joe *et al.* (1992). Our approach extends these methods to the temporal dependence case, but in so doing a new version of the associated multivariate threshold likelihood model is proposed. The fitted model is then used to calculate quantities of interest, including quantiles of the annual maximum distribution, the extremal index and the distribution of aggregate excesses. The main technique for this is simulation, using a random walk approximation developed by Smith (1992).

2 Approximate likelihoods for multivariate threshold and Markov models

2.1 A multivariate threshold model

Suppose $(Y_1, \ldots, Y_d)$ is a typical d–variate data point but that, in the spirit of threshold methods, for each j between 1 and d, we observe not Y_j but (δ_j, X_j), where $\delta_j = I\{Y_j \geq u_j\}$, with I the indicator function, and $X_j = \delta_j(Y_j - u_j)$ in relation to a fixed threshold u_j. We want to derive an approximation for the joint distribution of $\{(\delta_j, X_j),\ \ j = 1, \ldots, d\}$. The underlying assumption is that $(Y_1, \ldots, Y_d)$ lies in the domain of attraction of a multivariate extreme value distribution. Suitable conditions, given by Resnick (1987), are as follows:

Let F denote the joint distribution function of $(Y_1, \ldots, Y_d)$, let F_j denote the j'th marginal distribution function, and with $Z_j = 1/(1 - F_j)$ let $Z_j^{\leftarrow}$ denote the inverse function of Z_j. Now define a multivariate distribution F_* by

$$F_*(v_1, \ldots, v_d) = F[Z_1^{\leftarrow}(v_1), \ldots, Z_d^{\leftarrow}(v_d)].$$

Note that if F_i is continuous, then Z_i has the reciprocal of a uniform distribution, so that F_* is the joint distribution function of the transformed vector $(Z_1, \ldots, Z_d)$.

Resnick (1987, Prop. 5.15) showed that F is in the domain of attraction of a multivariate extreme value distribution if and only if, for all $\boldsymbol{v} = (v_1, \ldots, v_d) > 0$,

$$\lim_{t \to \infty} \frac{1 - F_*(tv_1, \ldots, tv_d)}{1 - F_*(t, \ldots, t)} = \frac{-\log G_*(v_1, \ldots, v_d)}{-\log G_*(1, \ldots, 1)}, \tag{2.1}$$

where G_* is a multivariate extreme value distribution function with unit Fréchet margins (i.e. distribution function $\exp(-1/x)$ for $0 < x < \infty$). A general characterization of G_* is the formula

$$G_*(v_1, \ldots, v_d) = \exp\{-V(v_1, \ldots, v_d)\}, \tag{2.2}$$

with

$$V(v_1, \ldots, v_d) = \int_{S_d} \max_{j=1,\ldots,d} (w_j/v_j) dH(\mathbf{w}),$$

where S_d is the unit simplex in $\mathbf{R}^d$ and H is a non–negative dependence measure on S_d satisfying

$$\int_{S_d} w_j dH(\mathbf{w}) = 1, \quad \text{for } j = 1, \ldots, d$$

(Pickands, 1981, and Coles and Tawn, 1991). An example which we shall use in much of this paper (when $d = 2$) is the logistic model

$$G_*(v_1, v_2) = \exp\{-(v_1^{-1/\alpha} + v_2^{-1/\alpha})^\alpha\} \tag{2.3}$$

where $0 < \alpha \leq 1$, which is discussed in Tawn (1988) and Shi *et al.* (1992).

We want to turn these asymptotic results into an explicit approximation for the joint distribution of $\{(\delta_j, X_j), \ j = 1, \ldots, d\}$. A suitable analogy here is the method of Davison and Smith (1990) in the univariate case. Motivated by results which give the GPD as an asymptotic distribution for exceedances over a high threshold, they proceeded as if the GPD were *exact* over a specified threshold. Initially, we make the same assumption for each marginal component, so that for sufficiently large u_j, the marginal distribution of $Y_j - u_j$ given $Y_j > u_j$ is taken as GPD. Thus

$$F_j(y) = 1 - \lambda_j\{1 + \xi_j(y - u_j)/\sigma_j\}_+^{-1/\xi_j}, \quad \text{for } y \geq u_j,$$

where $\lambda_j = 1 - F_j(u_j)$.

By analogy, we also take the limiting result (2.1) as an identity for sufficiently large t. In fact it is more convenient to treat (2.1) as an identity for some fixed $t = t_c$, provided the v_j are sufficiently large; this is clearly an equivalent interpretation. Specifically, taking $t_c = 1$ and $v_j \geq \lambda_j^{-1}$; $j = 1, \ldots, d$ ensures that the GPD approximation is applicable for each marginal component. Our additional assumption is that these levels are also sufficiently high for the asymptotic dependence structure to be a valid approximation through (2.1).

Suppose now, that $\delta_j = 1$; $j = 1, \ldots, d$, so that by the marginal assumptions,

$$Z_j(y) = \lambda_j^{-1}\{1 + \xi_j(y - u_j)/\sigma_j\}_+^{1/\xi_j}, \quad \text{for } y \geq u_j,$$

and hence

$$Z_j^{\leftarrow}(z) = u_j - \sigma_j\xi_j^{-1}\{1 - (\lambda_j z)^{\xi_j}\}, \quad \text{for } z \geq \lambda_j^{-1}.$$

Thus for $t_c = 1$ and $v_j \geq \lambda_j^{-1}$; $j = 1, \ldots, d$,

$$F_*(t_c v_1, \ldots, t_c v_d) = F[u_1 - \sigma_1\xi_1^{-1}(1 - (\lambda_1 v_1)^{\xi_1}), \ldots, u_d - \sigma_d\xi_d^{-1}(1 - (\lambda_d v_d)^{\xi_d})].$$

Treating (2.1) as an identity in this case then gives

$$1 - F[u_1 - \sigma_1\xi_1^{-1}(1 - (\lambda_1 v_1)^{\xi_1}), \ldots, u_d - \sigma_d\xi_d^{-1}(1 - (\lambda_d v_d)^{\xi_d})] = KV(v_1, \ldots, v_d) \tag{2.4}$$

where

$$K = \frac{1 - F(1, \ldots, 1)}{V(1, \ldots, 1)}.$$

The constant K is determined by the marginal distributions: setting $v_1 = \lambda_1^{-1}$ and $v_2 = \ldots = v_d = \infty$, we get from (2.4)

$$\lambda_1 = KV(\lambda_1^{-1}, \infty, \ldots, \infty) = K\lambda_1, \quad \text{i.e.} \quad K = 1$$

since $V(\lambda_1^{-1}, \infty, \ldots, \infty) = -\log G_*(\lambda_1^{-1}, \infty, \ldots, \infty) = \lambda_1$, from the fact that the margins of G_* are unit Fréchet. Translated back into the original coordinates, expression (2.4) gives the desired approximation:

$$F(y_1, \ldots, y_d) = 1 - V[\lambda_1^{-1}\{1 + \xi_1(y_1 - u_1)/\sigma_1\}_+^{1/\xi_1}, \ldots, \lambda_d^{-1}\{1 + \xi_d(y_d - u_d)/\sigma_d\}_+^{1/\xi_d}], \quad (2.5)$$

valid for $y_1 \geq u_1, \ldots, y_d \geq u_d$.

For example, in the $d = 2$ case with the logistic model given by expression (2.3), equation (2.5) becomes, for $y_1 \geq u_1, y_2 \geq u_2$,

$$F(y_1, y_2) = 1 - \left[\lambda_1^{1/\alpha}\{1 + \xi_1(y_1 - u_1)/\sigma_1\}_+^{-1/\alpha\xi_1} + \lambda_2^{1/\alpha}\{1 + \xi_2(y_2 - u_2)/\sigma_2\}_+^{-1/\alpha\xi_2}\right]^\alpha. \quad (2.6)$$

To obtain the joint density of $\{(\delta_j, X_j),\ j = 1, \ldots, d\}$ we use (2.5) directly in the case $\delta_1 = \ldots = \delta_d = 1$, and proceed by inclusion–exclusion relations in other cases. For example, with $d = 2$ we have

$$\begin{aligned} \Pr\{\delta_1 = 0, \delta_2 = 1, X_2 > x_2\} &= F(u_1, \infty) - F(u_1, u_2 + x_2) \\ &= V\{\lambda_1^{-1}, \lambda_2^{-1}(1 + \xi_2 x_2/\sigma_2)_+^{1/\xi_2}\} - \lambda_1, \end{aligned} \quad (2.7)$$

$$\Pr\{\delta_1 = 1, \delta_2 = 0, X_1 > x_1\} = V\{\lambda_1^{-1}(1 + \xi_1 x_1/\sigma_1)_+^{1/\xi_1}, \lambda_2^{-1}\} - \lambda_2, \quad (2.8)$$

$$\Pr\{\delta_1 = 0, \delta_2 = 0\} = F(u_1, u_2) = 1 - V(\lambda_1^{-1}, \lambda_2^{-1}). \quad (2.9)$$

Hence if we define $V_1(x, y) = -\frac{\partial}{\partial x}\{V(x, y)\}$, $V_2(x, y) = -\frac{\partial}{\partial y}\{V(x, y)\}$, and $V_{12}(x, y) = -\frac{\partial^2}{\partial x \partial y}\{V(x, y)\}$, the joint density of $(\delta_1, X_1, \delta_2, X_2)$ is given by

$$\begin{cases} (\lambda_1\lambda_2\sigma_1\sigma_2)^{-1}[t_1(x_1)]^{1-\xi_1}[t_2(x_2)]^{1-\xi_2}V_{12}\{\lambda_1^{-1}t_1(x_1), \lambda_2^{-1}t_2(x_2)\} & \text{if } \delta_1 = \delta_2 = 1, \\ (\lambda_2\sigma_2)^{-1}[t_2(x_2)]^{1-\xi_2}V_2\{\lambda_1^{-1}, \lambda_2^{-1}t_2(x_2)\} & \text{if } \delta_1 = 0, \delta_2 = 1, \\ (\lambda_1\sigma_1)^{-1}[t_1(x_1)]^{1-\xi_1}V_1\{\lambda_1^{-1}t_1(x_1), \lambda_2^{-1}\} & \text{if } \delta_1 = 1, \delta_2 = 0, \\ 1 - V(\lambda_1^{-1}, \lambda_2^{-1}) & \text{if } \delta_1 = \delta_2 = 0, \end{cases} \quad (2.10)$$

where

$$t_j(x_j) = (1 + \xi_j x_j/\sigma_j)_+^{1/\xi_j}, \quad j = 1, 2.$$

The procedure for $d > 2$ is similar, with each possible combination of $(\delta_1, \ldots, \delta_d)$ resulting in a different form of contribution to the likelihood function, constructed by taking suitable sums and differences of (2.5). We do not go into the details of this here, but return to this topic in Section 5.

2.2 Removing the marginal constraints

In some circumstances it may be over–restrictive to require that $v_j \geq \lambda_j^{-1}; j = 1, \ldots, d$ for the validity of (2.1) to be used as an identity since if (2.1) provides a good approximation over a larger region of the upper joint tail the proposed likelihood procedure will be inefficient. In principle, the functional form (2.5) could still hold over the larger region, but there is a complication: (2.5) assumes that the GPD holds for each marginal distribution within the range of validity of (2.5). This is only reasonable when we impose a threshold on *each* Y_j. However, the formula corresponding to (2.5) without assuming the GPD tail approximation is easily derived as

$$F(y_1, \ldots, y_d) = 1 - V\{[1 - F_1(y_1)]^{-1}, \ldots, [1 - F_d(y_d)]^{-1}\}. \qquad (2.11)$$

The asymptotic nature of (2.1) means we still require some threshold restriction on the vector $(y_1, \ldots, y_d)$. Suitable conditions to ensure the observation $(y_1, \ldots, y_d)$ is sufficiently far into the joint tail could be $\max(y_j - u_j) > 0$ or $\sum(y_j - u_j) > 0$.

