ENVIRONMETRICS
Environmetrics (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.794

Markov chain models for extreme wind speeds

Lee Fawcett and David Walshaw*,†

School of Mathematics and Statistics, Newcastle University, Newcastle NE1 7RU, U.K.

SUMMARY

Understanding and quantifying the behaviour of extreme wind speeds has important applications for design in civil engineering. As in the extremal analysis of any environmental process, estimates are often required of the probability of events that are rarer than those already recorded. Consequently, research has focused on the development of techniques that make optimal use of the available data. One such approach lies in *threshold methods*, which, unlike the more traditional annual maxima approach to the modelling of extremes, takes into consideration *all* extreme events, extreme in the sense that they exceed some high threshold. However, the implications of using all extremes in an analysis include problems of temporal dependence and non-stationarity. Several pragmatic ways of circumventing the problem of temporal dependence have been developed, though these often include the deletion of many extreme observations, for example, filter out a set of independent extremes. This paper looks at another approach to inference—one which explicitly *models* the temporal dependence of the process and so can use information on *all* extremes—and investigates the appropriateness of assumptions of short-term temporal dependence for wind speeds. We also examine the success of such methods at estimating some extreme events commonly studied for wind-speed data. Throughout this paper extreme wind speeds are analysed within a Bayesian framework, which can be argued to be particularly advantageous for extreme value analyses. For example, the objective of an extreme value analysis is usually an estimate of the probability of future events reaching extreme levels—something which is handled quite naturally in a Bayesian analysis through *predictive distributions*. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: Markov chain models; wind speeds; extreme value theory; temporal dependence; Bayesian inference; clusters

1. INTRODUCTION

Over the last decade or so, there has been a move away from the traditional 'Gumbel' framework for the statistical modelling of extreme values towards *threshold approaches*, which bring into use more extreme observations and so increase the precision of any analysis. However, the trade-off for this increased precision is the added complexity of near-certain temporal dependence between nearby observations. The most popular approach to circumvent this issue is to employ some filtering scheme which identifies threshold exceedances which are far enough apart to be deemed independent; such

*Correspondence to: David Walshaw, School of Mathematics and Statistics, Newcastle University, Newcastle NE1 7RU, U.K.
†E-mail: david.walshaw@ncl.ac.uk

Contract/grant sponsor: EPSRC.

Received 10 August 2005
Accepted 6 February 2006

‘declustering’ techniques are discussed, and implemented, in Davison and Smith (1990), Walshaw (1994), Coles (2001) and Fawcett (2005), and many others, which are cited within these few.

Such a pragmatic approach to temporal dependence, which is fairly straightforward to apply, ensures that the likelihood asymptotics are rigorous. However, declustering is wasteful of data, and estimation of model parameters has been shown to be sensitive to the filtering technique used to identify the independent threshold exceedances. Also, any information about dependence is lost through deletion.

In recent work, Fawcett and Walshaw (2006) argue that, when modelling extremes, it can be advantageous to make use of *all* extremes instead of employing some filtering scheme. Their work has shown that, instead of improving the estimation procedure by removing temporal dependence, filtering out a set of independent threshold exceedances can lead to substantial bias in parameter estimates, and especially in return level estimates. They show that if all threshold exceedances are used in the analysis, then model parameters are more accurately estimated and with greater precision (owing to the inclusion of more data). To account for the dependence in the data, appropriate adjustments to standard errors and confidence intervals can be achieved by employing methods due to Smith (1991). In this paper, we consider a more complex but potentially more informative approach, by explicitly *modelling* temporal dependence through Markov chain models.

The implementation and practical consequences of such models for wind speed data are explored here. We consider the ability of Markov chain models to estimate two *cluster functionals* of practical interest: (i) storm length and (ii) time between successive storms, and compare the estimation of these events for models which assume each of a first and second-order Markov structure.

In Section 2, we introduce the wind speed data used throughout this paper, and discuss some modelling issues relevant to these data. In Section 3, we build a Markov chain model, which allows for serial correlation in the wind speed data by invoking bivariate extreme value considerations for the temporal evolution of the process. Section 4 implements this first-order Markov chain model for the wind speed data, while Section 5 considers higher-order Markov chain models and explores the sensitivity of extreme event estimation to the choice of model order.

2. MODELLING EXTREME WIND SPEEDS

2.1. The wind speed data

The data used in this paper consist of hourly gust maximum wind speeds recorded at High Bradfield, a high altitude site on the Pennines, for the UK Meteorological Office. The data given to us were recorded over a 10-year period, from 1st January 1975 to 31st December 1984 inclusive, constituting approximately 86 000 observations after missing observations have been discounted. Figure 1 shows a time series plot of the hourly maxima at High Bradfield, a histogram of the data and a plot of the time series against the lag 1 time series. The first 3 years of data only are used in each case to best illustrate the relevant data characteristics. Table 1 gives summaries of the entire dataset. For a detailed description of the behaviour of the wind climate in the UK, see, for example, Shellard (1976); for a more involved discussion of the wind climate at high-altitude sites in the Pennines, see Smithson (1987).

