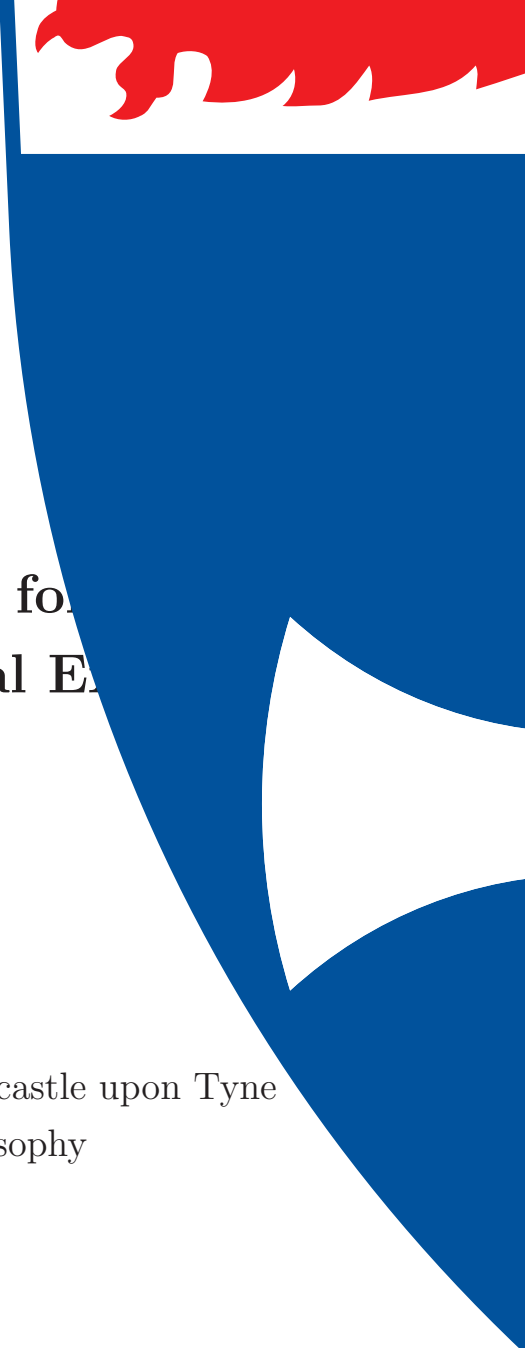




Statistical Methodology for Estimation of Environmental Effects

Lee Fawcett

A thesis submitted to the University of Newcastle upon Tyne
for the degree of Doctor of Philosophy

Statistical Methodology for the Estimation of Environmental Extremes

Lee Fawcett

Summary

In the U.K., the British Standards Institution is responsible for providing codes of practice to regulate structural design standards. At the design stage of any proposed construction, these codes of practice are used by engineers to assess the design strengths necessary for the structure to withstand the wind climate at that site. To date, contour maps displaying strength requirements based on gust speeds which are exceeded (on average) once every fifty years are used to this end, the gust speeds themselves being estimated as the result of fairly rudimentary extreme value analyses. Often, these analyses are based on fitting a limiting extreme value distribution to the set of annual gust maxima.

An improvement on the traditional “annual maxima” approach is to incorporate *all* extremes in the analysis, observations being extreme in the sense that they exceed some pre-determined high threshold. Such “threshold methods” have become increasingly popular over the last decade or so, their main advantage being an increase in precision due to the inclusion of more data. However, using *all* extremes instead of a filtered set of annual extremes can create its own problems; short-term temporal dependence and seasonal variation, for example, become more of an issue. Although a wide range of theoretical models for dependence and non-stationarity have been developed, the classes of limiting distributions under any given model are often too broad for the purposes of statistical inference. It is more common to employ statistical procedures which allow the application of results of extremes of independent and identically distributed sequences.

The issue of short-term temporal dependence is studied in great detail, using real-life wind speed data and simulated data. The inferential consequences of *removing*, *ignoring* and *modelling* such dependence are examined, with particular attention given to the estimation of extreme quantiles. We find that using all excesses – either by disregarding any temporal structure or modelling this structure – provide estimates of model parameters and extreme quantiles which are both more accurate and precise than estimates based on using a filtered set of threshold exceedances.

Until very recently, there has been little effort to implement extreme value models within the Bayesian framework. The advantage of Bayesian inference in the context of extremes seems obvious: since extremes are, by their very nature, scarce, the inclusion of any extra information expressed through prior distributions should add considerably to any analysis. The recent explosion in simulation-based inference through methods such as Markov chain Monte Carlo has made the estimation of complex models within the Bayesian framework possible. Even when genuine prior information is lacking, Bayesian inference is preferred by many as a more informative, complete form of inference than that offered by frequentist or likelihood based techniques. In the context of extremes, the extension to predictive densities within the Bayesian set-up also lends great appeal.

With this in mind, we extend our analysis to include Bayesian hierarchical modelling techniques, and construct a model which attempts to identify site and seasonal effects, while at the same time quantifying the degree of short-term extremal temporal dependence by assuming a Markovian dependence structure. We find that such structured statistical modelling leads to more informative, more accurate and more precise estimates of extremes than the simpler methods currently employed.

Acknowledgements

Firstly, I would like to thank my supervisor, Dave Walshaw, for all his support, encouragement and patience over the last few years. Thanks are due to my advisor Richard Boys, who has always been willing to help when asked, and Liz and Jonathan, whose help has also been appreciated. I am grateful to the U.K. Meteorological Office for supplying the wind speed data used in this thesis, and the EPSRC for funding my studies.

Loads of other people have supported me throughout my Ph.D – not least Jools – who has always listened and cheered me up when things haven’t been going so well. I am grateful to my Mam, Julie, Alan and Gill – and the rest of my family and friends – for their constant encouragement; cheers to Ian, John and Julia who were always willing to listen to my moans over a few pints! Marilyn – thanks for all those words of comfort. Finally, I should also thank Anne Guthrie and Kathy Pike for their motivation all those years ago.

Contents

1	Introduction	1
1.1	A brief history and basic concepts	1
1.2	Classical extreme value theory: key results	3
1.2.1	The ‘extremal types’ characterisation	4
1.2.2	The generalised extreme value distribution	5
1.2.3	Generalisation to the r –largest order statistics	7
1.3	Threshold models for extremes	8
1.4	Dependent and non–stationary sequences	10
1.4.1	Temporally dependent sequences	11
1.4.2	Non–stationarity	14
1.5	Structure of this thesis	15
2	Extreme wind speeds	18
2.1	Introduction	18
2.2	Data description	20
2.2.1	Wind: the variable	20
2.2.2	Site details and data collection	21
2.2.3	Modelling approach	23
2.3	Seasonal variation	24
2.3.1	Generation of a stationary series	24
2.3.2	Implementation for the Bradfield data	25
2.4	Temporal dependence	26
2.4.1	Ignoring dependence: using all threshold excesses	26
2.4.2	Removing dependence: using cluster peak excesses	27
2.5	Threshold selection	28
2.5.1	Implementation for the Bradfield data	28
2.6	Model fitting and verification	31
2.6.1	Maximum likelihood	31
2.6.2	Graphical evaluation	31

2.6.3	Results for the Bradfield data	32
2.7	Return level estimation	34
2.7.1	Methods of estimation	34
2.7.2	Standard errors	36
2.7.3	Implementation for the Bradfield data	36
2.8	Discussion	38
3	Simulation study	42
3.1	Introduction	42
3.2	The simulated data	43
3.2.1	Parametric models for the dependence function	44
3.2.2	GPD marginals	47
3.3	Simulation study details	47
3.4	Results	51
3.4.1	Estimates of σ^* and ξ	51
3.4.2	Return level estimates	53
3.4.3	Performance of estimators	55
3.4.4	Standard errors	56
3.4.5	Results at other levels of dependence	57
3.5	Extensions	60
3.5.1	Varied declustering parameter κ	60
3.5.2	“Automatic” declustering	62
3.5.3	Choice of cluster observation	64
3.6	A model for cluster peak excesses	67
3.7	Discussion	71
4	Markov chain models for threshold exceedances	74
4.1	Introduction	74
4.2	Markov chain models	74
4.2.1	Bivariate threshold excess model	75
4.2.2	Modelling	77
4.3	Implementation	78
4.3.1	Bradfield wind speed data	78
4.3.2	Simulated data	81
4.4	Further considerations	81
4.4.1	Different choices of bivariate extreme value limit	81
4.4.2	Higher-order models	83
4.5	Suitability of the (first-order) Markov assumption at Bradfield	84