In practice, to exploit (2.11), a hybrid approach to the marginal modelling is needed: the GPD model is appropriate for the j'th component when $y_j > u_j$, but some alternative distribution is needed when $y_j < u_j$. For example, it would be possible to assume a different parametric model for the marginal distribution in the range $y_j < u_j$, or alternatively to use the empirical distribution function and treat this as known in all subsequent inferences. This leads to an alternative form of the likelihood function. The advantage of this approach is that the approximation extends to a larger portion of the sample space, thus in principle making possible more sensitive inferences, and avoiding the inclusion–exclusion arguments which could become very restrictive when d is large. The disadvantage of this method, compared with that based on (2.5), is that some modelling of the marginal distribution below the threshold is needed.

Thus we have two forms of approximate likelihood, one using (2.5) and based only on observations which exceed the threshold, the other using (2.11) and requiring some assumptions about the marginal distributions below thresholds. The second method is in fact much closer to the previous multivariate threshold methods proposed by Coles and Tawn (1991), and by Joe *et al.* (1992). In those papers an approximation of the form (2.11) was made with $1 - F_j(u_j)$ replaced by $-\log F_j(y_j)$, $j = 1, \ldots, d$ and the region of validity of the approximation taken as $\max_j(y_j - u_j) > 0$. Furthermore, if the contribution to the likelihood, $[1 - V(\lambda_1^{-1}, \lambda_2^{-1})]^{n_{1,1}}$, by the $n_{1,1}$ observations with $\delta_1 = \delta_2 = 0$ is approximated by $\exp\{-nV(\lambda_1^{-1}, \lambda_2^{-1})\}$, the likelihood is exactly that of Coles and Tawn (1991, equation 5.4) which was based on a non–homogeneous Poisson process approximation. For the remainder of this paper we confine attention to the likelihood approximation (2.5).

2.3 Markov chain models

We now apply these ideas, in the case $d = 2$, to the model which is the central theme of the present paper; namely a stationary Markov chain, $\{Y_n\}$, with continuous state space. Such a chain is characterized by the joint distribution of (Y_n, Y_{n+1}), and we assume this to be in the domain of attraction of a bivariate extreme value distribution function. The joint density of $(Y_1, \ldots, Y_N)$ is given by

$$p(y_1) \prod_{j=2}^{N} p(y_j \mid y_{j-1}) \tag{2.12}$$

where $p(.)$ denotes the marginal (stationary) density and $p(y_j \mid y_{j-1})$ is the conditional density, evaluated at y_j, of Y_j given $Y_{j-1} = y_{j-1}$.

Suppose, now, we want an approximation to (2.12) based only on the high threshold exceedances. Because of stationarity, we can assume that the threshold u, the exceedance probability λ, and the GPD parameters (σ, k), are common to all observations. We therefore write $\delta_j = I\{Y_j > u\}, X_j = \delta_j(Y_j - u)$ and consider a likelihood function based only on $\{(\delta_j, X_j), j = 1, \ldots, N\}$. By rewriting (2.12) as

$$\prod_{j=2}^{N} p(y_{j-1}, y_j) / \prod_{j=2}^{N-1} p(y_j), \tag{2.13}$$

where $p(y_{j-1}, y_j)$ is the joint density of two consecutive observations, the problem reduces to that considered in Section 2.1. Specifically, the numerator of (2.13) is approximated as a product of terms of the form (2.10), while the denominator is replaced by the corresponding univariate approximation: $p(y_j) = \sigma^{-1}\lambda\{1 + \xi(y_j - u)/\sigma\}^{-1/\xi-1}$ if $y_j > u$ and by $p(y_j) = 1 - \lambda$ otherwise. Once again, it is also possible to consider a more general approximation based on (2.11) if we make some assumptions about the marginal distribution below the threshold, and this leads to a second form of approximate likelihood function. The two methods differ only in the way the numerator of (2.13) is modelled, not the denominator.

In principle, it might be possible to justify any of these approximate likelihoods by establishing standard properties such as consistency and asymptotic normality of the resulting estimators. Such a theory would, however, require detailed assumptions on the convergence rates to the separate marginal and dependence components of the bivariate extreme limit. This seems certain to involve considerable technical complications, and therefore we avoid further discussion of it here.

3 Examples of functionals of extreme events

In Section 1 it was indicated that there are many functionals of the values within each independent extreme event, other than the cluster maximum, which are of interest in applications. Here we present a general class of functional forms which is broad enough to include examples which address both general and problem–specific issues. Furthermore, this functional class is amenable to asymptotic treatment.

For the process $\{Y_n\}$ and the threshold u, the class of functionals we consider have the form

$$W_n(u) = \sum_{j=1}^{n-m+1} g\{(Y_j - u)_+, \ldots, (Y_{j+m-1} - u)_+\} \tag{3.1}$$

where $g : \mathbf{R}_+^m \to \mathbf{R}$ is a real-valued function for a fixed $m \geq 1$, and $g(0, \ldots, 0) = 0$. The structure of (3.1) confines attention to quantities that depend on the high-level exceedances of the process. Some examples are given as follows

(i) Let $m = 1$ and, for some $z > 0$, let $g(x) = 1$ if $x > z$; $g(x) = 0$ otherwise. Then $W_n(u)$ counts the number of exceedances of $u+z$ over the time period $1 \leq j \leq n$. In particular, $\Pr\{W_n(u) = 0\}$ is the probability that the maximum of the process up to time n does not exceed $u + z$.

(ii) Let $m = 2$ and, for some $z \geq 0$, let $g(x_j, x_{j+1}) = 1$ if $x_{j+1} > z$ and $x_j < z$; $g(.) = 0$ otherwise. Then $W_n(u)$ is the number of upcrossings by the process of the level $u + z$ in the period $1 \leq j \leq n$.

(iii) Let $m = 3$ and $g(x_j, x_{j+1}, x_{j+2}) = 1$ if, for some $z \geq 0$, $x_{j+1} > z$ and $x_{j+1} > \max(x_j, x_{j+2})$; $g(.) = 0$ otherwise. Then $W_n(u)$ is the number of local maxima of the process which exceed $u + z$ in the period $1 \leq j \leq n$.

(iv) Let $m > 1$, and for some $z > 0$, let $g(x_j, \ldots, x_{j+m-1}) = 1$ if $x_j > z, \ldots, x_{j+m-1} > z$; $g(.) = 0$ otherwise. Then $W_n(u)$ is the number of times, in the period $1 \leq j \leq n$, that there are m consecutive exceedances of $u + z$, counting overlaps. For example, a 'heatwave day' might be defined as a day on which the temperature is above $u + z$ and the previous $m - 1$ days have also had temperatures above $u + z$. Then $W_n(u)$ is the number of heatwave days that occur between day m and day n. Clearly, more complicated definitions of a heatwave, analogous to those discussed by Barnett and Lewis (1967), could also be incorporated within this structure.

(v) Let $m = 1$ and, for some $z > 0$, $g(x) = x - z$ if $x > z$; $g(x) = 0$ otherwise. Then $W_n(u)$ is the cumulative total of all excesses over $u + z$. In the hydrological context, this is a measure of the volume of overflow, which is sometimes taken directly as the main variable of interest (Anderson and Dancy, 1992).

(vi) The sea level at time t is the sum of a deterministic tidal component, T_t, and a stationary stochastic surge component, Y_t. The tidal series is periodic, so we take $T_1, \ldots, T_{n'}$ to be the values within a cycle. For simplicity, we take the surge to be independent of the tidal level, the dependent case following immediately. Tawn (1992) proposed obtaining the distribution of extreme sea levels by first obtaining the distribution of extreme surge levels and then modifying this to incorporate the tidal series. Given the upper tail of the marginal surge distribution, the dependence structure within extreme surge events, and the tidal series, the probability that the extreme sea level is less than some level, u^* say, can be calculated directly.

Specifically, we take the surge threshold to be u and consider only the case when $u^* > u + \max(T_1, \ldots, T_{n'})$. In this case let $g^*(x_j, i) = I\{u + x_j + T_{i+j-1} > u^*\}$ and $W_n(u, i) = \sum_{j=1}^{n} g^*(x_j, i)$. Then $W_n(u, i)$ is the total number of exceedances of u^* given that the first tide was T_i. Now the expected number, with respect to the timing of the first tide, of sea level exceedances of u^* is given by $W_n(u) = \sum_{i=1}^{n'} W_n(u, i)/n'$, so $\Pr\{W_n(u) = 0\}$ is the probability that the maximum of the sea level process at time n is less than u^*. Provided the temporal dependence model is suitable, this removes the requirement to estimate an additional parameter in Tawn's approach.

4 Extreme event theory for Markov chains

In this section we discuss how to calculate the limiting distribution of the class of functionals, given by (3.1), of the high–level exceedances of a Markov chain. First, we prove a limit theorem which shows that such functionals, under suitable asymptotic conditions, have a limiting compound Poisson distribution. Then we describe a simulation technique to evaluate the compounding distribution of the compound Poisson limit. The methods rely heavily on previous results for extreme value theory in Markov chains (O'Brien, 1987; Smith, 1992) as well as general theory leading to compound Poisson limits for extreme events (e.g. Hsing *et al.*, 1988). Throughout we assume the process $\{Y_n\}$ is a stationary aperiodic Harris chain with marginal distribution function F. Loosely, a Harris chain in continuous state space corresponds to an irreducible chain in discrete state space. Asmussen (1987) has precise definitions and a good discussion.

4.1 Asymptotic theory for functionals

For a functional $W_n(u)$, in the class defined by (3.1), we develop a limit theorem for the distribution of $W_n(u)$ when n and u are both large. The basic idea is that the high–level exceedances form clusters, and that the distribution in time of the clusters can be represented asymptotically as a Poisson process. Thus the limiting behaviour of $W_n(u)$ is that there are a Poisson–distributed number of independent clusters, and $W_n(u)$ is a sum of contributions from individual clusters.

Any aperiodic Harris chain is strong mixing (O'Brien, 1987) so, by a small extension of the usual definition of strong mixing, we can find a function $\phi(p)$ such that, if $\tilde{Z}_1$ and $\tilde{Z}_2$ are complex–valued random variables satisfying $|\tilde{Z}_i| \leq 1$ $(i = 1, 2)$, and such that $\tilde{Z}_1$ is a function of $\{Y_j,\ -\infty < j \leq n\}$ and $\tilde{Z}_2$ a function of $\{Y_j,\ n + p \leq j < \infty\}$, then

$$|\mathrm{E}(\tilde{Z}_1\tilde{Z}_2) - \mathrm{E}(\tilde{Z}_1)\mathrm{E}(\tilde{Z}_2)| \leq \phi(p), \tag{4.1}$$

and $\phi(p) \downarrow 0$ as $p \to \infty$.