2.2. A model for threshold exceedances

A natural way of modelling extremes of time series such as the hourly maximum wind speeds at High Bradfield is to use the generalised Pareto distribution (GPD) as a model for excesses over a high

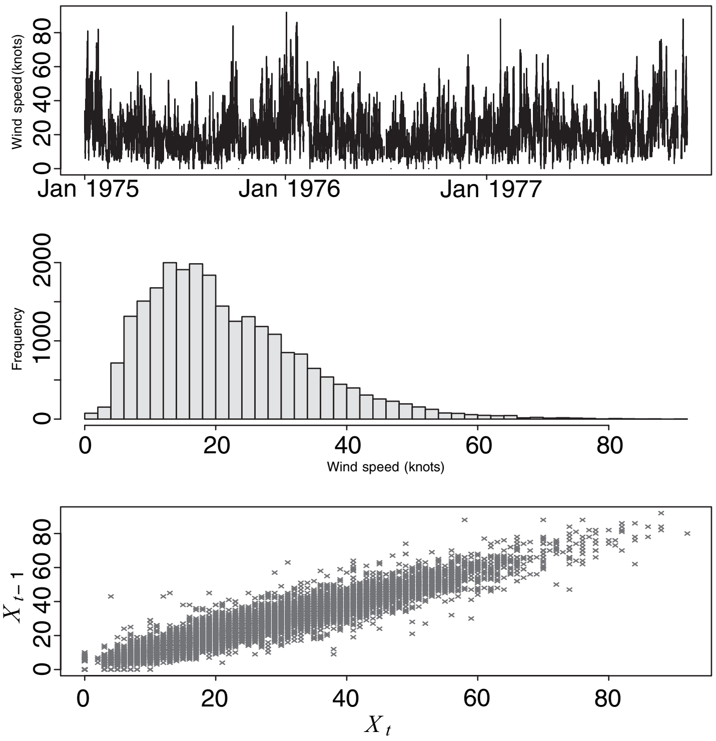

Figure 1. Time series plot, histogram and plot of the time series against the version at lag 1 for the hourly gust maxima at High Bradfield (1975–1977 inclusive).

Table 1. Summary statistics of High Bradfield hourly gust maxima (knots).

Mean	Standard Deviation	Median	LQ	UQ	Min.	Max.
22.28	12.24	20.0	13.0	29.0	0.0	99.0

threshold. Let $X_1, X_2, \ldots$ be a sequence of independent random variables with common distribution function F. Then, if it exists, the limiting distribution as $u \to \infty$ of $(X - u | X > u)$ is of GPD form:

$$G(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi} \tag{1}$$

where $a_+ = \max(0, a)$ and σ and ξ are scale and shape parameters respectively (see, for example, Coles (2001)).

2.3. Seasonal variation

The time series plot shown in Figure 1 illustrates the seasonal nature of the wind climate at High Bradfield (which is also typical of the seasonal behaviour of the wind climate at most sites in the UK), with the strongest gusts being typically recorded in the winter, particularly December and January. The summer months give rise to more stable climatic conditions and so lower wind speeds are generally observed. The modelling approach we adopt to overcome this problem of non-stationarity is to partition the annual cycle into 12 'seasons', using the calendar month as a natural way of doing this. Experience has shown that for wind speeds, as with many other environmental variables, dividing the year into 12 (roughly) equal-length seasons strikes a good balance between the two conflicting requirements of (a) reflecting reasonably accurately the continuous nature of seasonal changes in climate; and (b) retaining a substantial amount of data for analysis within each season (Walshaw, 1994). Thus, we denote by (σ_m, ξ_m) the GPD scale and shape parameters in month m, $m = 1, \ldots, 12$, fitting a separate GPD to each month of threshold exceedances.

2.4. Temporal dependence

Figure 1 shows a plot of the time series against the series at lag 1. The presence of substantial short-term serial correlation in the sequence of hourly gust maxima is obvious from this plot. The partial autocorrelation function for the hourly gust maxima (not shown) indicates a very large value at lag 1, with all lags greater than 1 showing negligible partial autocorrelation. Three options are available to circumvent the problems posed by short-term serial correlation. As already discussed, the most commonly adopted approach is to employ a declustering scheme to filter out a set of independent threshold exceedances. The second, shown to have substantial benefits over this approach, is to ignore any temporal dependence and fit Equation (1) to the set of all threshold exceedances, but then somehow account for this by adjusting the standard errors attached to maximum likelihood estimates. Though the plot of hourly gust maxima against their lag 1 counterparts in Figure 1, and the partial autocorrelation function, do not necessarily imply that a first-order dependence will hold for *extreme* values in the series, they do at least suggest this as a reasonable assumption for modelling which can (and in this paper will) be assessed later. Thus, a third approach is to model the temporal dependence in the series using a first-order Markov chain model.

3. FIRST-ORDER MARKOV CHAIN MODEL

Work by Fawcett (2005) shows that GPD parameter estimation, and return level estimation, can be sensitive to the choice of declustering scheme used to filter out a set of independent threshold exceedances. We can avoid declustering altogether if we are prepared to make stronger assumptions about our process. Based on the evidence in Figure 1, and inspection of the partial autocorrelation function for our hourly gust maxima, we might assume that the series of gust maxima $X_1, X_2, \ldots$ forms a stationary first-order Markov chain within each season. The stochastic properties of such a chain are completely determined by the joint distribution of successive pairs. Given a model $f(x_i, x_{i+1}; \boldsymbol{\psi})$ specified by parameter vector $\boldsymbol{\psi}$, it follows that the likelihood for $\boldsymbol{\psi}$ is given by

$$L(\boldsymbol{\psi}) = f(x_1; \boldsymbol{\psi}) \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \boldsymbol{\psi}) \bigg/ \prod_{i=1}^{n-1} f(x_i; \boldsymbol{\psi}). \tag{2}$$

To model threshold excesses, we can invoke bivariate extreme value considerations to model contributions to the numerator in the above expression; the denominator is simply replaced by the corresponding univariate densities based on Equation (1).