4.5.1	Conditional independence between observations at lag 2	84
4.5.2	The extremal index	87
4.5.3	The cluster length distribution	87
4.6	Second-order Markov chain model	88
4.6.1	Implementation for the Bradfield data	91
4.6.2	Model selection	92
4.6.3	Return level estimation	92
4.7	Discussion	94
5	Inference for the extremal index	96
5.1	Introduction	96
5.2	Estimation	97
5.2.1	Cluster size methods	97
5.2.2	Maxima methods	99
5.2.3	An intervals estimator	100
5.3	Performance of estimators	101
5.3.1	Processes with $\theta = 1$	102
5.3.2	Processes with $\theta < 1$	106
5.4	Application to wind speed data	109
5.4.1	Extremal index estimation	110
5.4.2	Consequences for return level estimation	111
5.5	Discussion	113
6	Bayesian inference for extremes	116
6.1	Introduction	116
6.1.1	General theory	116
6.1.2	Markov chain Monte Carlo	117
6.1.3	Bayesian inference for extremes	119
6.2	Implementation for threshold excesses	120
6.2.1	Analysis of simulated data	121
6.2.2	Analysis of Bradfield wind speed data	122
6.3	Bayesian inference for measures of extremal dependence	127
6.3.1	Inference for Markov chain models	127
6.3.2	Bayesian methods for the extremal index	134
6.3.3	A Bayesian declustering scheme for dependent extremes	138
6.4	Discussion	148

7	A hierarchical model for extreme wind speeds	150
7.1	Introduction	150
7.2	The wind speed data	151
7.3	A hierarchical model	154
7.3.1	Model specification	154
7.3.2	Inference	155
7.3.3	Results	157
7.3.4	Diagnostics	164
7.3.5	Return level estimation	165
7.3.6	Prediction	166
7.4	Simulation study	168
7.5	Discussion	171
A	Appendix to chapter 3	172
A.1	Simulation study results – first run, 15 studies	172
A.2	Distribution of cluster peak excesses	185

Chapter 1

Introduction

1.1 A brief history and basic concepts

Generally, extreme value theory can be defined as the study of the largest (or smallest) events occurring in any random process. The field has experienced a period of enormous growth over the last decade or so, with researchers playing a more significant role than ever in developing the applications and methodology of extreme value theory. Recent developments in simulation-based inference have enabled substantial increases in the complexity of statistical models that can be built and estimated, and have opened up the floodgates to the Bayesian approach to statistical inference of extreme values. As a branch of statistics, extreme value theory has found widespread application in the modelling of various natural and man-made phenomena. Common areas of application are environmental design and danger assessment, finance and insurance, survival analysis, and sport. For example, in environmental design, extreme value theory can be used to estimate the probabilities of extremes of various environmental processes, and so provide information which can aid the design of, amongst other things,

- sea-walls to withstand the highest sea-levels in a year;
- road surfaces to withstand the highest and lowest temperatures over several years;
- oil rigs, and other offshore structures, to stand up to the strongest sea currents;
- buildings and bridges so that they will not fail in the strongest winds;
- reservoir capacities to cope with extreme periods of drought.

In environmental danger assessment, extreme value techniques are used to help provide legislation, advice or public warnings relating to (for example) unacceptably high pollution levels, or the number of deaths which might occur due to a single natural or technological

disaster. In the field of finance, extreme value theory is often used to predict probabilities of rare events, such as the chance of a devastating stock market crash, or the chance of very large insurance claims in a short space of time. The study of extremes has also found its way into the survival literature, where the Weibull distribution – one of the ‘extreme value distributions’ – is often used to model remaining lifetimes. More light-hearted applications have addressed questions such as “is there an ultimate time in which the mile can ever be run?”, “what is the oldest age a human can ever live to?”, and “what is the distribution of the highest scores in a group of football matches?”. In most of these settings, the motivation for an extreme value analysis is to extrapolate a probability model beyond the range of observed data. In the absence of any physical or empirical basis for such an extrapolation, it is usual to fall back on asymptotic theory to provide tail models. Despite the asymptotics, the application of such models to domains on the fringe of, or beyond, observed data is seen by some as scientifically unreasonable; however, without any viable alternative, we are reliant upon such models to address real-life problems such as those already discussed. If the extremes of some sample appear to be well fitted by an asymptotic distribution, then there is reason to hope that arguments which suggest such a distribution have a sound basis, so that we can take high quantiles in the fitted distribution as serious estimates for levels which might be over-topped in any real-life scenario.

In one of the first books on the statistics of extreme value theory, Gumbel (1958) traces the origins back to 1709 when Nicolaus Bernoulli considered the problem of estimating the age of the longest survivor in a group of people. Research into extreme values as a subject in its own right began much later, between 1920 and 1940, when work by E.L. Dodd, M. Fréchet, E.J. Gumbel, R. von Mises and L.H.C. Tippett investigated the asymptotic distribution of the largest order statistic. The main theoretical result, which became known as the Extremal Types Theorem, was developed in stages by R.A. Fisher, L.H.C. Tippett and R. von Mises, and eventually proved by B. Gnedenko in 1943. Up until 1950, development in the field was largely theoretical. Although Gumbel’s book (1958) provided the starting point for the application of this theory – mainly to engineering problems – it was a while before sophisticated statistical models were developed (though Jenkinson worked on the application of extreme value models to extreme wind speeds in the 1950s). In the early 1970s, L. de Haan and J. Pickands generalised the theoretical results, and this gave a solid foundation for the advance of statistical models. The development of statistical methodology was still slow, but since the early 1980s, methods for the application of extreme value theory have become increasingly widespread. Current researchers who have played a significant role in developing the applications and methodology of extreme value theory include H. Rootzén, R.L. Smith, J.A. Tawn, A.C. Davison and S.G. Coles.

A broad and comprehensive literature on the subject of extreme value theory now exists. The book by Coles (2001), which adopts a heuristic approach to learning the subject in preference to a more rigorous mathematical exposition, provides an excellent starting point; for those who wish to divulge in the more demanding mathematical aspects of the probability theory behind extreme values of stochastic processes, the book by Leadbetter *et al.* (1983) is recommended. The seminal paper by Davison and Smith (1990) affords a discussion of models for excesses over high thresholds with the ability to account for serial dependence and seasonality, and was an important paper aimed at redirecting statisticians' thinking about extremes away from the traditional 'annual maxima' approach. Publications specifically dealing with extremes of multivariate events include papers by Smith *et al.* (1990), Coles and Tawn (1991), Tawn (1988) (specifically concerned with the bivariate case) and, more recently, the paper by Heffernan and Tawn (2004) which looks at a conditional approach for multivariate extremes. Papers which link the themes of extreme value modelling and Bayesian inference are sparse, and are restrictive in either the extreme value family adopted, or in the informativeness of the prior specification; for a review, see Coles and Powell (1996). Papers by Coles and Tawn (1996) and Smith and Walshaw (2003) make an attempt to elicit expert prior opinion, the latter in a multivariate framework. The Ph.D thesis of Walshaw (1991), and several subsequent papers by the same author, have investigated various methods for modelling extreme wind speeds; books by Lawson (1980) and Simiu and Scanlan (1996) give detailed accounts on the effects of extreme wind speeds on structures.