Given such a function ϕ, we can always define a sequence $\{p_n, \ n \geq 1\}$ such that

$$\frac{p_n}{n} \to 0, \quad \frac{n\phi(p_n)}{p_n} \to 0. \tag{4.2}$$

Let $\{u_n, \ n \geq 1\}$ denote an increasing sequence of threshold values such that

$$\lim_{n\to\infty} n\{1 - F(u_n)\} = \tau \qquad (0 < \tau < \infty). \tag{4.3}$$

We also assume

$$\lim_{p\to\infty} \lim_{n\to\infty} \sum_{i=p}^{p_n} \Pr\{Y_i > u_n \mid Y_1 > u_n\} = 0. \tag{4.4}$$

Condition (4.4) was also assumed in Smith (1992) where, following O'Brien (1974, 1987), it was used to justify defining the extremal index as

$$\theta = \lim_{p\to\infty} \theta_p, \text{ with } \theta_p = \lim_{u\to\infty} \theta_p(u) = \lim_{u\to\infty} \Pr\{Y_i \leq u, \ 2 \leq i \leq p \mid Y_1 > u\} \tag{4.5}$$

and $\theta_1(u) = 1$. Finally, we define

$$W^*_{n,p} = \sum_{j=1}^{p-m+1} g\{(Y_j - u_n)_+, \ldots, (Y_{j+m-1} - u_n)_+\}. \tag{4.6}$$

We assume

$$\mathrm{E}\{W^*_{n,p} \mid \max(Y_1, \ldots, Y_p) > u_n\} \leq M < \infty, \tag{4.7}$$

and that there exists a random variable W^*, which we term the cluster functional, such that, for all $w > 0$, as $n \to \infty$

$$\Pr\{W^*_{n,p_n} \leq w \mid \max(Y_1, \ldots, Y_{p_n}) > u_n\} \to \Pr\{W^* \leq w\}. \tag{4.8}$$

We then have:–

Theorem 1: As $n \to \infty$, the distribution of $W_n(u_n)$ converges to that of $\sum_{i=1}^{K} W_i^*$, where K is a Poisson random variable of mean $\theta\tau$, and $W_1^*, W_2^*, \ldots$ are independent (of each other and of K) random variables with the same distribution as W^* in (4.8). Thus,

$$\Pr\{\max(W_1^*, \ldots, W_K^*) \leq w\} = \exp\{-\tau\theta[1 - \Pr(W^* \leq w)]\}.$$

This result is proved in the Appendix.

To apply Theorem 1, it therefore remains to estimate both θ and the distribution of W^*, using the model as fitted by the methods of Section 2. Our methodology for this is simulation.

The extremal index θ may be defined by (4.5), and it also follows under condition (4.4) that (4.8) is equivalent to

$$\Pr\{W^* \leq w\} = \lim_{p\to\infty} \lim_{u\to\infty} \Pr\{W^*_{n,p} \leq w \quad | \quad \max(Y_1, \ldots, Y_p) > u\}. \tag{4.9}$$

The proof of (4.9) is also given in the Appendix.

Two other results are needed before we can define our simulation scheme. First, from (4.5), it follows that

$$\begin{aligned} \Pr\{\max(Y_1, \ldots, Y_r) > u\} &= \sum_{j=1}^{r} \Pr\{Y_j > u, \max_{i=j+1,\ldots,r} Y_i \leq u\} \\ &= \{1 - F(u)\} \sum_{j=1}^{r} \theta_{r-j+1}(u). \end{aligned}$$

Furthermore, since $\theta_p(u) \to \theta_p$ by (4.5), we have

$$\lim_{u\to\infty} \frac{\Pr\{\max(Y_1, \ldots, Y_r) > u\}}{1 - F(u)} = \sum_{j=1}^{r} \theta_j, \tag{4.10}$$

the real point being that the limit exists. Hence, for fixed r, the conditional limiting distribution of $\max(Y_1, \ldots Y_r) - u$, given that this quantity is positive, is the same as that for a single value $X_i = (Y_i - u)_+$, given X_i is positive. This is because, from (4.10) and (1.1), for large u, and $x > 0$

$$\frac{\Pr\{\max(Y_1, \ldots, Y_r) > x + u\}}{\Pr\{\max(Y_1, \ldots, Y_r) > u\}} \sim \frac{1 - F(u + x)}{1 - F(u)} \approx (1 + \xi x/\sigma)_+^{-1/\xi} \tag{4.11}$$

which is independent of r. This result is known (Hsing, 1987; Anderson, 1990), but is repeated here to make clear the connections with our other results.

The second result we need is the property that, under the assumptions we have made, a Markov chain in the tails 'looks like' a transformed random walk. Initially suppose $1 - F(y) \sim C\exp(-y)$ as $y \to \infty$. Under the assumption that (Y_1, Y_2) lies in the domain of attraction of a bivariate extreme value distribution, the limiting distribution function

$$H_F(x) = \lim_{z\to\infty} \Pr\{Y_2 \leq z + x \quad | \quad Y_1 = z\}$$

exists (Smith, 1992). Thus if we are given $Y_j = y_j$, for some known $y_j > u$, and want to simulate the distribution of $Y_{j+1}, \ldots, Y_r$, then it suffices to treat this as a random walk with step length having the distribution function H_F. By a trivial extension of the same result, if we want to simulate $Y_1, \ldots, Y_{j-1}$, from this same Y_j value, we can also use a random walk approximation, with step length having distribution function, H_B, given by

$$H_B(x) = \lim_{u\to\infty} \Pr\{Y_1 \leq u + x \quad | \quad Y_2 = u\}.$$

Here $H_B \neq H_F$ unless the Markov chain is time-reversible; that is unless in (2.2) $V(x, y) = V(y, x)$ for all x and y.

More generally, if the marginal distribution function is of the form

$$1 - F(u + x) \sim [1 - F(u)](1 + \xi x/\sigma)_+^{-1/\xi} \quad (x > 0)$$

for large u then it is possible to proceed by pointwise transformation to obtain, for given Y_n, that

$$\xi^{-1}\log\{[\xi(Y_{n+1}-u)+\sigma]/[\xi(Y_n-u)+\sigma]\} \text{ and } \xi^{-1}\log\{[\xi(Y_{n-1}-u)+\sigma]/[\xi(Y_n-u)+\sigma]\} \quad (4.12)$$

are independent random variables from distributions H_F and H_B respectively. The $\sigma = 1$ and $\xi = 0$ case for (4.12), taken as the limit as $\xi \to 0$, is the random walk described above. An alternative to transformation is to use a scheme described by Perfekt (1991) which effectively adapts the random walk approximation to any of the three domains of attraction of univariate extreme value theory.

The proposed simulation scheme for extreme events which exceed a high threshold is therefore as follows:-

1. Choose the function g, with the associated m, of interest. Select r, the simulation event length.
2. Generate the cluster maximum $M_r = \max(Y_1, \ldots, Y_r)$ given $M_r > u$. By (4.10), the limiting conditional distribution is GPD, so this step involves simulating a realization from a GPD with the required parameters.
3. Let J denote the index for which $Y_J = M_r$. By stationarity of the process, J is uniformly distributed over $\{1, 2, \ldots, r\}$.
4. Given $Y_J = y_J > u$, generate $(Y_{J+1}, \ldots, Y_r)$ from the limiting conditional distribution given $\max(Y_J, \ldots, Y_r) = y_J$. This may be done by using the random walk approximation (4.12) to generate a realization of $(Y_{J+1}, \ldots, Y_r)$, rejecting and resampling this whole vector (not Y_J) if $\max(Y_{J+1}, \ldots, Y_r) > y_J$, and so on until the vector is not rejected. That this gives the correct limiting conditional distribution is a trivial application of the rejection method of simulation theory.
5. Similarly, use the backward random walk to generate $(Y_1, \ldots, Y_{J-1})$ given $Y_J = y_J$ and $\max(Y_1, \ldots, Y_J) = y_J$.
6. Calculate $W^*_{n,r}$, with $u_n = u$, and also the total number of exceedances of the level u among $Y_1, \ldots Y_r$ (i.e. functional (iv) in Section 3); call this latter quantity T_r.
7. Repeat steps 2–6 L times to generate L independent event simulations. The parameter θ_r is estimated as the reciprocal of the sample mean of the L values of T_r.
8. Increase r until stability of the estimates is achieved.

5 Higher order models and model selection

In this section we discuss the constraints on, and selection of, parametric models for the dependence structure of extreme values from Markov chains of order $d \geq 2$.

First consider the extension of the first–order Markov model, described in Section 2.3, to a general d^{th}–order Markov chain, $\{Y_n\}$. The corresponding joint density of $(Y_1, \ldots, Y_N)$ can be written as

$$\prod_{j=d}^{N} p_d(y_{j-d+1}, \ldots, y_j) / \prod_{j=d}^{N-1} p_{d-1}(y_{j-d+2}, \ldots, y_j), \tag{5.1}$$

where p_j is the joint density of j consecutive observations. Using the same arguments as in Section 2, both the numerator and denominator of (5.1) may be approximated by multiplying together contributions similar in form to (2.10), derived from (2.5). Clearly, the exponent measures V_d and V_{d-1}, associated with the d^{th} and $(d-1)^{th}$ dimensional joint densities, see (2.2), are related by

$$V_{d-1}(x_1, \ldots, x_{d-1}) = V_d(x_1, \ldots, x_{d-1}, \infty) \tag{5.2}$$

$\forall (x_1, \ldots, x_{d-1}) \in \mathbf{R}_+^{d-1} \backslash \{\mathbf{0}\}$. Therefore, given a d-dimensional parametric model for V_d, the joint density (5.1) can be calculated. Some models, though, are not given explicitly in terms of the exponent measure, but instead are defined by the associated dependence measure H_d (see (2.2)). In order to evaluate (5.1) directly for such models, it is necessary to obtain an analogous relationship to (5.2), which relates the dependence measure of lower–dimensional marginal models to the dependence measure of the d-dimensional model. Denoting by h_d and h_{d-1} the d and $(d-1)$–dimensional densities of H_d and H_{d-1}, where H_{d-1} is the dependence measure of the $(d-1)$–dimensional marginal model of H_d, it can be shown from (5.2) and (2.2) that

$$h_{d-1}(w_1, \ldots, w_{d-1}) = \int_0^1 (1-q)^{d-1} h_d\{(1-q)w_1, \ldots, (1-q)w_{d-1}, q\} dq, \tag{5.3}$$

with $w_1 + \ldots + w_{d-1} = 1$. The proof of this result is similar to the proof of Theorem 2 in Coles and Tawn (1991). As an illustration of (5.3) consider the Dirichlet model, with parameters $(\alpha_1, \ldots, \alpha_d)$,

$$h_d(\mathbf{w}) = \left\{ \prod_{j=1}^{d} \frac{\alpha_j}{\Gamma(\alpha_j)} \right\} \frac{\Gamma(\boldsymbol{\alpha}.\mathbf{1}+1)}{(\boldsymbol{\alpha}.\mathbf{w})^{d+1}} \prod_{j=1}^{d} \left(\frac{\alpha_j w_j}{\boldsymbol{\alpha}.\mathbf{w}} \right)^{\alpha_j - 1}; \quad \mathbf{w} \in S_d, \alpha_j > 0,\ j = 1, \ldots, d. \tag{5.4}$$

then h_{d-1} has a Dirichlet model with parameters $(\alpha_1, \ldots, \alpha_{d-1})$.

As with standard time series modelling, the selection of the order of the process is difficult. In our temporally dependent extreme value problem testing between orders $d-1$ and d is equivalent to testing for conditional independence of variables within a unit simplex domain, subject to normalization constraints. Standard likelihood procedures can be used for such tests; however, typically models will be non–nested, in which case we suggest the use of a combination of the usual time series procedures, such as *AIC* and *BIC*, and informal model goodness–of–fit assessment based on comparisons of predicted and observed functionals of interest (see Section 3). The same procedures will be used to select a model for the dependence structure given the order of the process.

Thus it is clear that the estimation procedure discussed in Section 2.1 extends to higher order Markov chains. The extension of the limit theory of Section 4 and the calculations via the simulation step are not presented here, however these follow as a consequence of the formulation of Section 4 applied to results given by Yun (1993). Here though we concentrate on models for H_F and H_B in the $d = 2$ case.

By the definition of H_F and H_B, from property (2.2) it is easily shown that

$$H_F(x) = \int_{a(x)}^{1} w dH(w) \quad \text{and} \quad H_B(x) = \int_{0}^{a(-x)} (1-w) dH(w) \tag{5.5}$$

with $a(x) = \exp(-x)/[1+\exp(-x)]$. Thus the models for H, or equivalently V, estimated using the procedure of Section 2, provide models for H_F and H_B. In this paper we will restrict attention to four specific models, others being given by Coles and Tawn (1991), Smith (1993), and Coles and Walshaw (1993).