3.1. Bivariate threshold excess model

Suppose that a GPD with parameters (σ, ξ) is a suitable model for threshold excesses of a threshold u by a variable X. Then from Equation (1), and for $x > u$,

$$\Pr(X > x|X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1/\xi}. \tag{3}$$

Now

$$\Pr(X > x|X > u) = \frac{\Pr(X > x, X > u)}{\Pr(X > u)} \tag{4}$$

$$= \frac{\Pr(X > x)}{\Pr(X > u)} \tag{5}$$

since $x > u$. Thus,

$$\Pr(X < x) = 1 - \lambda_u\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1/\xi} \tag{6}$$

where $\lambda_u = \Pr(X > u)$.

Now suppose $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are independent realisations of a random variable (X, Y) with joint distribution function F. For suitable thresholds u_x and u_y, the marginals for $X - u_x$ and $Y - u_y$ each have an approximation of the form given in Equation (6), with respective parameter sets $(\lambda_{u_x}, \sigma_x, \xi_x)$ and $(\lambda_{u_y}, \sigma_y, \xi_y)$. The transformations

$$\tilde{X} = -\left(\log\left\{1 - \lambda_{u_x}\left[1 + \xi_x\left(\frac{X-u_x}{\sigma_x}\right)\right]^{-1/\xi_x}\right\}\right)^{-1} \qquad \text{and} \tag{7}$$

$$\tilde{Y} = -\left(\log\left\{1 - \lambda_{u_y}\left[1 + \xi_y\left(\frac{Y-u_y}{\sigma_y}\right)\right]^{-1/\xi_y}\right\}\right)^{-1} \tag{8}$$

then induce a variable $(\tilde{X}, \tilde{Y})$ whose distribution function $\tilde{F}$ has margins that are approximately standard Fréchet for $X > u_x$ and $Y > u_y$ (Coles, 2001). It can be shown that the joint distribution function $G(x, y)$ for a bivariate extreme value distribution with standard Fréchet margins has the representation

$$G(x, y) = \exp\{-V(x, y)\} \tag{9}$$

for $x > 0, y > 0$, where

$$V(x, y) = 2\int_0^1 \max(q/x, (1-q)/y)\mathrm{d}H(q), \tag{10}$$

and H is a distribution function on [0,1] satisfying the mean constraint

$$\int_0^1 q\mathrm{d}H(q) = \frac{1}{2}. \tag{11}$$

There is no characterisation of the complete family of distributions specified by Equation (9), and so model choice involves specifying an appropriate sub-family through the choice of H.

3.2. The logistic family

One standard class of parametric families for G is the logistic family, for which

$$V(x, y) = \left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha} \tag{12}$$

for $x > 0$, $y > 0$ and $\alpha \in (0, 1]$. Independence and complete dependence correspond to $\alpha = 1$ and $\alpha \to 0$ respectively. The set-up of this bivariate threshold excess model, using the logistic model, is discussed in greater detail in Coles (2001), and has been used to measure spatial dependence of extremes for pairs of sites in work by, amongst others, Smith and Walshaw (2003). In a time series context and for the first-order Markov chain model being set up here, x and y can be replaced with successive values in time x_i and x_{i+1} (respectively). Then, contributions to the numerator in Equation (2) can be found by differentiation of Equation (9) with respect to both x and y if $x > u_x$ and $y > u_y$, with appropriate censoring if one of either x or y falls below its corresponding threshold. If both x and y lie sub-threshold, the contribution to the numerator in Equation (2) is given by the distribution function evaluated at the thresholds u_x and u_y. Contributions to the denominator in Equation (2) are given by the univariate densities specified by Equation (1). Though the logistic model is the most widely used for modelling multivariate extremes, other models are available which allow for asymmetry in the dependence structure; we consider some of these in Section 5 of this paper.

4. INFERENCE

4.1. Prior information and Bayesian sampling

We now fit the first-order Markov chain model to the wind speed data at High Bradfield. Thresholds u_m, $m = 1, \ldots 12$, are chosen for each month m using mean residual life plots—for more information on these, see, for example, Coles (2001). In the absence of any expert prior information, we specify non-informative prior distributions for the GPD scale and shape parameters and the logistic dependence parameter. For computational convenience, and to retain the positivity of the scale parameter, we work with $\eta = \log(\sigma)$ and specify the following (independent) prior distributions:

$$\pi(\eta_m) \sim N(0, 10000); \tag{13}$$

$$\pi(\xi_m) \sim N(0, 100); \tag{14}$$

$$\pi(\alpha_m) \sim U(0, 1); \tag{15}$$

$m = 1, \ldots, 12$. After setting initial values for both η_m and ξ_m, a simple Metropolis step[1] is used to generate successive draws from the posterior distribution. At each iteration, the simulated values for η_m and ξ_m are used to transform the data to standard Fréchet; a Metropolis step is then similarly used to generate a posterior draw for the logistic dependence parameter α_m using the Markov chain likelihood in Equation (2). Within each of the Metropolis steps, a random walk procedure is used to generate candidate values for each of the parameters, the variances of the innovations being tuned to maximise the efficiency of the algorithm.

For data from each month, for all but nine pairs of observations, likelihood contributions to the Markov chain model component are given by differentiation of Equation (9). However, there are discontinuities between years; for example, we can safely assume the Markov property for consecutive pairs throughout January 1975, but a discontinuity will arise between the last observation from this month (which *should* form a pair with the first observation in February 1975) and the first observation in January 1976. For each of these nine pairs we assume independence, that is assign a value of 1 to α_m in that pair's contribution to the likelihood.