1.2 Classical extreme value theory: key results

In this section, we summarise some of the theoretical results which are fundamental to 'classical' extreme value theory. The generalised extreme value distribution, which uses block (or annual) maxima, is outlined, and statistical techniques for fitting this model discussed. We then show, in subsequent sections, how this model can be extended to allow several of the largest order statistics (occurring over a fixed time) to be incorporated into an extremal analysis. Finally, the situation where extremes are no longer independent and may originate from non-stationary series is considered.

Suppose that $X_1, X_2, \dots, X_n$ is a sequence of independent and identically distributed (IID) random variables with common distribution function F . One way of characterising extremes is by considering the distribution of the maximum order statistic

$$M_n = \max\{X_1, X_2, \dots, X_n\}. \quad (1.1)$$

This is trivial (in principle), since

$$\begin{aligned}\Pr\{M_n \leq x\} &= \Pr\{X_1 \leq x, X_2 \leq x, \dots, X_n \leq x\} \\ &= \Pr\{X_1 \leq x\} \times \Pr\{X_2 \leq x\} \times \dots \times \Pr\{X_n \leq x\} \\ &= \{F(x)\}^n.\end{aligned}$$

However, in practice the distribution function F is unknown. This leads to an approach based on asymptotic arguments – specifically, limiting distributions for M_n as $n \rightarrow \infty$.

1.2.1 The ‘extremal types’ characterisation

The obvious question now, is, what possible distributions might be considered candidates for the distribution for M_n as $n \rightarrow \infty$? It might also be convenient to find out if we can formulate this set of candidate distributions into a single class – say G – which is independent of F ; if so, we can estimate the distribution of M_n using G , without any reference to F . Clearly, the limiting distribution of M_n is degenerate, since M_n converges with probability 1 to the upper endpoint of F . This is analogous to the sample mean $\bar{X}$ converging to the population mean μ with certainty in the Central Limit Theorem; here, the degenerate limit is prevented by allowing a linear rescaling, so that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1).$$

Similarly, we look for limiting distributions as $n \rightarrow \infty$ of normalised sequences $(M_n - b_n)/a_n$. The univariate classification (due to Fisher and Tippett in 1928 and formally proved by Gnedenko in 1943) is summarised by the ‘extremal types theorem’; before stating the result, it is important to define an equivalence class of distributions: F and F^* are *type equivalent* if there exists constants a and b such that $F^*(ax + b) = F(x)$, for all x .

Theorem 1.1 (*Extremal types theorem*)

If there exist sequences of constants $a_n > 0$ and b_n such that, as $n \rightarrow \infty$,

$$\Pr\{(M_n - b_n)/a_n \leq x\} \rightarrow G(x) \tag{1.2}$$

for some non-degenerate distribution G , then G is of the same type as one of the following distributions:

$$I : G(x) = \exp\{-\exp(-x)\} \quad -\infty < x < \infty; \tag{1.3}$$

$$II : G(x) = \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0, \alpha > 0; \end{cases} \tag{1.4}$$

$$III : G(x) = \begin{cases} \exp\{-(-x)^\alpha\} & x < 0, \alpha > 0 \\ 1 & x \geq 0. \end{cases} \tag{1.5}$$

Conversely, each of these distributions, G , may appear as a limit for the distribution of $(M_n - b_n)/a_n$ and, in fact, does so when G itself is the distribution of X in equation (1.1).

The three types of distribution in Theorem 1.1 (*I*, *II* and *III*) have become known as the Gumbel, Fréchet and Weibull types (respectively), and are known collectively as the *extreme value distributions*. It should be noted that Theorem 1.1 does not *ensure* the existence of a non-degenerate limit for M_n ; nor does it specify *which* of types *I*, *II* or *III* is applicable *if* a limit distribution *does* exist (i.e. in which *domain of attraction* the distribution of G lies). However, when such a distribution does exist, we find that, by analogy with the Central Limit Theorem, the limiting distribution of sample maxima follows one of the distributions given in Theorem 1.1 no matter what the parent distribution F . See Leadbetter *et al.* (1983) for a detailed exposition of the existence of limits and domains of attraction.

An equivalent characterisation to the extremal types theorem is obtained via the class of *max-stable distributions*. A distribution G is said to be *max-stable* if, for $n = 2, 3, \dots$, there are constants $\alpha_n > 0$ and β_n such that $G^n(\alpha_n x + \beta_n) = G(x)$. The following theorem (Theorem 1.4.1 in Leadbetter *et al.* (1983)) establishes the connection between the max-stability property and extreme value theory:

Theorem 1.2

A distribution is max-stable if, and only if, it is of the same type as an extreme value distribution.

It follows from Theorem 1.2 that the distribution of the maximum of independent observations will be of the same type as that of the underlying population if, and only if, the underlying population itself has a distribution of extreme value type.

1.2.2 The generalised extreme value distribution

In practice, it is inconvenient to work with three separate classes of limiting distribution (as in theorem 1.1). However, there exists a parameterisation which encompasses all three types of extreme value distribution. Von Mises (1954) and Jenkinson (1955) independently derived the *generalised extreme value distribution* (GEV), denoted $\mathcal{G}(\mu, \sigma, \xi)$, whose cumulative distribution function is given by

$$G(x; \mu, \sigma, \xi) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\}, \quad (1.6)$$

where $a_+ = \max(0, a)$. The situation where $\xi = 0$ is not defined in (1.6), but is taken as the limit as $\xi \rightarrow 0$, given by

$$G(x; \mu, \sigma) = \exp \left\{ -\exp \left(\frac{x - \mu}{\sigma} \right) \right\}. \quad (1.7)$$

The parameters μ ($-\infty < \mu < \infty$), σ (> 0) and ξ ($-\infty < \xi < \infty$) are location, scale and shape parameters respectively. The value of the shape parameter ξ differentiates between the three types of extreme value distribution in Theorem 1.1: $\xi = 0$, leading to (1.7), corresponds to the Gumbel distribution (type *I*); $\xi > 0$ and $\xi < 0$ correspond to the Fréchet (type *II*) and Weibull (type *III*) distributions respectively.

A typical application of the generalised extreme value distribution is to fit (1.6) to a series of annual (or block) maxima. Estimates of extreme quantiles of these annual maxima are then obtained by inversion of (1.6), to give

$$q_r = \mu - \frac{\sigma}{\xi} \left[1 - \{-\log(1 - 1/r)\}^{-\xi} \right] \quad (1.8)$$

for $\xi \neq 0$, where $G(q_r) = 1 - 1/r$. In the case of $\xi = 0$, inverting (1.7) gives

$$q_r = \mu - \sigma \log [-\log(1 - 1/r)]. \quad (1.9)$$

In extreme value terminology, q_r is the return level associated with the r -year return period, and practitioners commonly extrapolate the relationships given in (1.8) and (1.9) to obtain return level estimates considerably beyond the range of data observed.

Numerous ways of fitting (1.6) to a set of annual maxima have been considered, though numerical maximum likelihood estimation has become generally accepted as the most robust procedure (Hosking (1990) argues that using L -moments can provide a more robust estimative procedure in small samples; applications of L -moments can be found in Guttman *et al.* (1993) and Hosking and Wallis (1993)). A potential difficulty with the use of likelihood methods for the GEV concerns the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid. Such conditions are not satisfied by the GEV because the end-points of the GEV distribution are functions of the parameter values: $\mu = \sigma/\xi$ is an upper end-point of the distribution when $\xi < 0$, and a lower end-point when $\xi > 0$. This violation of the usual regularity conditions means that the standard asymptotic likelihood results are not automatically applicable. Smith (1985) found that, for $-1 < \xi \leq -0.5$, maximum likelihood estimators are obtainable but, in general, do not have all of the standard asymptotic properties: estimators are ‘super-efficient’ (i.e. have variances smaller than the Cramér–Rao lower bound) and are not asymptotically normal. Further, he found that maximum

likelihood estimators for $\xi \leq -1$ do not, in general, exist. However, when $\xi > -0.5$, though the classical regularity conditions are still not satisfied, the information matrix is finite and the classical asymptotic properties continue to hold.