1. Bilogistic model:

   $$h(w) = (1-\alpha_1)(1-s)s^{1-\alpha_1}/\{(1-w)w^2[\alpha_1(1-s)+\alpha_2 s]\}, \quad 0 < \alpha_i \leq 1 \quad i = 1,2, \tag{5.6}$$

   where $s \equiv s(w)$ is the root of

   $$(1-\alpha_1)(1-w)(1-s)^{\alpha_2} = (1-\alpha_2)ws^{\alpha_1}, \tag{5.7}$$

   (Joe *et al.*, 1992). Thus, $H_F(x) = s_*^{1-\alpha_1}$ and $H_B(x) = (1-s_+)^{1-\alpha_2}$, where $s_* = s(a(x))$ and $s_+ = s(a(-x))$, with $a(.)$ and $s(.)$ defined by (5.5) and (5.7).

2. Negative Bilogistic model: As (5.6), but with $-\infty < \alpha_i < 0, i = 1,2$, (Coles and Tawn, 1994), so $H_F(x) = 1 - s_*^{1-\alpha_1}$ and $H_B(x) = 1-(1-s_+)^{1-\alpha_2}$.

3. Dirichlet model: Defined by the measure density (5.4), so $H_F(x)$ and $H_B(x)$ must be evaluated numerically from (5.5).

4. Asymmetric Logistic model: $V(x,y) = (1-\theta)/x + (1-\phi)/y + [(\theta/x)^{1/\alpha} + (\phi/y)^{1/\alpha}]^\alpha$, $0 \leq \theta \leq 1, 0 \leq \phi \leq 1, 0 < \alpha \leq 1$, (Tawn, 1988). Unlike models 1–3 neither $H_F(x)$ nor $H_B(x) \to 0$ as $x \downarrow -\infty$, since there is positive probability of an infinite step out of the upper tail of the random walk. Thus $H_F(-\infty) = 1-\theta$, $H_B(-\infty) = 1-\phi$ with $H_F(x) = 1-\theta+\theta^{1/\alpha}(\theta^{1/\alpha}+\phi^{1/\alpha}\exp\{-x/\alpha\})^{\alpha-1}$ and $H_B(x) = 1-\phi+\phi^{1/\alpha}(\phi^{1/\alpha}+\theta^{1/\alpha}\exp\{-x/\alpha\})^{\alpha-1}$ for $z > -\infty$.

6 Application to Temperature Data

Our data are daily minimum temperatures, recorded to the nearest degree Fahrenheit, at Wooster, Ohio. Wooster is one of 138 high–quality stations reported in the U.S. Historical

Climatology Network (Quinlan *et al.*, 1987) for which long–term daily data are available. Our reasons for the selection of this particular site are due to it being representative of the network with a long and good quality data series (June 1893 — December 1987 with only 45 missing values). Furthermore, some exploratory analysis of this series is already available (Grady, 1992). Restriction to daily minimum temperatures does not greatly inhibit the scientific usefulness of the study, since most practical issues relating to cold weather focus on this quantity, whilst trends in global average surface temperature have also been primarily attributed to increases in minimum temperatures (Karl *et al.*, 1992).

Findings in Coles *et al.* (1993) suggest that there is evidence for a quadratic trend in the Wooster series and that the series is approximately stationary over the winter months December to February, during which all the yearly extreme minimum temperatures occur. To simplify presentation we focus here on the winter months only and assume stationarity throughout this season and over years. A detailed analysis of the complete yearly series, accounting for seasonality, is deferred to a companion paper (Coles *et al.*, 1993).

We denote by $Y_1, Y_2, \ldots$ the negated series of daily minimum temperatures. Negating the series in this way means that attention is turned to maxima rather than minima. As with most contemporary extreme value analysis, a primary step is the evaluation of a threshold above which the asymptotic model is treated as exact. For this analysis we need a threshold which satisfies both marginal and joint requirements of the model. That is, all exceedances should follow a GPD with the dependence of successive exceedances being Markov with dependence determined by a multivariate extreme value model. If the asymptotic model is applicable above a given threshold each model component should be invariant to a higher threshold selection. Thus, to maximize efficiency, we need to find the lowest threshold above which model stability is achieved.

By use of a mean residual life plot (Davison and Smith, 1990) a range of candidate thresholds based on marginal stability is identified. Then, for a series of thresholds within this range, a first–order Markov model with logistic dependence model (2.3) is repeatedly fitted. The resulting parameter estimates appear stable with respect to thresholds above 5.5 (i.e. $-5.5°$F), yielding 232 exceedances — an average of less than one per month. We illustrate here the stability of the dependence parameter, α, as viewed by the variation in the corresponding extremal index with threshold level shown in Figure 1. After allowing for sampling variability it is evident that θ decreases for thresholds below 5.5 corresponding to stronger dependence, whereas higher thresholds yield similar estimates with poorer precision. Thus we take $u = 5.5$ and $\hat{\lambda} = 232/8485$.

Under the assumption of a first–order process and with the stated threshold choice, each of the dependence models discussed in Section 5 is considered within the context of the Markov model. The resulting model fits are given in Table 1.

Based on the marginal parameter estimates and the AIC or BIC criteria, there are only small differences between the models. Each model exhibits slight asymmetry corresponding to extreme

Model	σ	ξ	dependence parameters	NLLH	θ	m_u	m_s
bilogistic	5.85 (0.57)	-0.190 (0.068)	$\alpha = 0.785(0.035)$ $\beta = 0.651(0.061)$	1475.77	0.62	1.06	0.77
negative bilogistic	5.80 (0.58)	-0.175 (0.068)	$\alpha = -1.089(0.195)$ $\beta = -2.119(0.362)$	1476.46	0.61	1.07	0.81
Dirichlet	5.88 (0.55)	-0.186 (0.069)	$\alpha = 0.689(0.178)$ $\beta = 0.349(0.060)$	1476.14	0.62	1.07	0.80
asymmetric logistic	5.83 (0.58)	-0.186 (0.075)	$\alpha = 0.694(0.064)$ $\theta = 0.796(0.175)$ $\phi = 1.000(0.171)$	1476.40	0.63	1.05	0.73
logistic	5.85 (0.58)	-0.191 (0.070)	$\alpha = 0.730(0.026)$	1477.17	0.62	1.06	0.76

Table 1: Summary of Markov model fits based on various choices for bivariate extreme value limit. NLLH is the negative log–likelihood of the fitted model, θ is the corresponding extremal index estimate and m_u and m_s are the mean number of threshold upcrossings and mean intra–cluster standard deviation respectively.

events rising at a faster rate than they decline, although the asymmetry is not significant when judged by a formal likelihood ratio test for the logistic model within the bilogistic family. One way of comparing the impact of the different dependence models is to examine summary statistics of extreme events generated from the fitted models using the simulation procedure described in Section 4, with the random walk step lengths given in Section 5. To illustrate, Table 1 gives, for each dependence model, the simulated values of the extremal index (θ), the mean number of threshold upcrossings (m_u) and the mean standard deviation of excesses within a cluster of exceedances (m_s). The similarity of these values across models confirms that dependence model selection is a relatively unimportant aspect of the modelling procedure here; hence we restrict further analysis to the logistic model only.

We now examine the suitability of the assumption that extremes of the process are first–order Markov by looking for conditional independence between observations of lag 2 within extreme events. One approach is to consider the trivariate distribution of consecutive triples of extremes; under the first–order Markov model this distribution should lie in the domain of attraction of a trivariate extreme value distribution with special structure as follows. First transform the observations as

$$Y_i^* = \lambda^{-1}[1 + \xi(Y_i - u)/\sigma]^{1/\xi}; \quad i = 1, \ldots, N, \tag{6.1}$$

so that the Y_i^* have a unit Frechet distribution. In fact, because we need to look at triples where perhaps only one or two of the observations has exceeded the threshold, equation (6.1) is applied with the fitted tail model extrapolated back below the threshold u (see Coles and Tawn, 1994). Now, under the first–order Markov assumption with logistic model for transitions, the distribution of $(Y_i^*, Y_{i+1}^*, Y_{i+2}^*)$ is in the domain of attraction of the time–series logistic model

(Coles and Tawn, 1991) with parameters $r_1 = r_2 = 1/\alpha$. In particular, this means that if we set $R_i = Y_i^* + Y_{i+1}^* + Y_{i+2}^*$ and $W_{i,j} = Y_{i+j-1}^*/R_i$; $j = 1,2,3$ then the density of the angular components $(W_{i,1}, W_{i,3})$ is of known form. We can therefore compare contours of the fitted density model with the empirically obtained points for which at least one component exceeds the threshold; this is done in Figure 2, with contours drawn at heights $1, 2, \ldots 10$. The data appear consistent with the model contours though of course more formal goodness of fit tests based on this plot could be developed to further assess this.

A simpler and somewhat more *ad hoc* procedure is to look for serial correlation in the differences of the successive random walk steps during extreme events. The first lag auto–correlation is 0.12 which is not significantly different from zero. Our conclusion is that there is no reason to doubt the assumption of a first–order Markov process in the extremes. Further examination of this assumption is obtained below by comparing the model–based estimates of various quantities of interest with their empirical counterparts.

6.1 The Mean Cluster Size

This functional is just θ^{-1} (Leadbetter, 1983); equivalently, we therefore focus on estimation of θ. Using the random walk, θ can be calculated using the Weiner–Hopf method of Smith (1992) or the simulation–based method discussed here. Generally the latter will be adequate, although here we have used the former to produce the smooth curve shown in Figure 3, which illustrates for the logistic model the near–linear relationship between θ and the extremal coefficient, $V(1,1)$, over most of the range of θ.

For a given threshold, an empirical estimate of θ can be obtained simply as the reciprocal of the sample mean cluster size. However, this requires some method of identifying clusters; there are many possibilities for this (Smith and Weissman, 1994). Figure 1 shows θ estimated using a 'fixed storm length' definition, whereby extreme events have a maximum duration of specified length. In Figure 1 the maximum lengths are 2, 5 and 10 days plotted as $+, \bullet$ and $\times$ respectively. Each empirical estimator follows the profile of the fitted model, with a maximum duration of 5 days apparently doing best. Figure 1 clearly illustrates the difficulty of analysing extreme events without a formal temporal model: results are sensitive to the choice of declustering scheme, whilst without a model for reference or additional physical information, there is little basis for making an appropriate choice.

6.2 The distribution of the number of exceedances per cluster

¿From the random walk, the distribution, π, of the number of threshold exceedances per cluster can easily be obtained. As with the extremal index this functional is invariant to threshold choice. For the fitted logistic model π is given in Table 2. Unlike some analytical examples (Hsing *et al.*, 1988), π is not geometric, having both a heavier tail and larger $\pi(1)$. Also shown in Table 2 are the empirical estimates of π, based on the declustering scheme which gave the

	$\pi(j)$			
j	Model–based	$u = 5.5$	$u = 7.5$	$u = 10.5$
1	0.6585	0.4478	0.6095	0.6721
2	0.2002	0.4179	0.2857	0.2951
3	0.0787	0.1119	0.0857	0.0164
4	0.0330	0	0.0095	0
5	0.0154	0.0223	0.0095	0.0164
≥ 6	0.0144	0	0	0

Table 2: Empirical and model–based estimates of the distribution, π, of the number of the exceedances of the threshold, u, per cluster

best empirical estimate of θ, for various choices of threshold u. According to the asymptotic arguments on which our model is based, the distribution of the number of events within a cluster should be independent of the threshold, once thresholds are sufficiently high. Indeed, our model threshold of 5.5 was selected so as to give stability in the mean cluster size (Figure 1). It seems, however, that this level does not ensure stability of the distribution as a whole, and that somewhat higher levels may be required to achieve this. This explains, in part, the poor agreement between the model–based and empirical estimates of π. The agreement is better when the empirical estimates are obtained at higher thresholds, though in each case the model under–predicts the proportion of two–day events observed. This indicates that the first–order Markov model may not give a sufficiently detailed description of the true temporal dependence for such detailed aspects of temporal behaviour to be accurately estimated.