4.2. *Return levels*

Let $X_1, X_2, \ldots, X_n$ be the first n observations from a stationary sequence with marginal distribution function F. Standard arguments in Leadbetter *et al.* (1983, Ch. 3) show that, for large n and x, it is typically the case that

$$\Pr\{\max(X_1, X_2, \ldots, X_n) \leq x\} \approx \{F(x)\}^{n\theta} \tag{16}$$

where $\theta \in (0, 1)$ is the *extremal index* and is a measure of the degree of extremal dependence in the series, and can be interpreted as the reciprocal of the mean cluster size; for a more detailed discussion of the extremal index, see Coles (2001). Setting $x = z_r$ in Equation (16), equating this to $1 - r^{-1}$ and solving for z_r, gives the r-year *return level* of the process, or the value which is exceeded (on average) once every r years. In the context of extreme wind speeds, the 50-year return level is used by the British Standards Institution (1997) to produce contour maps displaying the strength requirements for buildings and other such large structures so that this level of wind speed can be withstood. Thus, the accuracy and precision of return level estimation is an important design requirement.

4.3. *Predictive analysis*

The objective of most extreme value analyses is to obtain an estimate of the probability of future events reaching extreme levels; conveniently, *prediction* is neatly handled within a Bayesian setting. For example, we have seen in Subsection 2.2 that a suitable marginal model for the threshold excess Y of a process is $Y \sim \text{GPD}(\sigma, \xi)$. Estimation of $\boldsymbol{\psi} = (\sigma, \xi, \alpha)$ could be made on the basis of previous observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ using the Markov chain model approach with logistic dependence parameter α quantifying the degree of extremal dependence between successive extremes. Allowing for uncertainty in parameter estimation and future observations,

$$\Pr\{Y \leq y | x_1, \ldots, x_n\} = \int_{\boldsymbol{\Psi}} \Pr\{Y \leq y | \boldsymbol{\psi}\} \pi(\boldsymbol{\psi} | \boldsymbol{x}) \mathrm{d}\boldsymbol{\psi} \tag{17}$$

[1] Details of MCMC techniques are now extensively published (Smith and Roberts (1993), for example) and so are omitted here.

gives the distribution of a future threshold excess. Solving

$$\Pr\left\{Y \leq z_{r,\text{pred}} | x_1, \ldots, x_n\right\} = 1 - r^{-1} \tag{18}$$

for $z_{r,\text{pred}}$ therefore gives an estimate of the r-year return level, which incorporates uncertainty due to model estimation. Although Equation (17) is analytically intractable, it can be approximated since we have estimated the posterior distribution using MCMC. After removal of the 'burn-in' period, the MCMC procedure gives a sample $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(B)}$ that can be regarded as realisations from the stationary distribution $\pi(\boldsymbol{\psi}|\boldsymbol{x})$. Thus

$$\Pr\left\{Y \leq z_{r,\text{pred}} | x_1, \ldots, x_n\right\} \approx \frac{1}{B} \sum_{i=1}^{B} \Pr\left\{Y \leq z_{r,\text{pred}} | \boldsymbol{\psi}^{(i)}\right\} \tag{19}$$

which we can set equal to $1 - r^{-1}$ and solve for $z_{r,\text{pred}}$ using numerical methods (Coles, 2001).

4.4. Results for High Bradfield

Though not shown here, the MCMC sample paths showed rapid convergence to their apparent stationary distributions, with good mixing properties. Table 2 shows posterior means (and standard deviations in parentheses) for the GPD scale and shape parameters and the logistic dependence parameter (for each month), as well as the thresholds used. Table 2 shows that the posterior draws for the logistic dependence parameter typically range between 0.3 and 0.45, representing strong (or at least moderate) temporal dependence for all months. Though these values are of interest in their own right, the increased complexity (and computational expense) of modelling temporal dependence in this way can only be justified if it can be shown that there are real practical advantages over the simpler methods which *ignore* dependence or *filter out* a set of independent exceedances.

As discussed earlier, estimates of return levels are used in a practical setting to specify design requirements for buildings and other such structures. The annual exceedance rate of z_r is given by

Table 2. Posterior means (and standard deviations) of the GPD scale and shape parameters and logistic dependence parameter for each month $m = 1, \ldots, 12$, Also shown are the monthly thresholds.

Month (m)	u_m	σ_m	ξ_m	α_m
1	59.8	8.078 (0.640)	−0.084 (0.057)	0.346 (0.016)
2	45.7	8.123 (0.665)	−0.090 (0.057)	0.368 (0.018)
3	51.3	8.898 (0.668)	−0.339 (0.051)	0.429 (0.019)
4	43.4	8.390 (0.665)	−0.259 (0.054)	0.360 (0.018)
5	35.4	6.843 (0.586)	−0.105 (0.063)	0.354 (0.018)
6	37.9	3.932 (0.350)	0.076 (0.068)	0.449 (0.019)
7	35.1	7.451 (0.566)	−0.397 (0.050)	0.424 (0.020)
8	35.2	6.955 (0.610)	−0.086 (0.063)	0.352 (0.018)
9	35.1	9.894 (0.405)	−0.133 (0.025)	0.294 (0.009)
10	48.4	7.718 (0.634)	−0.217 (0.060)	0.425 (0.018)
11	52.4	7.349 (0.594)	−0.046 (0.052)	0.416 (0.019)
12	53.8	8.332 (0.746)	−0.103 (0.069)	0.370 (0.017)

$$\sum_{m=1}^{12}\{1 - G_m(z_r)^{h_m\theta_m}\}, \quad m = 1, \ldots, 12 \tag{20}$$

where $\{1 - G_m(z_r)^{h_m\theta_m}\}$ is the annual exceedance rate of z_r in month m (from Equation (16)), G_m is the GPD distribution function in month m with parameters σ_m and ξ_m (from Equation (1)) and h_m is the number of hours in month m. By setting this equal to r^{-1}, the solution for z_r is the r-year return level to a good approximation, provided r is large (Coles, 2001). The extremal index in Expression (20) ensures we incorporate the temporal dependence of the series into the estimation of z_r, and is implicitly defined through the value of α_m. Thus, for each posterior draw for α_m we obtain a corresponding posterior draw for θ_m through simulation; for a more detailed discussion about how this is done, see Fawcett (2005). Expression (20) can then be set equal to r^{-1} and solved for z_r for each posterior draw of σ_m, ξ_m and θ_m to obtain posterior draws for the r-year return level z_r.