The recent explosion in simulation-based inference and the advent of Markov chain Monte Carlo algorithms have offered an alternative way to make inferences from the likelihood function, which many practitioners prefer. Although (in most environmental applications) instances of $\xi < -0.5$ are rare, they are not unheard of, and so here the Bayesian framework provides a viable inferential alternative to maximum likelihood estimation. A maximum likelihood estimate (or posterior Bayesian estimate) for the r -year return level q_r can be obtained by substituting maximum likelihood estimates (or posterior means, perhaps) for μ , σ and ξ into (1.8) or (1.9), and standard errors for estimates of q_r estimated via the delta method in the maximum likelihood approach (Rao, 1973). Asymmetry in the likelihood surface suggests that confidence intervals based on the profile likelihood, rather than the usual Wald interval ($\text{MLE} \pm 1.96 \times \text{estimated standard error}$), are often much more accurate, though in a Bayesian setting this does not apply and we can obtain confidence intervals directly from the posterior distribution for q_r .

Diagnostic checks for model goodness-of-fit usually consist of a comparison of empirical and model-based estimates of the distribution of M_n . Obviously, the model is required for extrapolation, but it should also fit reasonably well to the most extreme of the observed data. Comparisons of quantiles based on the GEV and the empirical distribution function in a quantile-quantile plot give an assessment of model fit. In the same way, empirical and model-based distributions can be compared in a probability plot.

1.2.3 Generalisation to the r -largest order statistics

Given a sequence of IID random variables $X_1, X_2, \dots$ we have shown how the generalised extreme value distribution (expression 1.6) can be used to model the set of normalised annual maxima. This approach is highly inefficient since all but the maximum in each year (or block) are discarded – other observations which could be considered extreme are simply thrown away because they are not as extreme as the maximum value in that year. A generalisation of the result in expression (1.6) attempts to overcome this, by incorporating the largest r order statistics from each year, where r is any positive integer. If we denote the r largest order statistics in an IID sample by $M^{(1)} \geq M^{(2)} \geq \dots \geq M^{(r)}$, for $r \geq 1$, then the technique here is to obtain the limiting joint distribution of

$$\left(\frac{M_n^{(1)} - b_n}{a_n}, \frac{M_n^{(2)} - b_n}{a_n}, \dots, \frac{M_n^{(r)} - b_n}{a_n} \right). \quad (1.10)$$

It can be shown that the complete class of limiting non-degenerate joint distributions is in fact given by the probability density function

$$f(x_1, x_2, \dots, x_r; \mu, \sigma, \xi) = \sigma^{-r} \exp \left\{ - \left[1 + \xi \left(\frac{x^{(r)} - \mu}{\sigma} \right) \right]_+^{-1/\xi} - \left(1 + \frac{1}{\xi} \right) \sum_{j=1}^r \log \left[1 + \xi \left(\frac{x^{(j)} - \mu}{\sigma} \right) \right]_+ \right\} \quad (1.11)$$

for $j = 1, \dots, r$. As before, the case $\xi = 0$ is taken as the limit as $\xi \rightarrow 0$ in (1.11), to give

$$f(x_1, x_2, \dots, x_r; \mu, \sigma, \xi) = \sigma^{-r} \exp \left\{ - \exp \left[- \left(\frac{x^{(r)} - \mu}{\sigma} \right) \right] - \sum_{j=1}^r \left(\frac{x^{(j)} - \mu}{\sigma} \right) \right\}. \quad (1.12)$$

Obviously, the case where $r = 1$ is equivalent to the annual maxima approach, for which the GEV holds as the limiting distribution. The increased precision over the traditional annual maxima approach (due to more extremes being incorporated into the analysis) has obvious appeal; however, Smith (1986) shows that as r increases, the rate of convergence to the limiting distribution decreases rapidly, and so the number of order statistics to include must be considered carefully. Such methods must also take account of serial correlation, and are vulnerable to the effects of seasonal variation. Papers in the Journal of Hydrology by Smith (1986) and Tawn (1988b) illustrate the use of r -largest methods; Coles' web notes (1999) also apply the technique to the 10 largest sea levels each year in Venice. However, apart from these few examples, little else can be found in the literature.

1.3 Threshold models for extremes

Generalisation of the classical annual maxima approach for modelling extreme values to the r -largest order statistics method has been discussed, the main advantage being the inclusion of more extreme data in the analysis, leading to more precise inferences on the extremal behaviour of the process under study. However, only the r largest values within each year (or block) are used, and any other extremes discarded. In the present section, we discuss an approach which aims to include *all* extreme values in the analysis, extreme in the sense that they exceed some pre-determined high threshold level.

Threshold methods developed rapidly during the 1980s, culminating in the Davison and Smith (1990) paper which applied these techniques to an environmental data set which displayed short-term serial dependence and seasonal variation. Since then, threshold

methods have become the standard tool for many practitioners involved in modelling extreme values. Relative to the traditional annual maxima and r -largest approaches, threshold methods maximise efficiency by using all extreme values in their analysis; however, as shall be seen in chapter 2 when these methods are applied to extreme wind speed data, the fact that we use all extremes can itself create problems.

The generalised Pareto distribution

Ignoring, for now, the practical implications of using all our extreme data, again consider a sequence of IID random variables $X_1, X_2, \dots, X_n$ with common distribution function F . Then for a sufficiently large threshold u , the distribution of $(X - u)$, conditional on $X > u$, is approximately

$$G(y; \bar{\sigma}, \xi) = 1 - \left(1 + \frac{\xi y}{\bar{\sigma}}\right)_+^{-1/\xi}, \quad (1.13)$$

where $\bar{\sigma}$ (> 0) and ξ ($-\infty < \xi < \infty$) are scale and shape parameters respectively. The shape parameter ξ in the GPD takes exactly the same value as that for the corresponding GEV distribution; the scale parameter $\bar{\sigma}$ is equal to $\sigma + \xi(u - \mu)$, where σ and μ are the scale and location parameters (respectively) in the corresponding GEV distribution (see expression 1.6). Specifically, G is defined on $0 < y < \infty$ if $\xi > 0$, and $0 < y < -\bar{\sigma}/\xi$ if $\xi \leq 0$. The case $\xi = 0$ is interpreted as the limit $\xi \rightarrow 0$, and is the exponential distribution with rate $1/\bar{\sigma}$. This is known as the *generalised Pareto distribution*, or GPD. The GPD is a limiting distribution for excesses over thresholds if, and only if, the parent distribution lies in the domain of attraction of one of the three extreme value distributions (see Theorem 1.1). However, since the limiting distribution of sample maxima follows one of the distributions given in Theorem 1.1 no matter what the parent distribution, the GPD is the *only* non-degenerate limiting distribution for excesses over thresholds of IID sequences. Until this point we have used the notation $\bar{\sigma}$ to denote the scale parameter for the GPD, so as to distinguish it from the corresponding parameter of the GEV distribution. For notational convenience we now drop this distinction, using σ to denote the scale parameter within either family.

The GPD yields several important properties. The first of these, known as the ‘threshold stability property’, is that if $(X - u)$ follows a generalised Pareto distribution (conditional on $X > u$), then $(X - u')$ also follows a generalised Pareto distribution for any $u' > u$. In other words, once a suitably high enough threshold has been found such that the GPD may be assumed a valid model for excesses over that threshold, then the GPD holds for excesses over any higher threshold too. Another property unique to the GPD is that if $N \sim \text{Poisson}$, and $X_1, \dots, X_N$ are IID random variables following a GPD, then

$\max\{X_1, \dots, X_N\}$ has the GEV distribution. A proof of the threshold stability property is given in section 3.3, and, as will be shown, this property is exploited in graphical procedures for threshold selection and assessing the fit of the GPD. The second property suggests that, if the exceedances of u occur as a Poisson process with threshold excesses which are IID and generalised Pareto distributed, then the maximum value over any block size has a generalised extreme value distribution. Thus, if we assume extreme events occur over time as a Poisson process, models which fit the GEV to sets of block maxima are consistent with models which fit the GPD to sets of threshold excesses.