6.3 Return Levels

The distribution of the annual minimum temperature is an essential design input for many manufacturing processes. Quantiles of this distribution are termed 'return levels'. Using standard theory for extremes of stationary processes (Leadbetter *et al.*, 1983), we take

$$\Pr\{\max_{i=1,\ldots,m} Y_i < y\} \approx \{F(y)\}^{m\theta}, \tag{6.2}$$

where $m(= 90)$ is the number of days in the winter period. This result also follows as a consequence of Theorem 1 applied to functional (i) of Section 3. Using the GPD tail approximation for F, the return level y_p, defined such that (6.2) equals $1 - p$, satisfies

$$y_p = u - \sigma\xi^{-1}[1 - \{\lambda^{-1}[1 - (1 - p)^{1/(m\theta)}]\}^{-\xi}] \tag{6.3}$$

for $y_p > u$. Substitution of the maximum likelihood estimates (mles) of (λ, σ, k), together with the random walk estimate of θ based on the logistic model with the mle of the parameter α, leads to the mle $\hat{y}_p$ of y_p. Ideally we would like to use profile likelihood based confidence intervals for

$\hat{y}_p$. However, here this requires considerable additional computation relative to the delta method and so seems unwarranted for an exploratory analysis. Calculation even of the standard error of y_p is complicated, since we only have the hessian matrix corresponding to the vector $(\lambda, \sigma, k, \alpha)$. But by exploiting the linearity of Figure 3, we have $\theta \approx -1.34 + (1.81) \times 2^{\alpha}$, from which standard errors for y_p follow. The resulting estimates, plotted as a function of $-\log[-\log(1-p)]$, together with pointwise 95% confidence intervals, are given in Figure 4.

Also shown are return level estimates based on the cluster maxima approach of Davison and Smith (1990), for which the GPD parameter estimates are $\hat{\sigma} = 7.16(0.76)$, $\hat{\xi} = -0.344(0.066)$ and $\hat{\lambda} = 134/8458$. Clearly, a consequence of the declustering is that some distortion of the tail has occurred, with compensatory changes in the estimates of scale and shape. The cluster maxima model therefore gives comparable estimates within the range of the data, but under–estimation in the extreme tail by comparison with the Markov model. The apparent improvement in precision of the cluster maxima approach, as measured by confidence intervals obtained using the delta method, is an artefact of the shape parameter estimate being greater for that model. Results in Davison and Smith (1990) suggest that if profile likelihood based confidence intervals had been used the width of confidence intervals would be more comparable.

6.4 Aggregate Excesses and Cold–Waves

The functionals examined in Sections 6.1 — 6.3 relate only to the peak and duration of extreme events. There are many problems for which the cumulative effect of extreme values is a better measure of the severity of an extreme event. In such cases the aggregate excess functional, W_E (functional (v) in Section 3), is the main quantity of interest.

Anderson and Dancy (1992) derive the asymptotic distribution of this functional when $z = 0$, from the point process characterization of Hsing (1987). No specific forms for the dependence structure are assumed, and consequently the limiting distributional family is very broad. By contrast, our fitted Extremal Markov model completely specifies the aggregate excess distribution, which can be obtained by simulation, based on importance sampling of the cluster maximum to ensure accurate tail estimates. For the temperature data, Figure 5 compares the model–based distribution function of W_E, with the corresponding empirical distribution function; in each case we take $z = 0$, and the empirical estimation is based on the previously discussed declustering scheme. Furthermore, by Theorem 1, the distribution of the winter period maximum aggregate excess, W_E^*, is given by

$$\Pr(W_E^* \leq w) = \exp\{-90\theta\lambda(1 - W_p)\},$$

where $W_p = \Pr\{W_E \leq w\}$.

The joint distribution of the aggregate excess and the cluster maximum, W_M, can also be obtained from the above simulation scheme. This distribution has a singular component, with probability $\pi(1)$, for $W_E = W_M$, and a density on $\{w : W_E > W_M\}$. In Figure 6 model–based densities are compared with the corresponding empirical observations of the cluster maximum and aggregate excess.

As a final example, we consider the cold–wave functional discussed in Section 3. The stochastic behaviour of this characteristic is highly dependent on the localized structure of extreme events. We fix $m = 3$ and consider various levels of cold–wave definition $z > 0$. By simulating clusters of exceedances from the fitted model we obtain an estimate of $C(z)$, the probability a cluster contains no cold waves at level $u + z$. Then by Theorem 1, the probability of no cold waves of level $u + z$ in a winter period is given by $\exp\{-90\theta\lambda C(z)\}$. This probability is plotted as a function of $u + z$ in Figure 7.

6.5 Comparison of model and empirical estimates

One of the earliest papers to address the kind of questions we have studied in this paper was Barnett and Lewis (1965). In that paper, motivated by an industrial problem, they wanted to compute probabilities of various events related to low temperatures, such as whether at any time during the winter the temperature dropped below a certain critical level for three consecutive hours. Their approach was essentially to condition on the annual minimum value, which they modelled by a Gumbel distribution, and to use empirical approaches to obtain the required conditional probabilities. A disadvantage of their method, as they themselves quite openly pointed out, was that without any model for the conditional probabilities their method required some rather *ad hoc* assumptions that they could not justify on general grounds.

¿From the more modern point of view of threshold methods, we could consider the conditional distribution of an event of the form described given the mimimum temperature within a cluster, rather than given the mimimum temperature over the whole year, and methods of the form described by Barnett and Lewis should still be applicable. This could be considered an empirical approach to the problem, as opposed to the model–based approach which has been the main theme of this paper. The caveat of empirical estimators in this context is that clusters are intrinsically rare, so that estimates tend to be unstable and partially dependent on the declustering scheme adopted. We have shown how an appeal to asymptotic model structure can avoid these difficulties, but it is still valuable to compare model–based and empirical estimates. For example, in the cold–wave example of Section 6.4, the empirical estimate of $C(u)$ is $123/134 = 0.92$, by comparison with a value of 0.87 obtained from the fitted model.

Generally, where comparison is possible, the fitted model appears to capture the key empirical features of functionals of interest. When discrepancies do arise, as highlighted by the empirical and model estimates of the cluster size distribution in Table 2, the dilemma is whether to attribute the discrepancy to inadequacies of the asymptotic model, or to instability of empirical estimates. Potential deficiencies of the asymptotic model include the possibility of non–stationarity of the process over the winter period (cf. Coles *et al.*, 1993), the possible need for a higher–order Markov model, and inconsistencies due to our model being continuous and the data discrete. The choice of threshold $u = 5.5$ was made so as to ensure stability of empirical estimates of the mean cluster size, but the results of Section 6.2 suggest that a higher threshold may be required to obtain stability of other cluster characteristics. Thus, for example, there is significant evidence that the model under–estimates the aggregate excess distribution at all levels except the tails of the distribution. This is perhaps not too serious, since if the relevant functional is the aggregate excess then the upper tail is the region of most interest.

Appendix

Proof of Theorem 1: Let $\{t_n\}$ denote another sequence of integers such that $t_n \to \infty, t_n/p_n \to 0$ and $n\phi(t_n)/p_n \to 0$. To see that such a sequence must exist, observe that if we can find a sequence $\{p_n\}$ satisfying (4.2), then there also exists a sequence satisfying (4.2) and $n\phi(p_n/\delta)/p_n \to 0$ for any $\delta > 0$. By choosing a sequence $\{\delta_n\}$ so that $\delta_n \to \infty$ sufficiently slowly, we can ensure $n\phi(p_n/\delta_n)/p_n \to 0$. Let $t_n = 1 + \text{int}(p_n/\delta_n)$, where $\text{int}(x)$ denotes integer part of x.

With t_n so defined, let $k_n = \text{int}(n/(p_n + t_n))$ and write the index set $\{1, \ldots, n\}$ as $A_1 \cup B_1 \cup A_2 \cup B_2 \cup \ldots \cup B_{k_n}$ where the A's and B's are successive 'long' and 'short' blocks of lengths p_n and either t_n or $t_n + 1$; thus $A_1 = \{1, 2, \ldots, p_n\}$, B_1 is either $\{p_n + 1, \ldots, p_n + t_n\}$ or $\{p_n + 1, \ldots, p_n + t_n + 1\}$, and so on. Write

$$W_n(u_n) = \sum_{i=1}^{k_n} (W_{n,i}^{(1)} + W_{n,i}^{(2)})$$

where

$$W_{n,i}^{(1)} = \sum_{j \in A_i} g\{(Y_j - u_n)_+, \ldots, (Y_{j+m-1} - u_n)_+\},$$

$$W_{n,i}^{(2)} = \sum_{j \in B_i} g\{(Y_j - u_n)_+, \ldots, (Y_{j+m-1} - u_n)_+\},$$

and observe that by (4.3), (4.7), and the fact that $k_n t_n\{1 - F(u_n)\} \to 0$, we have

$$\sum_{i=1}^{k_n} W_{n,i}^{(2)} \to_p 0.$$

Hence it suffices to calculate the asymptotic distribution of $\sum W_{n,i}^{(1)}$. Let $\{W_{n,i}^{(0)}, 1 \le i \le k_n\}$ denote *independent* random variables with the same (common) marginal distribution as $W_{n,i}^{(1)}, 1 \le i \le k_n$. Define

$$S_{n,k}^{(j)} = \sum_{i=1}^{k} W_{n,i}^{(j)}, \quad T_{n,k}^{(j)} = \sum_{i=k+1}^{k_n} W_{n,i}^{(j)}, \quad j = 0, 1 \text{ and } 0 \le k \le k_n,$$

$(S_{n,0}^{(j)} = 0, T_{n,k_n}^{(j)} = 0, \ \ j = 0, 1)$ and

$$\psi_k(t) = \text{E}\left[\exp\{it(S_{n,k}^{(0)} + T_{n,k}^{(1)})\}\right], 0 \le k \le k_n.$$

Thus ψ_0 is the characteristic function of $\sum W_{n,k}^{(1)}$, ψ_{k_n} is the characteristic function of $\sum W_{n,k}^{(0)}$, and

$$| \psi_{k+1}(t) - \psi_k(t) | \leq \phi(t_n), \quad 0 \leq k \leq k_n - 1.$$

Hence

$$| \psi_0(t) - \psi_{k_n}(t) | \leq k_n \phi(t_n) \to 0.$$

Thus the asymptotic distribution of $\sum W_{n,k}^{(1)}$ is the same as that of $\sum W_{n,k}^{(0)}$. To conclude the proof, define $I_{n,k}$ to be 1 if $Y_j > u_n$ for some j such that either $j \in A_k$ or $j - m + 1 \in A_k$, and 0 otherwise. By standard results in extreme value theory (e.g. Leadbetter, 1983; or Hsing *et al.*, 1988), $\sum I_{n,k}$ converges to a Poisson random variable with mean $\theta\tau$. The result then follows from this fact, the independence of summands, and (4.8).

Proof of (4.9): More precisely, we want to show that if (4.4) holds and if the limiting distribution defined by (4.9) exists, then this limiting distribution also satisfies (4.8).