Figure 2 shows posterior densities for the 10-, 50-, 200- and 1000- year return levels, constructed after the removal of burn-in; the numerical posterior summaries are given later in Table 4 (we report the posterior median rather than the mean due to the severe asymmetry shown in Figure 2). Also shown in Table 4 are the corresponding *predictive* return levels which have taken into account uncertainty in parameter estimation and future observations. Figure 3 shows a plot of these predictive return levels (on the usual -log{-log($1 - r^{-1}$)} scale, given in bold); the corresponding posterior medians for some standard return levels are also shown for comparison, with their associated 95 per cent credible

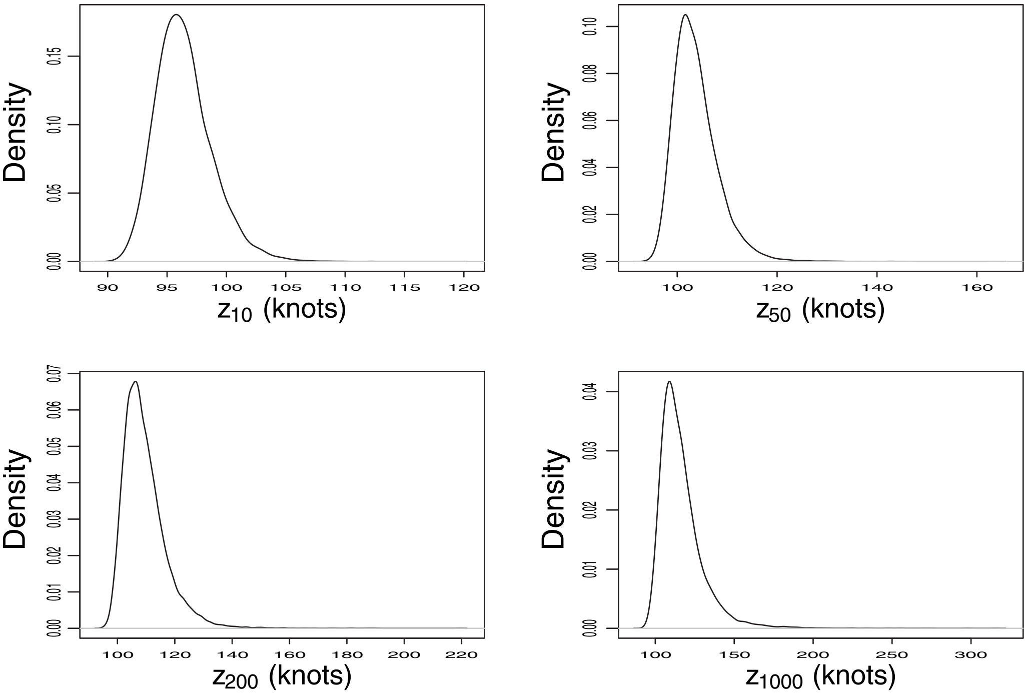

Figure 2. Posterior densities for the 10-, 50-, 200- and 1000-year return levels (z_{10}, z_{50}, z_{200} and z_{1000} respectively).

Table 3. Posterior means (and 95 per cent credible intervals) for dependence parameters in January.

Model	α_1	β_1
Logistic	0.346 (0.312, 0.379)	
Bilogistic	0.413 (0.371, 0.455)	0.405 (0.359, 0.448)
Dirichlet	0.423 (0.374, 0.466)	0.430 (0.381, 0.482)

Table 4. Posterior medians (and 95 per cent credible intervals) for return levels based on the first and second-order Markov chain model; Corresponding *predictive* values are also shown; wind speeds are in knots.

	z_{10}	z_{50}	z_{200}	z_{1000}
First-order model	96.21 (91.69, 102.80)	103.14 (96.61, 115.92)	108.15 (99.63, 131.64)	113.55 (104.57, 149.31)
Predictive ($z_{r,\text{pred}}$)	*100.71*	*111.96*	*124.28*	*144.94*
Second-order model	96.79 (92.03, 102.11)	103.28 (97.53, 115.42)	108.15 (99.91, 129.98)	113.41 (106.47, 148.90)
Predictive ($z_{r,\text{pred}}$)	*100.02*	*111.43*	*123.37*	*143.78*