As with fitting the GEV, most practitioners use numerical maximum likelihood estimation to fit (1.13) to a set of threshold excesses. For $-1 < \xi \leq -0.5$, maximum likelihood estimators exist (in large enough samples), though in general (as before) do not possess all of the standard asymptotic properties; when $\xi \leq -1$, maximum likelihood estimators do not, in general, exist. For $\xi > -0.5$, maximum likelihood estimators are asymptotically normal and efficient (Smith, 1985). Luckily, in most environmental applications, values of $\xi \leq -0.5$ are rare, but do occur from time to time. Again, the Bayesian methodology provides a framework within which this problem can be avoided. However, in a Bayesian setting, use of the $\text{GPD}(\sigma, \xi)$ model may be restrictive since the scale parameter σ is dependent on the choice of threshold level u ; an uninformative prior for σ then becomes informative at higher thresholds. To overcome this, the GPD can be reparameterised with scale and shape parameters $\tilde{\sigma}$ and ξ (respectively, ξ remaining unchanged), where $\tilde{\sigma} = \sigma - \xi u$. Under this parameterisation, both parameters are threshold-independent. Section 2.7 shows how estimates of extreme quantiles can be obtained through inversion of (1.13).

1.4 Dependent and non-stationary sequences

The asymptotic results discussed so far have assumed the underlying process to be independent and identically distributed. They also assume this process is stationary. In practice, extreme value data – particularly environmental time series – exhibit some form of departure from this ideal. The most common forms are:

- Local temporal dependence, where successive values of the time series are dependent, but values farther apart are independent (to a good approximation);
- Long term trends, where the underlying distribution changes gradually over time;
- Seasonal variation, where the underlying distribution changes periodically through time.

These departures can be handled through a combination of extending both the theory and the modelling. However, although a wide range of theoretical models for non-stationarity have been studied, only in a few cases have these been used for statistical modelling; the results have generally been too specific to be of use in modelling data for which the form of non-stationarity is unknown. Over the last decade or so, it has been more usual for practitioners to employ statistical procedures which allow the existing results to be applied. The rest of this section discusses some of the theoretical results which can be applied to non-IID data, and some of the techniques used in practice which allow the application of these results.

1.4.1 Temporally dependent sequences

The book by Leadbetter *et al.* (1983) considers, in great detail, properties of extremes of dependent processes. A key result often used is ‘Leadbetter’s $D(u_n)$ condition’, which ensures that long-range dependence is sufficiently weak so as not to affect the asymptotics of an extreme value analysis. This condition is stated more formally in Definition 1.1 below.

Definition 1.1

A stationary series $X_1, X_2, \dots$ is said to satisfy the $D(u_n)$ condition if, for all $i_1 < \dots < i_p < j_1 < \dots < j_q$ with $j_1 - i_p > l$,

$$\left| \Pr \{X_{i_1} \leq u_n, \dots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \dots, X_{j_q} \leq u_n\} - \Pr \{X_{i_1} \leq u_n, \dots, X_{i_p} \leq u_n\} \Pr \{X_{j_1} \leq u_n, \dots, X_{j_q} \leq u_n\} \right| \leq \alpha(n, l), \quad (1.14)$$

where $\alpha(n, l) \rightarrow 0$ for some sequence l_n such that $l_n/n \rightarrow 0$ as $n \rightarrow \infty$.

For sequences of independent variables, the difference in probabilities expressed in (1.14) is exactly zero for *any* sequence u_n . More generally, we will require that the $D(u_n)$ condition holds only for a specific sequence of thresholds u_n that increases with n . For such a sequence, the $D(u_n)$ condition ensures that, for sets of variables that are far enough apart, the difference in probabilities expressed in (1.14), while not zero, is sufficiently close to zero to have no effect on the limit laws for extremes. The $D(u_n)$ condition is a weaker condition than that of *strong mixing*, first introduced by Rosenblatt (1956) and discussed in detail in Leadbetter *et al.* (1983).

Theorem 1.3

Let $\tilde{X}_1, \tilde{X}_2, \dots$ be a stationary series satisfying Leadbetter's $D(u_n)$ condition, and let $\tilde{M}_n = \max\{\tilde{X}_1, \dots, \tilde{X}_n\}$. Now let $X_1, X_2, \dots$ be an *independent* series with X having the same distribution as $\tilde{X}$, and let $M_n = \max\{X_1, \dots, X_n\}$. Then if M_n has a non-degenerate limit law given by $\Pr\{(M_n - b_n)/a_n \leq x\} \rightarrow G(x)$, it follows that

$$\Pr\{(\tilde{M}_n - b_n)/a_n \leq x\} \rightarrow G^\theta(x) \quad (1.15)$$

for some $0 \leq \theta \leq 1$.

The parameter θ is known as the *extremal index*, the concept of which originated in papers by Cartwright (1958), Newell (1964), Loynes (1965), O'Brien (1974) and Leadbetter *et al.* (1983). Due to the max-stability property mentioned in Theorem 1.2, and since G in Theorem 1.3 is necessarily an extreme value distribution, then the distribution of maxima in processes displaying short-range temporal dependence (characterised by the extremal index θ) is also a GEV distribution; the powering of the limit distribution by θ only affects the location and scale parameters of this distribution. Consequently, regardless of any extremal dependence in the process, the theory developed for independent processes can be applied to processes which display short-term temporal dependence; if we are not interested in the degree of extremal clustering of the process, we can therefore ignore the extremal index and simply fit a GEV to the set of annual maxima in the usual way. The only difference is that our implied n – the number we are taking the maximum over – is now effectively reduced due to the dependence, so convergence of maxima to the limit distribution will be slower.

Though the modelling procedure for fitting the GEV to a set of annual maxima is unchanged for series which display short-term temporal dependence, some revision is needed of the threshold exceedance approach. If all threshold exceedances are used in our analysis, and the GPD fitted to the set of threshold excesses, the likelihoods we use will be incorrect since they assume independence of sample observations. In practice, several techniques have been developed to circumvent this problem; the most commonly adopted being to formally identify clusters of extremes, and fit the GPD to the set of ‘independent’ *cluster peak* excesses. This approach is easy to implement, but similar to the annual maxima approach it is wasteful of data, with all but the maximum value within each cluster being discarded. Cluster identification is also often arbitrary, though Ferro and Segers (2003) propose an “automatic” declustering scheme (see section 3.5.2). Other approaches include fitting the GPD to the set of all threshold excesses (assuming independence), but then appropriately adjusting the inference to take into account the reduction in information. This approach avoids the need to decluster (and so make arbitrary decisions about

what constitute independent clusters of extremes), but the likelihood asymptotics are less rigorous and, as with the filtering approach, we lose all information about the dependence of the process. A third, more computationally expensive method, is to explicitly model the temporal dependence of the process via a parametric stochastic model (for example, a Markov chain) – the obvious advantage here being the fact that dependence is properly accounted for through the likelihood. The inferential consequences of all three of these approaches (*removing*, *ignoring* and *modelling* temporal dependence) are investigated in chapters 2, 3 and 4 of this thesis; practical issues, such as declustering techniques (used when removing dependence), variance inflation procedures (used when ignoring dependence) and the implementation of Markov chain models, are considered in detail when they are used. Procedures for estimating the extremal index are considered in chapter 5.

The role of the extremal index

The role of the extremal index is easily seen through the following simple examples.

Example 1.1 *Independent process*

Consider an independent process, i.e. a process for which $X_i = \tilde{X}_i$. Then, since $G(x) = G^\theta(x)$, the extremal index $\theta = 1$.