Fix $w > 0$. If (4.9) holds, then there exists some sequence $p_n' \to \infty$ such that

$$\Pr\{W^* \leq w\} = \lim \Pr\{W_{n,p_n'}^* \leq w \mid \max(Y_1, \ldots, Y_{p_n'}) > u_n\}. \tag{6.4}$$

One way to characterise the limit in (4.11) is to condition on the first time, I, and the last time, J, at which the level u_n is exceeded, where $1 \leq I \leq J \leq p_n'$. The probability that $W_{n,p_n'}^* \leq w$ given I and J is a function of $J - I$, say q_{J-I}. Define $\theta_p(u) = \Pr\{Y_2 \leq u, \ldots, Y_p \leq u \mid Y_1 > u\}$ as earlier, and also $\overline{\theta}_p(u) = \Pr\{Y_1 \leq u, \ldots, Y_{p-1} \leq u \mid Y_p > u\}$. Then

$$\Pr\{W_{n,p_n'} \leq w \mid \max(Y_1, \ldots, Y_{p_n'}) > u\} =$$

$$\frac{\sum_{J=1}^{p_n'} \sum_{I=1}^{J} \overline{\theta}_I(u_n) \theta_{p_n'-J+1}(u_n) q_{J-I} \Pr\{Y_J > u_n \mid Y_I > u_n\}}{\sum_{J=1}^{p_n'} \theta_{p_n'-J+1}(u_n)}. \tag{6.5}$$

We claim that (4.12) is asymptotically indistingishable from

$$\theta \sum_{r=0}^{\infty} q_r \Pr\{Y_{r+1} > u_n \mid Y_1 > u_n\}. \tag{6.6}$$

Since our proof will work equally well with p_n' replaced by p_n, this will suffice to show that (4.9) and (4.11) lead to the same limit. Using the fact that $\theta_j(u_n)$ for $1 \leq j \leq p_n'$ is bounded below and above, and that $\sum_{J=1}^{p_n'} \theta_{p_n'-J+1}(u_n)$ diverges as $n \to \infty$, it suffices to show that

$$\sum_{I+1}^{J} \{\overline{\theta}_I(u_n) - \theta\} q_{J-I} \Pr\{Y_J > u_n \mid Y_I > u_n\}$$

becomes negligible for sufficiently large $J \leq p_n$. However, this follows directly from (4.4) and the fact that the definition (4.5) works equally well in reverse time.

Acknowledgements

We are grateful to Clive Anderson for pointing out that equation (2.5) is valid for non–equal λ_i and Thomas Karl of the National Climate Data Center for the temperature data. R.L.S. was partially supported by the NSF. J.A.T. and S.G.C. were each partly funded by a Nuffield Foundation grant and the analysis used equipment provided by the SERC under the Complex Stochastic Systems Initiative.

References

Anderson, C. W. (1990). Discussion of paper by A. C. Davison and R. L. Smith. *J. Roy. Statist. Soc., B*, **52**, 425–426.

Anderson, C. W. and Dancy, G. P. (1992). The Severity of Extreme Events. Submitted.

Asmussen, S. (1987). *Applied Probability and Queues.* Wiley, Chichester.

Barnett, V. and Lewis, T. (1967). A study of low–temperature probabilities in the context of an industrial problem (with discussion). *J. Roy. Statist. Soc., A*, **130**, 177–206.

Coles, S. G., Tawn, J. A. and Smith, R. L. (1993). A seasonal Markov model for extremely low temperatures. Submitted.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *J. R. Statist. Soc., B,* **53**, 377–392.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural design. *Appl. Statist.* To appear.

Coles, S. G. and Walshaw, D. (1993). Directional modelling of extreme wind speeds. *Appl. Statist.* To appear.

Grady, A. M. (1992). Modeling daily minimum temperatures: an application of the threshold method. Proceedings of the *5th Int. Meeting on Statist. Clim.*, ed. F. Zwiers, Canadian Climate Centre.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc., B*, **52**, 393–442.

Hsing, T. (1987). On the characterisation of certain point processes. *Stoch. Proc. Appl.*, **26**, 297–316.

Hsing, T. (1991). Estimating parameters of rare events. *Stoch. Proc. Appl.*, **37**, 117–139.

Hsing, T., Husler, J. and Leadbetter, M. R. (1988). On the exceedance point process for a stationary sequence. *Probab. Theory and Related Fields*, **78**, 97–112.

Joe, H., Smith, R. L. and Weissman, I. (1992). Bivariate threshold methods for extremes. *J. Roy. Statist. Soc., B*, **54**, 171–183.

Karl, T. R., Kukla, G., Razuvayez, V. N., Changery, M. J., Quayle, R. G., Heim, R. R., Easterling, D. R. and Fu, C. B. (1991). Global warming: evidence for asymmetric diurnal temperature change. *Geophys. Res. Lett.*, **18**, 2253—2256

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. *Z. Wahrscheinlichkeitsth.*, **65**, 291–306.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series*, Springer Verlag, New York.

Leadbetter, M. R., Weissman, I., de Haan, L. and Rootzén (1989). On clustering of high levels in statistically stationary series. Proceedings of the *4th Int. Meeting on Statist. Clim.*, ed. J. Sansom, New Zealand Meteorological Service.

O'Brien, G. L. (1974). The maximum term of uniformly mixing stationary processes. *Z. Wahrscheinlichkeitsth.*, **30**, 57–63.

O'Brien, G. L. (1987). Extreme values for stationary and Markov sequences. *Ann. Probab.*, **15**, 281–291.

Perfekt, R. (1991). Extremal behaviour of stationary Markov chains with applications. Submitted.

Pickands, J. (1971). The two–dimensional Poisson process and extremal processes. *J. Appl. Probab.*, **8**, 745–756.

Pickands, J. (1981). Multivariate extreme value distributions. *Proc. 43rd Sess. Int. Statist. Inst.*, 859–878.

Quinlan, F. T., Karl, T. R. and Williams, C. N., Jr. (1987). *United States Historical Climatology Network HCN) serial temperature and precipitation data.* NDP–019, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Resnick, S.I. (1987). *Extreme Values, Point Processes and Regular Variation.* Springer Verlag, New York.

Shi, D. Smith, R. L. and Coles, S. G. (1992). Joint versus marginal estimation for bivariate extremes. Submitted.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground level ozone. *Statist. Sci.*, **4**, 367–393.

Smith, R. L. (1992) The extremal index for a Markov chain. *J. Appl. Prob.*, **29**, 37—45.

Smith, R. L. (1993). Max–stable processes and spatial extremes. Submitted.

Smith R. L. and Weissman, I. (1994). Estimating the extremal index. To appear in *J Roy. Statist. Soc., B.*

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, **75**, 397–415.

Tawn, J. A. (1992). Estimating probabilities of extreme sea levels. *Appl. Statist.*, **41**, 77–93.

Tong, H. (1990). *Non–linear Time Series: a Dynamical Systems Approach.* Oxford: Oxford University Press.

Yun, S. (1993). The extremal index of a higher–order stationary Markov chain. Manuscript in preparation.

Figure Headings

Figure 1 Estimated extremal index, θ, plotted against the threshold, u, used in its estimation. The curve corresponds to estimates obtained using the Markov model with transitions described by the fitted logistic model. Associated 95% confidence intervals are shown by the broken lines. The points $+, \bullet, \times$ correspond to empirical estimates of θ obtained using a fixed storm length declustering scheme of 2, 5 and 10 days respectively.

Figure 2 Simplex plot of $W_{i,1}$ v $W_{i,3}$. Contours at heights $1, 10(1)$, are based on the trivariate time series logistic model and the points to observed values.

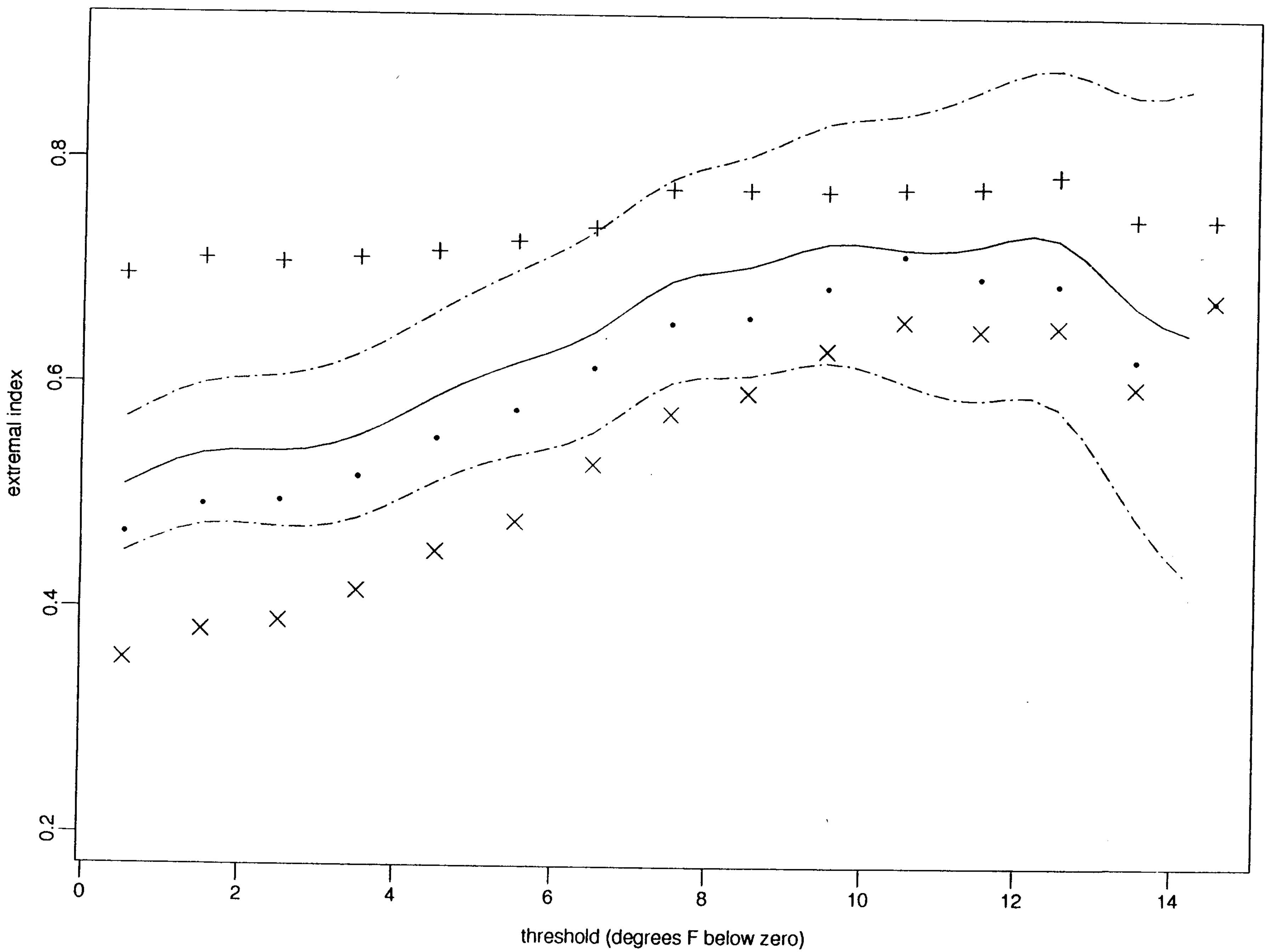
Figure 3 The relationship between the extremal index, θ, and the extremal coefficient, 2^α, for the Markov logistic model, with parameter α.

Figure 4 Return levels, y_p, together with associated 95% confidence intervals. *All exceedances* and *cluster maxima* correspond to the Markov model and the Davison and Smith (1990) approaches respectively.