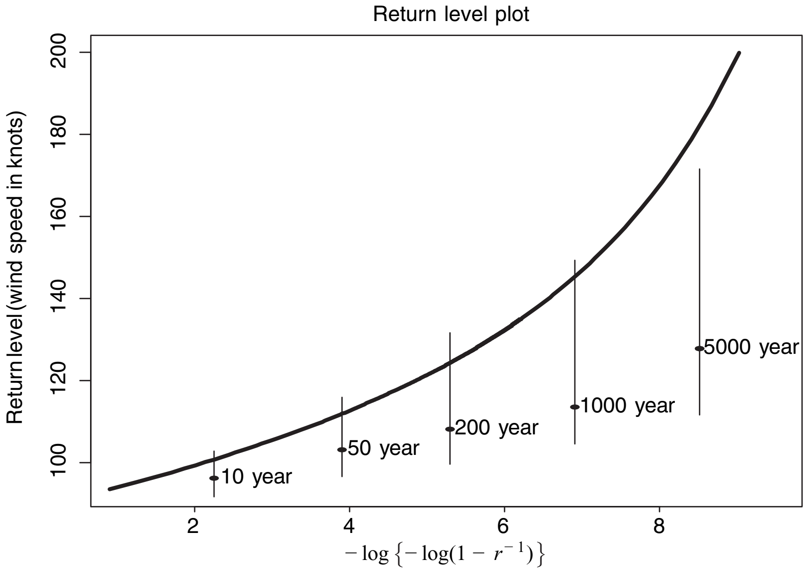

Figure 3. Predictive return level curve (bold line) for High Bradfield, Also shown, for comparison, are posterior medians for some standard return levels with their 95 per cent credibility bands.

intervals. Clearly, designing to this lower level could lead to substantial underprotection due to uncertainty in parameter estimation (particularly for longer period return levels, for which even designing to the upper limit of the 95 per cent credible interval might be inadequate).

5. FURTHER MODELLING CONSIDERATIONS

We now investigate the sensitivity of inferences to the choice of model for consecutive pairs of hourly wind speeds observed at High Bradfield. We also consider the possibility that a higher-order Markov chain model might be more appropriate for the temporal dependence contained in our data, and investigate the ability of such models to capture certain cluster characteristics, information about which is completely lost if temporal dependence is either ignored, or filtered out via some declustering scheme.

5.1. Alternative models for temporal dependence

The main reason for the popularity of the logistic model (12) is its flexibility, with all levels of positive temporal dependence—complete independence to perfect dependence—catered for by the dependence parameter α. The derivation of the joint distribution function for the bivariate logistic model from Equation (10) is not obvious, but it can be shown that Equation (12) can be obtained by letting H have the density function

$$h(q) = \frac{1}{2}(\alpha^{-1} - 1)\{q(1-q)\}^{-1-1/\alpha}\{q^{-1/\alpha} + (1-q)^{-1/\alpha}\}^{\alpha-2} \tag{21}$$

on $0 < q < 1$ (Coles, 2001). The mean constraint (11) is automatically satisfied for this model because of the symmetry about $q = 0.5$. However, one limitation of the logistic model due to the symmetry of h is the fact that the variables x and y in Equation (12) are exchangeable. The following two models are suggested as alternatives to the symmetric logistic model, to account for the possibility of an asymmetric dependence structure.

5.1.1. The bilogistic model. A generalisation of the logistic model that allows for asymmetry in the dependence structure is the bilogistic model, derived by Joe *et al.* (1992). Here, h has the density function

$$h(q) = \frac{1}{2}(1-\alpha)(1-q)^{-1}q^{-2}(1-a)a^{1-\alpha}\{\alpha(1-a) + \beta a\}^{-1} \tag{22}$$

on $0 < q < 1$, where α and β are parameters such that $0 < \alpha < 1$ and $0 < \beta < 1$, and $a = a(q, \alpha, \beta)$ is the solution of

$$(1-\alpha)(1-q)(1-a)^{\beta} - (1-\beta)qa^{\alpha} = 0. \tag{23}$$

When $\alpha = \beta$, this model reduces to the logistic model. The value $\alpha - \beta$ determines the extent of asymmetry in the dependence structure.

5.1.2. The Dirichlet model. An alternative (asymmetric) model proposed by Coles and Tawn (1991) is the Dirichlet[2] model, where

$$h(q) = \frac{\alpha\beta\Gamma(\alpha+\beta+1)(\alpha q)^{\alpha-1}\{\beta(1-q)\}^{\beta-1}}{2\Gamma(\alpha)\Gamma(\beta)\{\alpha q+\beta(1-q)\}^{\alpha+\beta+1}}, \tag{24}$$

and where $0 < q < 1$, and the parameters satisfy $\alpha > 0$ and $\beta > 0$. As with the bilogistic model, the Dirichlet model is symmetric in the case $\alpha = \beta$.

Table 3 shows the results of fitting a Markov chain model to the High Bradfield wind speed data using the three models for dependence discussed so far—the (symmetric) logistic model, the bilogistic model, and the Dirichlet model. We show posterior means and 95 per cent credible intervals for January (hence the parameter subscript 1), though similar model comparisons are observed for other months. The similarity between the dependence parameters within both the bilogistic and Dirichlet models suggests that in fact there is nothing to be gained from allowing asymmetry in the dependence structure; indeed, for both the bilogistic and Dirichlet fits the 95 per cent credible intervals for α_1 and β_1 overlap substantially, and this is repeated across all other months. Recall that, when $\alpha = \beta$, the Dirichlet model is symmetric and the bilogistic model actually reduces to the (symmetric) logstic model given in Equation (12).

5.2. Higher-order Markov chain models

Extensions of the first-order Markov model to a general *d*th order Markov chain are straightforward. The right-hand-side of Equation (2) is simply replaced by

$$\prod_{i=1}^{n-d} f_d(x_i,\ldots,x_{i+d-1};\boldsymbol{\psi}) \Bigg/ \prod_{i=d}^{n-d} f_{d-1}(x_i,\ldots,x_{i+d-2};\boldsymbol{\psi}) \tag{25}$$

where f_i is the joint density of i consecutive observations. Testing between orders $d-1$ and d is equivalent to testing for conditional independence of variables within a unit simplex domain (see Coles and Tawn, 1991). In this Section, we use this approach to compare first-and second-order Markov assumptions for the wind speeds observed at High Bradfield. We also examine the ability of both a first- and second-order Markov chain model to capture information on the clustering of extremes in the wind speed data, which can also be used as a model diagnostic. We then compare return level inference based on fitting the first-order Markov chain model (Subsection 4.4) with that, which uses a second-order Markov assumption.