Example 1.2 *Simple dependence process*

Let $X_1, X_2, \dots, X_{n+1}$ be an IID sequence with distribution function $F_X(x) = \exp\{-1/x\}$, $x \geq 0$. Define a sequence $Z_1, Z_2, \dots, Z_n$ by $Z_i = \max\{X_i, X_{i+1}\}$, $i = 1, 2, \dots, n$. The marginal distribution of Z_i is

$$\begin{aligned} F(z) &= \Pr(Z_i \leq z) \\ &= F_X(z)^2 \\ &= \exp\{-2/z\}. \end{aligned}$$

Define the sequence $\tilde{Z}_1, \tilde{Z}_2, \dots, \tilde{Z}_n$ to be IID with marginal distribution function F . Then, using $a_n = n$ and $b_n = n$, we have

$$\begin{aligned} \Pr\{(\tilde{M}_n - b_n)/a_n \leq z\} &= F^n(nz + n) \\ &= \exp\{-2(1+z)^{-1}\} \\ &= G(z), \end{aligned}$$

and

$$\begin{aligned} \Pr\{(M_n - b_n)/a_n \leq z\} &= \Pr(\max\{X_1, X_2, \dots, X_{n+1}\} \leq (nz + n)) \\ &\rightarrow \exp\{-(1+z)^{-1}\} \quad \text{as } n \rightarrow \infty \\ &= G^{\frac{1}{2}}(z). \end{aligned}$$

Thus, the extremal index for this process is $\theta = 1/2$. More generally, increased dependence in the extreme levels of the sequence corresponds to decreased θ . An alternative interpretation of θ is that θ^{-1} is the limiting mean size of clusters of exceedances above a threshold, as that threshold tends to the upper endpoint of Z_1 (Dixon, 1998). In the above examples, the limiting mean size of clusters is 1 and 2 for the independent and simple dependence processes respectively.

1.4.2 Non-stationarity

As already discussed, asymptotic results of processes with specific forms of non-stationarity are usually *too* specific to be of use in data for which the form of non-stationarity is unknown. This section discusses how we can deal with two sources of non-stationarity – trend and seasonality – by modelling through the parameters of the standard models.

Trends

Trends can be accommodated by fitting the models described so far, but with time-dependent covariates. For example, to allow for a linear trend in the underlying level of extremal behaviour, we could use:

$$\mu(t) = \beta_0 + \beta_1(t), \quad (1.16)$$

where t represents time; or to allow for changing variability, we could use

$$\sigma(t) = \beta_0 + \beta_1(t). \quad (1.17)$$

The issue of trend is not considered in this thesis, as the datasets used are not long-term enough to exhibit any noticeable change in mean level.

Seasonality

The most widely adopted technique to deal with data which vary seasonally is to partition the data into seasons (within which we can assume the data to be homogeneous), and perform a separate extremal analysis on each season. Examples of such an approach can be found in Smith (1989) and Walshaw (1994). These seasons might be, for example, ‘winter’ and ‘summer’, or ‘dry’ and ‘wet’, where the seasonal variation is clearly understood. However, for data which exhibit less defined seasons, we can fit to separate months or years. Disadvantages of this approach are that a separate set of extremal parameters require estimating for each season, and that recombining these estimates is often non-trivial. To overcome these disadvantages, another approach is to allow the extremal parameters to vary continuously throughout the period of seasonality – for example,

within the year. Fourier forms can be fitted to the parameters, and a model selected based on likelihood ratio tests. However, in his Ph.D thesis, Walshaw (1991) suggests that inferences are barely altered in relation to a piecewise seasonality approach (for extreme wind gusts, anyway), and that the significant increase in computation time incurred by fitting continuously varying parameters is therefore not worthwhile.

1.5 Structure of this thesis

One of the primary objectives of this thesis is to formulate statistical models for extreme wind speeds which improve upon models currently used by practitioners in various disciplines. Civil engineers are particularly interested in being able to obtain accurate and precise estimates of long-period gust return levels; at present, “once-in-fifty-year gust speeds” are used to provide strength requirements for proposed buildings at a particular site, and these quantities are estimated using rudimentary extreme value analyses based on fitting the Gumbel distribution (type *I*) to sets of annual maxima. We aim to show how these return levels can be estimated more accurately and efficiently through more structured statistical modelling, with the potential for financial savings and, given the increased frequency and strength of storms in the U.K., improved safety.

Chapter 2 gives a brief introduction to the changing climate of the British Isles, with particular attention given to the extremal behaviour of the wind climate. A simple threshold model is then fitted to some real-life wind speed data from the U.K., by using the generalised Pareto distribution (expression 1.13) as a model for excesses over a high threshold. The issue of threshold selection is addressed, as well as the modelling issues of seasonal variation and short-range temporal dependence. The effects of both *ignoring* and *removing* the latter on parameter estimation are examined, with particular regard given to return level estimation. Simple diagnostic procedures to determine model goodness-of-fit are also introduced.

Chapter 3 investigates the inferential effects of *ignoring* and *removing* temporal dependence further, by fitting models to simulated data for which the true parameter values are *known*. Bivariate extreme value theory is drawn upon for simulating Markov chains within which successive pairs of realisations are modelled by a bivariate extreme value distribution. Again, particular consideration is given to the effects of both approaches on return level estimation, with return levels estimated for sequences displaying various strengths of extremal dependence.

Chapter 4 tackles the issue of temporal dependence from a different angle, and considers models which take into account the dependence aspects of a process. Markov chain models for threshold exceedances are introduced, and fitted to both simulated data and the real-life wind speed data discussed in chapter 2. Models which account for first- and second-order dependence are investigated, and attention given to different dependence structures. The suitability of a ‘Markov assumption’ for the real-life wind speed data is assessed through various formal and *ad hoc* diagnostic procedures. Again, we lend particular consideration to the estimation of return levels from such models.

Chapter 5 addresses the issue of making inferences about the extremal index. Various estimation procedures are reviewed, and the performance of these compared in a simulation study which examines sequences for which $\theta = 1$ and $\theta < 1$. Again, these are also implemented for the wind speed data, and attention given to the consequences of the different estimators for return level inference.

Chapter 6 introduces the Bayesian framework for extremal inference. We begin by outlining the general theory of Bayesian inference, and describe various Markov chain Monte Carlo algorithms for simulating from the posterior distribution. We then go on to discuss the particular advantages of adopting a Bayesian approach when analysing extreme values, and illustrate the use of Bayesian techniques for threshold excesses via simulated data and the wind speed data introduced in chapter 2. This Bayesian approach to inference is compared and contrasted to the classical maximum likelihood approach used in chapters 2 and 3 (for the real-life data and simulated data respectively), again with specific attention given to the effects on return level inference. We then focus attention on Bayesian inference for measures of extremal dependence, and investigate the advantages of a Bayesian approach to fitting Markov chain models to threshold excesses, as well as looking at ways of making inferences about the extremal index in the Bayesian setting. A Bayesian declustering scheme for temporally dependent extremes is then suggested, and applied to the real-life wind speed data. This allows us to study the posterior distribution of cluster functionals, of which we consider two: mean cluster size, and mean number of threshold upcrossings.

Finally, chapter 7 examines the possibility of pooling information from other wind speed stations across the U.K., in a fully Bayesian hierarchical model for threshold excesses. Site and seasonal effects are quantified, and temporal dependence accounted for through a Markov chain model set-up. This approach is compared to a site-by-site analysis modelled in the usual maximum likelihood framework. Imposing the Markov assumption as a

form of temporal dependence, we also extend the usual return level inference to examine predictive return levels.