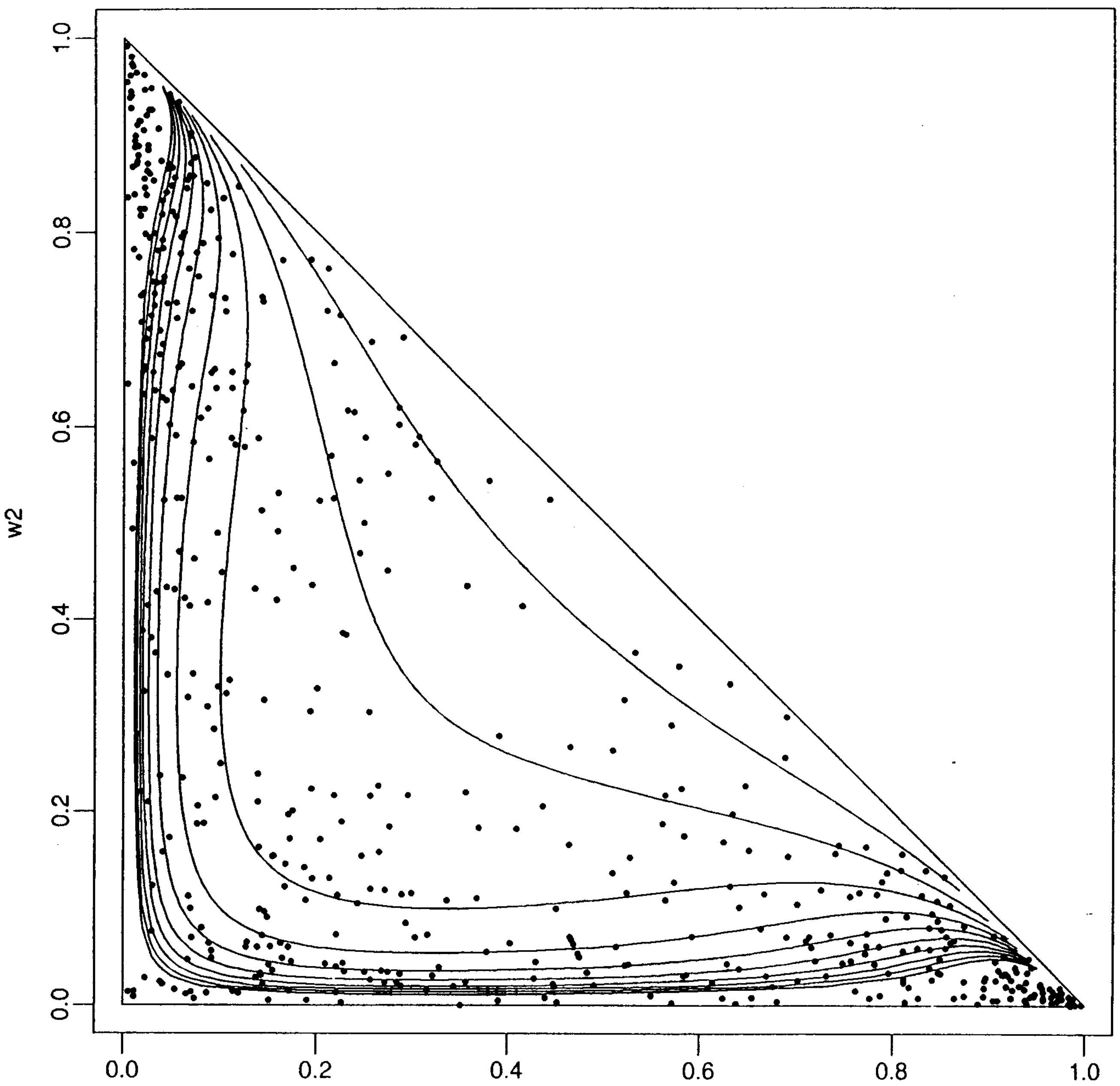
Figure 5 The distribution of aggregate excesses, $\Pr\{W_E \leq w\}$, estimated from the Markov model (solid line) and empirically (broken line).

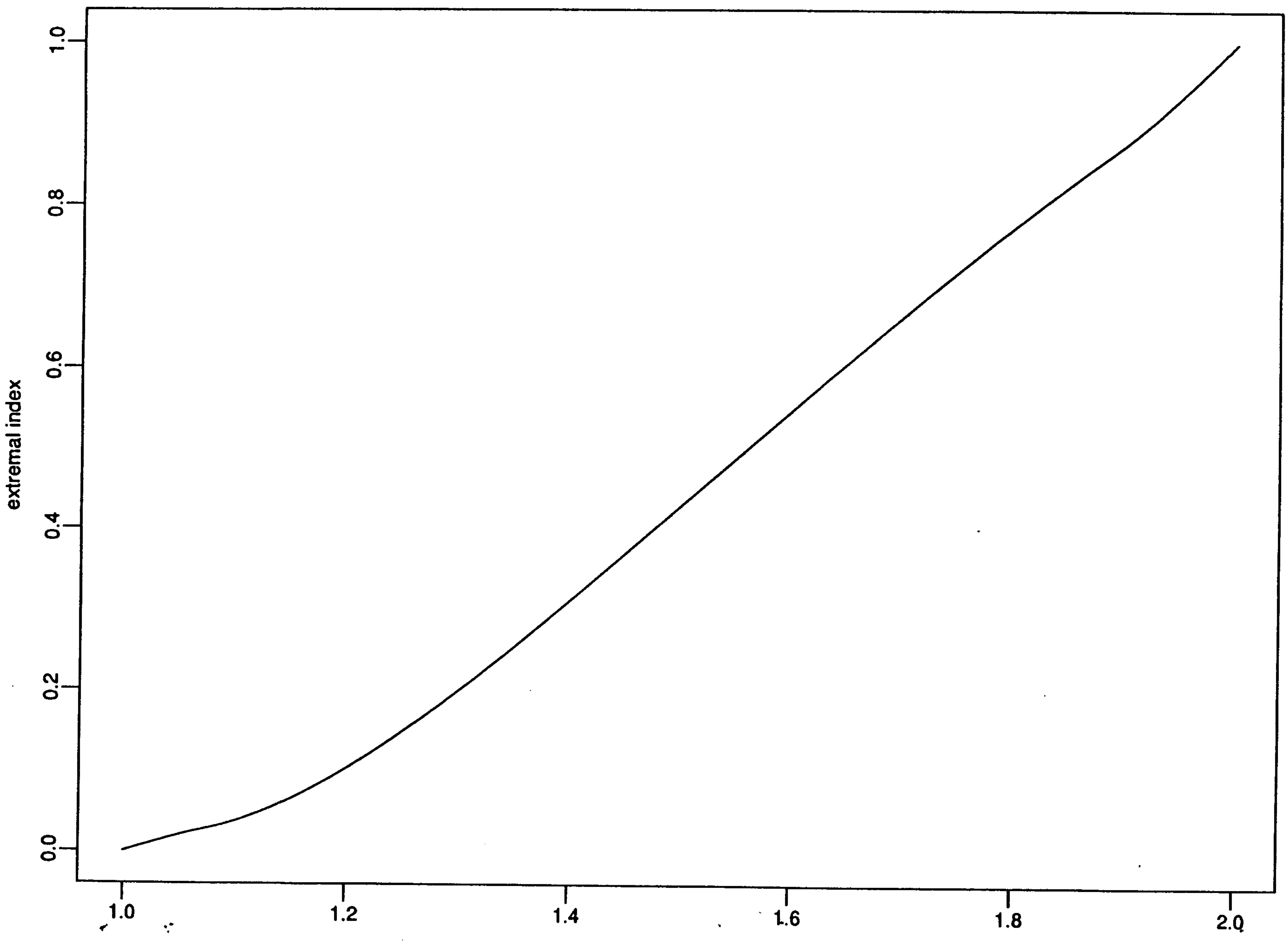
Figure 6 The joint distribution of the cluster maximum and aggregate excess. Plot (a) shows the singular component (i.e. cluster maximum = aggregate excess) occurring with probability $\pi(1)$, whereas plot (b) shows the continuous component. The fitted model and observations are given in each case — in (a) the observations are shown via a crude histogram.

Figure 7 Probability of no cold waves in a winter period for various levels, $u + z$, of cold wave definition.

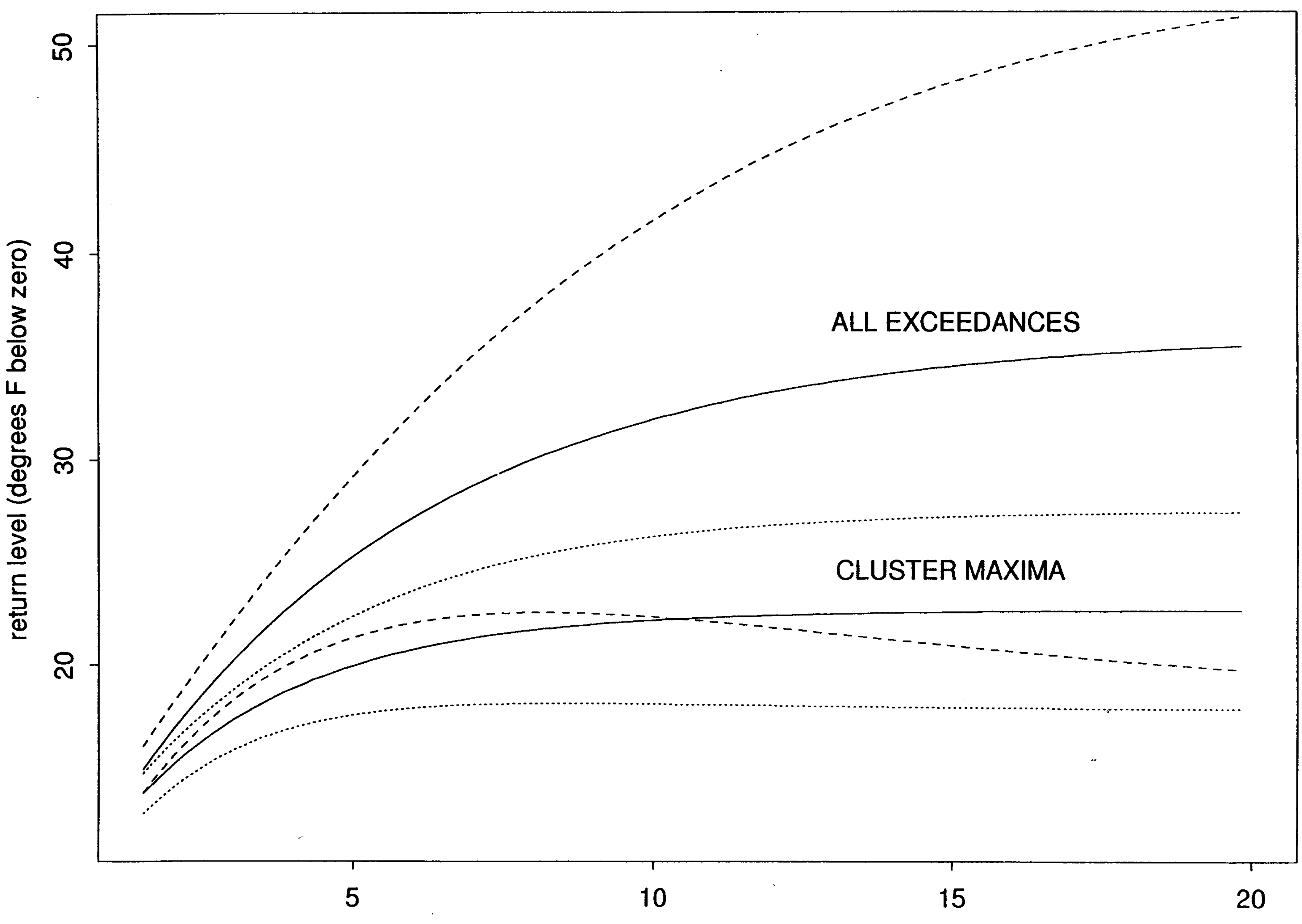


extremal index
0.2
0.4
0.6
0.8
0
2
4
6
8
10
12
14
threshold (degrees F below zero)