5.2.1. Fitting a second-order Markov chain model. Using the inferential procedures outlined in Section 4, we now fit a second-order Markov chain model (with logistic dependence structure) to the High Bradfield wind speed data. Posterior means for α (not shown) range from 0.482 (September) to 0.621 (October), where α is now a measure of the joint dependence of successive *triples* of extremes. These values are higher than those based on the first-order fit, reflecting a decay in dependence with an increase in range.

[2] Coles and Tawn use the name Dirichlet since the model is developed by transformation of the standard Dirichlet family of distributions.

5.2.2. Comparison of first- and second-order Markov assumptions. We now assess the suitability of the first-and second-order Markov assumptions for our wind speed data by looking for conditional independence of observations at lag 2. We also use both the first and second-order Markov chain models to estimate two cluster functionals commonly studied in the hydrological literature: (i) storm length and (ii) duration between storms.

5.2.2.1. Conditional independence between observations at lag 2. We first compare the suitability of a first and second-order Markov structure by considering the trivariate distribution of consecutive triples of extremes. Creating a series $\tilde{X}_i$ by transforming the data to unit Fréchet (at sufficiently high levels), under the fitted model the distribution of $(\tilde{X}_i, \tilde{X}_{i+1}, \tilde{X}_{i+2})$ is in the domain of attraction of the time series logistic model of Coles and Tawn (1991), with parameters determined by α. This means that, if we set $R_i = \tilde{X}_i + \tilde{X}_{i+1} + \tilde{X}_{i+2}$, and $W_{i,j} = \tilde{X}_{i+j-1}/R_i$, $j = 1, 2, 3$, then the density of the angular components $(W_{i,1}, W_{i,3})$ is of known form. A plot of $W_{i,1}$ against $W_{i,3}$ (which we call a 'simplex plot' here) can help determine the appropriateness of a first-order Markov assumption relative to a second-order one. If the majority of points lie towards the edges of the plot, pairwise dependence (only) is implied, supporting the adequacy of the first-order Markov chain model; a cluster of points on the interior suggests a three-way dependence (i.e. second-order Markov chain model). For more information on using such plots for model selection, see Coles and Tawn (1991).

Figure 4 shows two simplex plots for the High Bradfield data, one for the Markov chain model fit for January and one for July, though similar plots were obtained for all 12 months. In both we see a tendency for points to lie on the interior of the plot, suggesting that a second-order Markov assumption for the temporal structure of our data might be preferable to a first-order assumption (though in some months this is more obvious than others; for example, the simplex plot for July seems to show a more pronounced clustering of points on the interior than that for January).

5.2.2.2. Storm length. A general summary of cluster behaviour is the distribution, ω, of the number of exceedances per cluster. In a climatological setting, this is sometimes referred to as the 'storm length' (Brabson and Palutikof, 2000). A comparison between the model-based and empirical estimates of ω is made in Figure 5 (top), which shows density plots of ω for the wind-speed data in

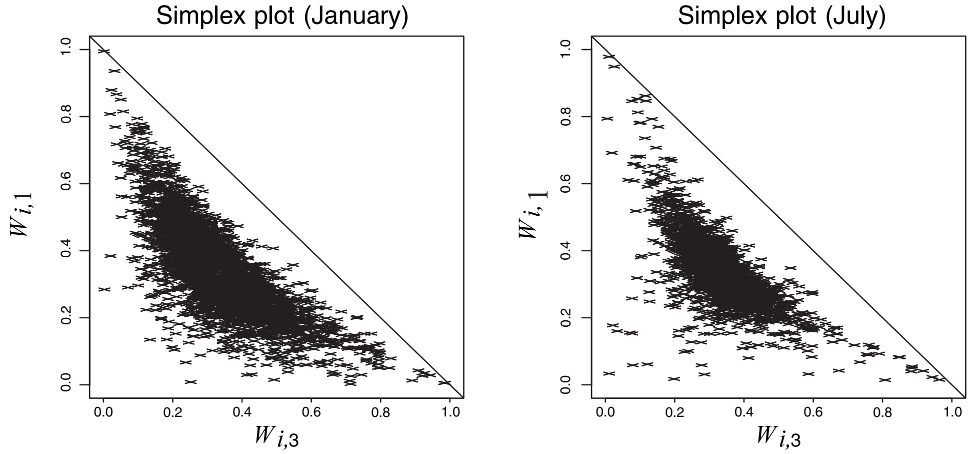

Figure 4. Simplex plots of $W_{i,1}$ versus $W_{i,3}$ for the High Bradfield data: January and July.