Chapter 2

Extreme wind speeds

2.1 Introduction

Recently in the U.K., public awareness of the issues surrounding climate change has been heightened by the increased frequency of severe storms and flooding. The effects of ‘global warming’ on temperatures and sea-levels have been well-documented: average annual temperatures across the U.K. are expected to rise by between 2°C and 3.5°C by the mid 2080s, depending on future levels of greenhouse gas emissions; summer precipitation will decline and heavy winter precipitation will become more frequent, with an associated rise in sea-levels (a rise of between 26cm and 86cm above the current level in southeast England by 2080 is expected (DEFRA, 2003)). However, as the storms of October 1987 and October 2002 have demonstrated, increased winter precipitation is also often accompanied by severe wind gusts. In October 1987, wind speeds in excess of 110 miles per hour were recorded over parts of southern England, killing eighteen people and destroying buildings and bridges. Gust speeds and mean wind speeds were as great as those which can be expected to recur, on average, no more frequently than once every 200 years (see figure 2.1). Yet only fifteen years later, the U.K. was hit by another storm of similar ferocity, with severe wind damage to buildings in the eastern counties of England. More recently, in March 2004, winds as strong as 85 miles per hour swept the U.K., causing three deaths in East Anglia and damage to two bridges in the same area. Figure 2.3 shows the damage caused by these winds at a school in Killingworth, Northumberland, where part of the roof was ripped off by gusts exceeding 80 miles per hour; had the storm occurred on a school day, casualties would have been inevitable. Meteorologists predict that in the future such storms will occur with greater frequency, and strength.

Though other disciplines (such as oceanography and forestry) have a significant interest in the extremal behaviour of the wind climate, civil engineering provides the greatest

Figure 2.1: Left: Maximum gusts (miles per hour) during the 1987 storm. Right: Wind damaged building in Southampton after the 1987 storm

Figure 2.2: Uprooted trees caused severe structural damage in London and the Home Counties during the winds of October 2002

Figure 2.3: Structural damage to Bailey Green Primary School in Killingworth, Northumberland, after the gales of March 2004

impetus for studying extreme wind speeds. In structural planning, codes of practice exist regulating the design strengths of buildings according to estimates of the strongest gusts which are likely to occur at any site. Extreme value analyses of wind speeds are routinely used to estimate these gusts; given the expected increase in frequency of storms such as those of 1987 and 2002, the accuracy and reliability of such analyses is becoming more important, in the interests of both human safety and financial savings. Until very recently, extreme value analyses of wind speed data have been fairly rudimentary, almost always using simple model structures and procedures for estimation. In the U.K., contour maps are produced by the British Standards Institution, which display strength requirements based on ‘once-in-50-year gust speeds’, defined as the speed which is exceeded (on average) once every fifty years. This is more commonly referred to as the *50-year return level* gust. The maps themselves are the result of simple extreme value analyses carried out on medium to long term records (typically between 15 and 60 years), collected at many stations throughout the U.K. (see Shellard, 1976). Here, the Gumbel distribution has been used to model annual gust maxima.

The principal aim of this chapter is to demonstrate, through a real-life dataset, a typical threshold-based approach to modelling extreme wind speeds. Practical solutions to the various problems outlined in the previous chapter are demonstrated, with particular attention given to seasonality and short-term temporal dependence. The effects of removing temporal dependence and modelling a set of independent cluster maxima (Davison and Smith, 1990) are investigated. The consequences of such a scheme on parameter estimation – and in particular return level estimation – are explored, and compared with an approach which ignores temporal structure and simply fits the GPD to all threshold excesses.

2.2 Data description

2.2.1 Wind: the variable

The behaviour of wind velocity as a continuous random variable demonstrates many characteristics which distinguish it from other environmental variables. In common with other such variables, extreme wind speeds often occur seasonally and display short-term temporal dependence. However, unlike sea-level, for example, wind velocity cannot be broken down into distinct components (temporal variability in sea-level processes is essentially due to two different physical mechanisms: astronomically induced tides, and meteorological storm surges. Several authors (e.g. Tawn and Vassie, 1989) advocate separate analyses for these constituent parts); nor does it arrive in clearly identifiable episodes,

unlike rainfall. In comparison to other environmental phenomena, wind velocity is also relatively well-behaved. It is not subject to *very* violent departures from the norm – though a wind velocity of 200 miles per hour may sound quite severe, such departures from the mean level are quite small relative to those demonstrated by rainfall, flow-rates and pollution concentrations. Wind velocity is also one of the most rapidly varying of all environmental variables. Many distinct gusts are often observed in periods as short as a few minutes, and so we are provided with a large amount of information in a relatively short time period. This strengthens the arguments supporting the use of limiting asymptotic distributions for the most extreme gusts, and enables us to make inferences about long-period return levels from comparatively few measurements.

2.2.2 Site details and data collection

The data considered in this chapter have been collected at High Bradfield, in the Peak District national park, for the U.K. Meteorological Office. The measuring station is situated in an open area, about 380 metres above sea-level. The basic instrumental method for measuring wind in the U.K. is the Munro cup anemometer and vane. Traditionally, this was connected to an anemograph and dial (figure 2.4 shows a section of the anemograph trace for Bradfield), but with the advent of automatic systems it is now usually attached to a processing device, such as the DALE logger. The Munro cup has a large inertia and so a relatively slow response time; however, since 1998, the Munros have been replaced by a newer, lightweight anemometer with better response characteristics. For comparison with wind speed records at other sites (U.K. and worldwide), individual anemometers are designated as having an ‘effective height’ above ground level, which is the “height over open terrain in the vicinity of the instrument” (Shellard, 1976). Meteorologists then reduce hourly mean speeds and hourly gust maxima to their ‘standard’ effective height of 10 metres, using the power law formulae:

$$\begin{aligned} V_{10} &= V_z(10/z)^{0.170} && \text{for mean speeds} \\ V_{10} &= V_z(10/z)^{0.085} && \text{for gust maxima,} \end{aligned}$$

where V_{10} is the speed at 10 metres and V_z is the speed at z metres (see Shellard, 1976). At some sites, the exponents need adjusting to account for variations in surface roughness.

Available records for High Bradfield span a period of ten complete years, from January 1st 1975 to December 31st 1984. For each hour, a set of summary statistics describing the wind is recorded. Up to and including 2nd November 1981, these consist of the maximum gust speed in knots (recorded to the nearest integer) and the hourly mean speed. From 3rd November 1981, the direction of the maximum gusts are also noted. In this

Figure 2.4: Section of Bradfield anemograph trace

Figure 2.5: Time series plot of hourly gust maxima at Bradfield (1975–1978), and histogram of entire dataset

Mean	St. dev.	Median	LQ	UQ	Min.	Max.
22.28	12.24	20.0	13.0	29.0	0.0	99.0

Table 2.1: Summary statistics of Bradfield hourly gust maxima (knots)

chapter, we look at the set of hourly gust maxima, constituting some 87,672 observations. Before automatic processing, the direction vane and rotating cups each generated an anemograph trace, from which the digital measurements were extracted by eye. For hourly gust maxima, it is easy to read from the trace plot the precise magnitude (which is then recorded to the nearest knot). Theoretically, this is the greatest instantaneous wind velocity attained in the given hour; in practice, because of inertia in the anemometer, the maximum recorded is actually the peak value of a moving average taken over two or three seconds (BADC, 2003). A general discussion of the wind climate of high altitude sites in the southern Pennines is given by Smithson (1987); more general analyses and modelling of the wind climate of the British Isles can be found in S.G. Smith (1984) and Shellard (1976). The unit of speed used at U.K. stations is the knot (0.515ms^{-1}), which is the unit of measurement used most often in this thesis. Where other measurements have been used, 1 knot is also equal to 1.152 miles per hour (mph) and 1.854 kilometres per hour (kmph). Figure 2.5 shows a time series plot and histogram of the first three years of data from Bradfield; table 2.1 gives some summaries.