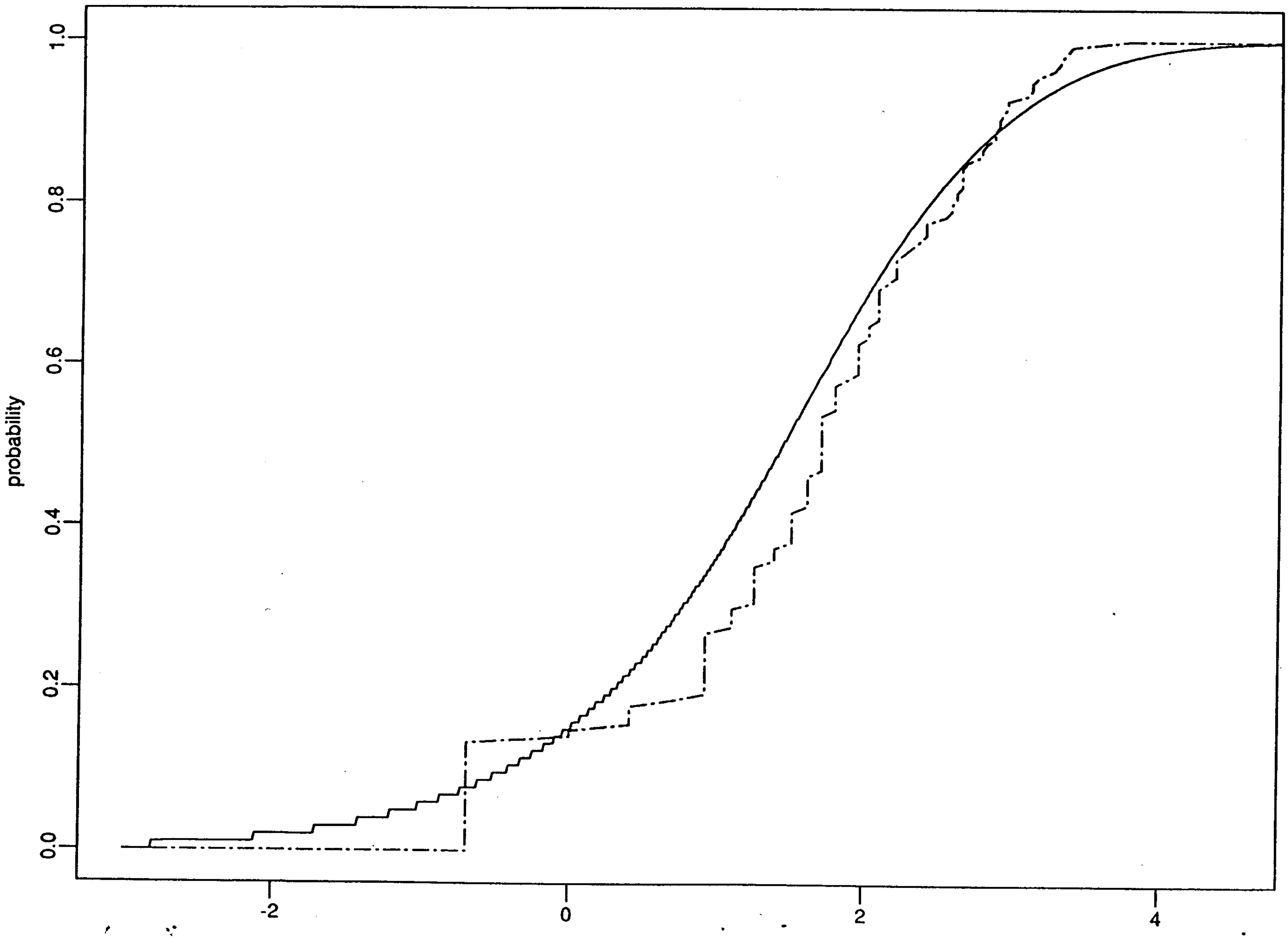

w2
0.0
0.2
0.4
0.6
0.8
1.0



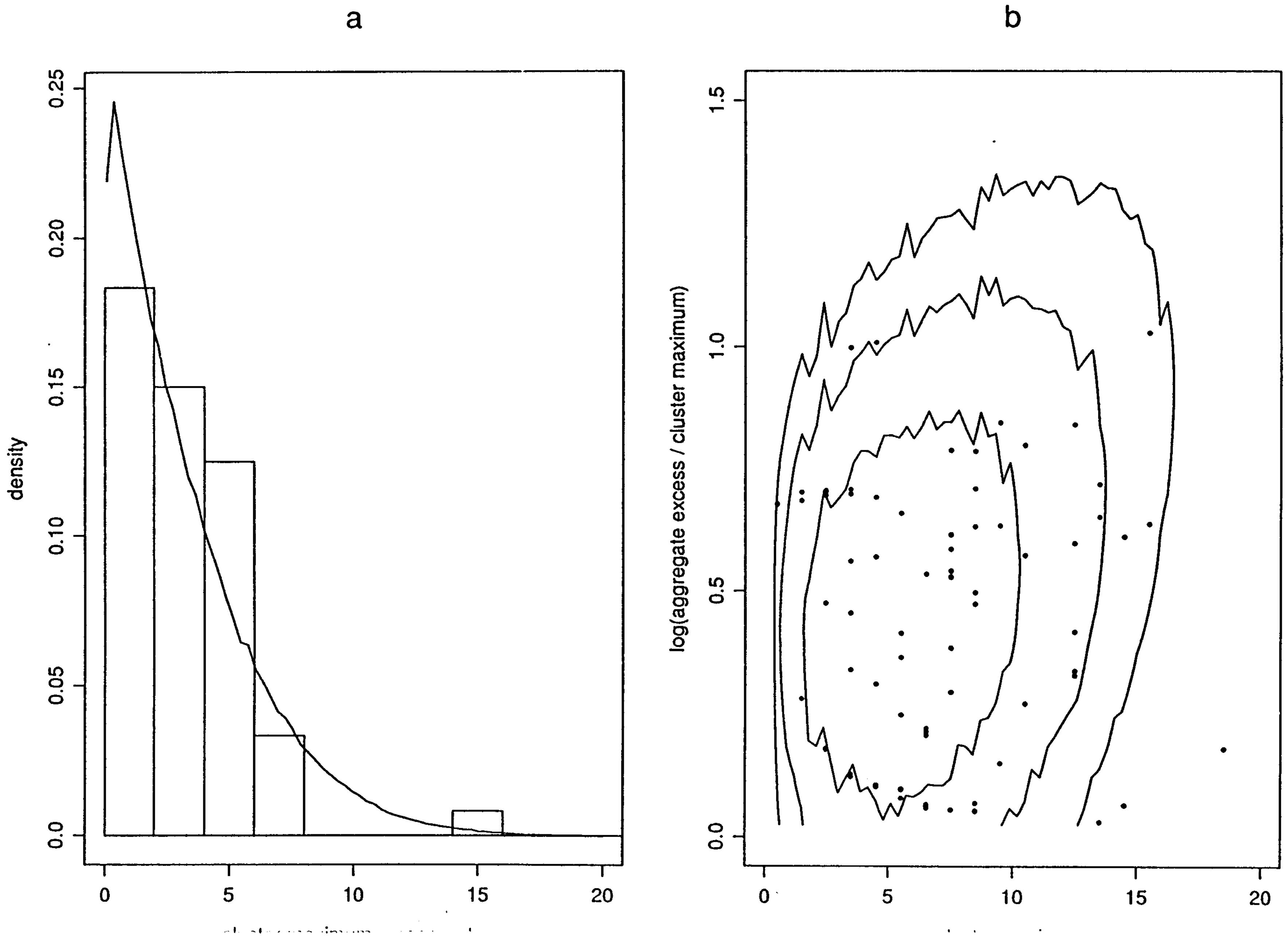
extremal index
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0

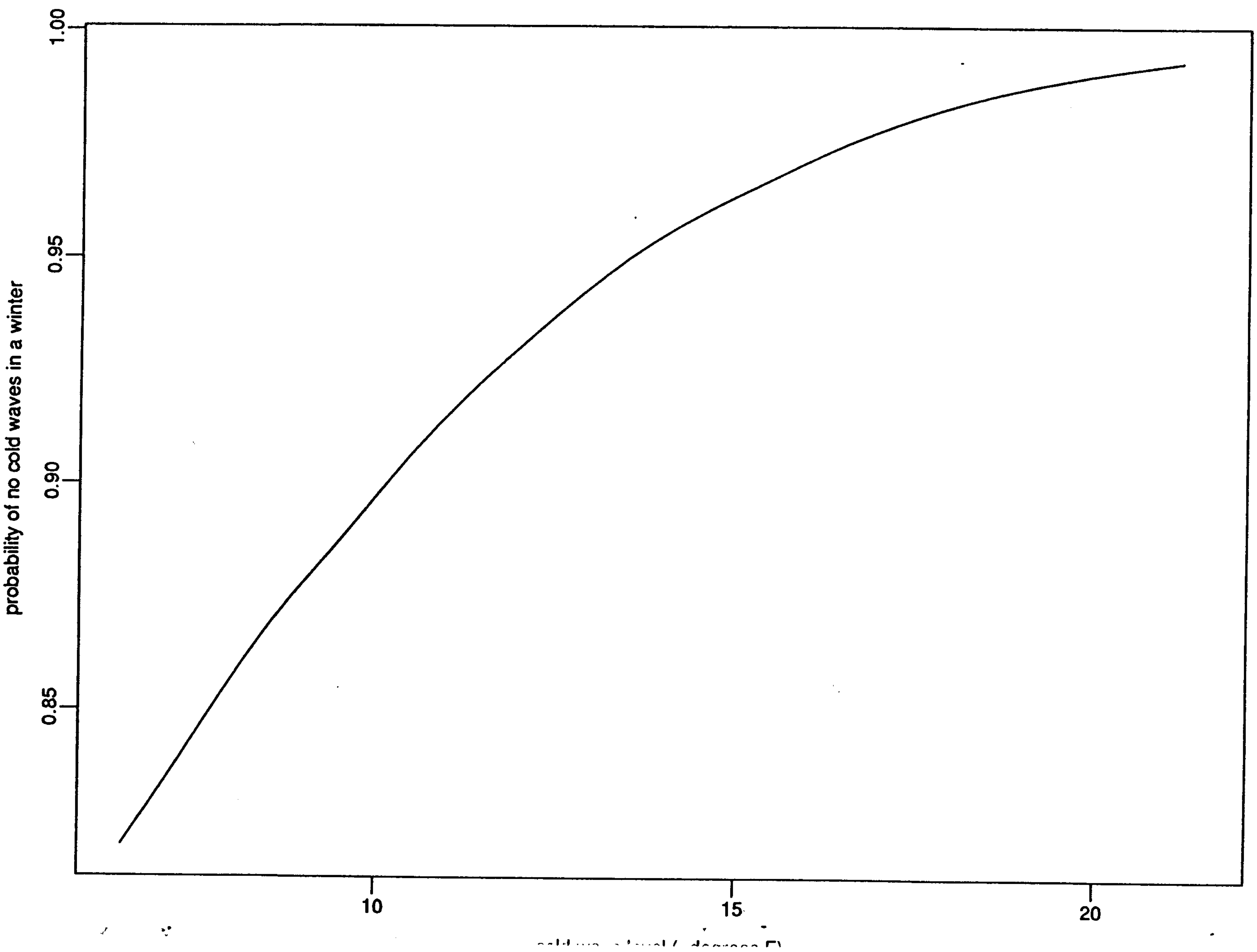


return level (degrees F below zero)
20
30
40
50
ALL EXCEEDANCES
CLUSTER MAXIMA
5
10
15
20


probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
0
2
4

Joint distribution of cluster maximum and aggregrate excess





probability of no cold waves in a winter
0.85
0.90
0.95
1.00
10
15
20