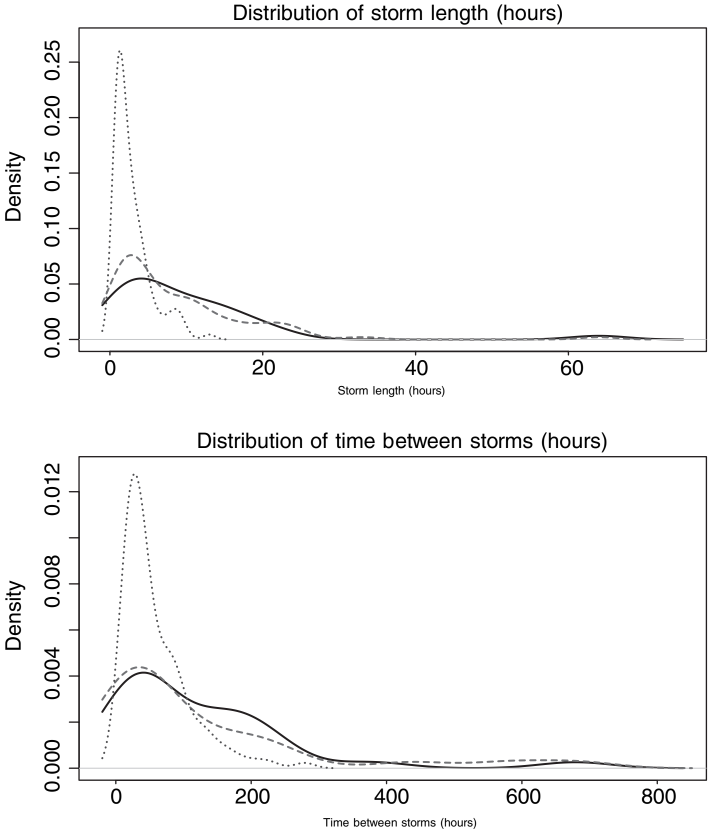

Figure 5. Model-based and empirical estimates of the storm length distribution ω (top) and the distribution of the duration between storms ρ (bottom); the dotted and dashed lines correspond to the first-and second-order fits respectively; the bold lines correspond to the observed distribution.

January (solid line), along with the corresponding density plots based on a simulated first-order Markov chain (dotted line) and a second-order Markov chain (dashed line), the posterior means of α_1 from each of the first and second-order fits to the wind-speed data being used to simulate the chains (with the same marginal GPD parameters as the wind-speed data also). Similar comparisons between the empirical and model-based distribution of ω were made for other months (though storm lengths were generally smaller during summer months). The plot in Figure 5 suggests that the second-order model fits storm lengths much better than does the first-order model, storms typically being of much shorter duration in the first-order model than those observed at High Bradfield.

5.2.2.3. Duration between storms. The performance of the first-and second-order models is also assessed by comparing empirical and model-based estimates of the times between successive storms, the distribution of which we denote by ρ, again a quantity often studied in wind climatology. Figure 5 (bottom) shows the empirical distribution of ρ (solid line) along with model-based estimates using the first- and second-order Markov chain models (dotted and dashed lines respectively). Again, we see a clear preference for the second-order model, which allows for a (much) heavier tail in the distribution of ρ than does the first-order model.

5.2.3. Return level estimation. The diagnostics used in 5.2.2 seem to indicate a preference for the second-order Markov chain model over the simpler first order model for the temporal dependence in the extremes of the wind-speed data. Thus, if we were interested in the clustering behaviour of the wind-speed data *per se*, we would use the second-order model as a basis for our inference. In this situation, the increased complexity of allowing for longer-range dependence in the extremes (with the associated computational expense) is clearly justified. However, if return level estimation is used as a basis for model order selection, Table 4 suggests that increasing the order of our model is unnecessary. Table 4 reports posterior summaries for some return levels obtained for the second-order Markov chain fit, with summaries from the first-order model obtained in Subsection 4.4 also given. We see that the first- and second-order Markov chain models give very similar estimates of return levels, both in terms of posterior location and variability.

ACKNOWLEDGEMENTS

We thank the UK Meteorological Office for supplying the wind-speed data used in this paper. Lee Fawcett's work was funded by an EPSRC Doctoral Training Award.

REFERENCES

Brabson BB, Palutikof JP. 2000. Tests of the generalized Pareto distribution for predicting extreme wind speeds. *Journal of App Met* **39**: 1627–1640.

British Standards Institution 1997. Code of Basic Data for one Design of Buildings: CP 3, Ch. V, Loading; part 2, Wind Loads. British Standards Institution, London.

Coles SG. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer: London.

Coles SG, Tawn JA. 1991. Modelling extreme multivariate events. *J. R. Statist. Soc. B* **53**: 377–392.

Davison AC, Smith RL. 1990. Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc. B* **52**: 393–442.

Fawcett L. 2005. Statistical methodology for the estimation of environmental extremes. *Ph.D Thesis*. University of Newcastle: Newcastle.

Fawcett L, Walshaw D. 2006. Improved estimation for temporally clustered extremes. (in press).

Joe H, Smith RL, Weissman I. 1992. Bivariate threshold models for extremes. *J. R. Statist. Soc. B* **54**: 171–183.

Leadbetter MR, Lindgren G, Rootzén H. 1983. *Extremes and Related Properties of Random Sequences and Series*. Springer-Verlag: New York.

Shellard HC. 1976. Wind. In *The Climate of the British Isles,* Chandler TJ, Gregory S (eds). Longman: London; 39–73.

Smith AFM, Roberts GO. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B* **55**: 3–23.

Smith EL, Walshaw D. 2003. Modelling Bivariate Extremes in a Region. *Bayesian Statistics* **7**: 681–690.

Smith RL. 1991. Regional estimation from spatially dependent data. http://ww.stat.unc.edu/postscript/rs/regest.pdf.

Smithson PA. 1987. An analysis of wind speed and direction at a high-altitude site in the Southern Pennines. *Meteorological Magazine* **116**: 74–85.

Walshaw D. 1994. Getting the most from your extreme wind data: a step by step guide. *Journal of Research of the National Institute of Standards and Technology* **99**: 399–411.