2.2.3 Modelling approach

We consider the problem of estimating return levels for the gusts at Bradfield. The data available have been collected over a period of ten years – relatively short in comparison to the return periods of interest; in such cases, a classical ‘annual maxima’ approach is unworkable due to the sparsity of data. Other methods which make use of several order statistics from each year (the ‘ r -largest’ approach, for example (Tawn, 1988b)) can give viable estimates of 50-year return level gusts from relatively short runs of data; however, such methods must take into account serial correlation, and are vulnerable to the effects of seasonal variability. Although both these issues also affect threshold methods, their main advantage over the use of order statistics is their flexibility in the choice of extreme events for analysis; the number of events (such as storms) which occur over a fixed period are allowed to vary according to the strength of the wind during that time, whereas the r -largest order statistics method uses observations from fixed time intervals. To this end, the generalised Pareto distribution will be used to model sets of threshold excesses. The following sections will detail how seasonal variability and short-term correlation are dealt with, as well as demonstrating a procedure for the selection of a suitably high threshold.

2.3 Seasonal variation

It is usual in strongly seasonal climates for the occurrence of extreme winds to be confined to a certain part of the yearly cycle. In the U.K., for example, it is very unusual for wind damage to occur outside the period October–March. Such seasonal variation is evident in the time series plot for Bradfield (figure 2.5), which shows distinct cycles in the data. From this plot, we can see that winter months generally display stronger gusts than summer months; this is a phenomenon one might expect, since prolonged, anticyclonic periods are more prevalent during June, July and August than in winter months. A model which takes account of seasonal variability will identify all gusts which are large *given the time of year* as extreme. There is only a point to modelling the extremes which occur in summer months, for example, if we believe that they can help us understand what happens in winter months, where genuinely large events *can* occur. For this to be realistic, we must assume that the same mechanism is responsible for the generation of large gusts throughout the year, and it is just the scale of this mechanism which changes. Indeed, in temperate climates (such as that of the U.K.), the same alternating sequence of anticyclones and depressions leads to most of the storms which occur throughout the year (Walshaw, 1994); the seasonal variability comes from the severity of these systems.

For wind speed data, there is no natural partition of the year into separate seasons. Here, we take our seasonal unit to be one month; Walshaw (1994) argues that by dividing the year into twelve equal length seasons, we can strike 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. Davison and Smith (1990) discuss two approaches for dealing with seasonally varying data. The first involves the removal of seasonal components to generate a *prewhitened* series; any subsequent analysis is then performed on this residual series. The second involves a separate GPD being fitted to each set of seasonal excesses, assuming stationarity within each season. Here, we adopt the latter approach, though the method of prewhitening is also used in the initial stages to aid the threshold selection procedure.

2.3.1 Generation of a stationary series

With monthly varying data, an appropriate set of prewhitening operations can be obtained by transformations t_m for each month m . If we allow both the shape and scale parameters of the generalised Pareto distribution to vary between months, it can be shown that a set of transformations $t_m(x) = a_m \log(x - c_m) + b_m$ can be chosen to force the distribution

of excesses over a *single* threshold to be homogeneous GPD over all months (assuming that the original excesses are GPD – see Walshaw (1994)). If only the scale is allowed to vary between months, it is sufficient to set $c_m = 0$ and use the set of transformations $t_m(x) = a_m \log(x) + b_m$. A good approximation to the appropriate transformations is obtained by making the upper tails of the empirical monthly distributions of gust maxima coincide with each other. For example, for the case which allows both the scale and shape parameters to vary, estimates are made by transforming three high quantiles from each month to three distinct arbitrary points. The precise choice of quantiles is not critical; practically, we need to move as far as possible into the upper tails of the distributions, while still retaining a substantial amount of information between them (and above the largest). We can transform the empirical 90, 95 and 99% quantiles ($q_{1,m}$, $q_{2,m}$ and $q_{3,m}$ respectively) to the corresponding quantiles of the unit exponential distribution (Q_1 , Q_2 and Q_3), by solving

$$a_m \log(q_{1,m} - c_m) + b_m = Q_1, \quad (2.1)$$

$$a_m \log(q_{2,m} - c_m) + b_m = Q_2 \quad \text{and} \quad (2.2)$$

$$a_m \log(q_{3,m} - c_m) + b_m = Q_3, \quad (2.3)$$

where $a_m > 0$, b_m and c_m , and $m = 1, \dots, 12$.

2.3.2 Implementation for the Bradfield data

For each month in the Bradfield data, 10 years of hourly gust maxima give approximately 7300 observations. This gives about 730 points lying above the 90% quantile, 365 above the 95% quantile and 73 points above the 99% quantile. Therefore, when estimating these quantiles' theoretical values via their empirical equivalents, the sampling error is fairly small. So for each month m , $m = 1, \dots, 12$, equations (2.1), (2.2) and (2.3) are solved simultaneously for a_m , b_m and c_m , and the transformation $a_m \log(x - c_m) + b_m$ is applied to all the data within that month. The resulting residual series lies in the range $[-5.511, 6.462]$, and is approximately stationary in the upper tails. Figure 2.6 shows a plot of the transformed series. If the first approach suggested by Davison and Smith is adopted, then an extreme value analysis of this transformed series is carried out, and the seasonal components replaced prior to the calculation of return levels (by applying the inverse of the prewhitening operations). In the separate months model, this prewhitened series is used to select a single threshold, which is then translated into a monthly varying threshold for the raw data by applying the inverse of the prewhitening operations. An extreme value analysis is then carried out on each month separately, *using the true observations*. The prewhitened series is also useful to assess the fit to all seasons simultaneously, instead of producing quantile–quantile plots, for example, for each season individually.

Figure 2.6: Time series plot of prewhitened Bradfield data

2.4 Temporal dependence

In most environmental applications, short-term dependence at extreme levels is common. Wind velocity is no exception, with a tendency for large gusts to occur in clusters as storms. To account for this, some modification is needed to threshold-based techniques; if maximum likelihood is used, for example, to estimate the parameters of the GPD, then assuming independence will result in an under-estimation of the associated standard errors as there is less information in the data than the likelihood requires. Various solutions to this problem have been suggested; the two we consider here are:

1. *Ignore* any temporal dependence in the series and proceed with the analysis. Then inflate the standard errors of the estimates accordingly;
2. *Remove* temporal dependence from the series prior to analysis by filtering out a set of independent threshold excesses. This can be done by formally identifying clusters of extremes which are deemed to be correlated, and discarding all but the largest observation (the “cluster peak”) within each cluster.

2.4.1 Ignoring dependence: using all threshold excesses

This approach fits the GPD to the set of all threshold excesses under the assumption of independence. As already discussed, the likelihood asymptotics are somewhat less rigorous under such an approach – the main effect being the under-estimation of standard errors. However, this method does have some appeal; the standard likelihood is easily applied to all the extreme data without the need to decluster the series (see 2.4.2), and there is an increase in precision since all excesses above the threshold are incorporated into the analysis. Smith (1990) has also suggested a procedure in which the usual asymptotic like-

likelihood calculations are supplemented by empirical information on dependence, to produce a modified covariance matrix for the parameters, which is approximately correct after the dependence has been taken into account. Under the model fitting procedure which assumes independence, we denote by H the observed information matrix. If independence were a valid assumption, then the covariance matrix of the maximum likelihood estimates would be approximately H^{-1} . However, to account for dependence, Smith (1990) shows that this approximation should be replaced by $H^{-1}VH^{-1}$, where V is the covariance matrix of the likelihood gradient vector. The effect of this modification on the standard errors attached to the maximum likelihood estimates of the parameters is to inflate their values, in comparison to those obtained under the assumption of independence.

2.4.2 Removing dependence: using cluster peak excesses

This approach (suggested by Davison and Smith, 1990) is probably one of the most commonly used methods to deal with temporal dependence of environmental time series. The idea is to formally identify ‘clusters’ of extremes, and then use only the maximum from each cluster in the analysis. A simple way of determining such clusters is to specify a threshold u , and define consecutive exceedances of u to belong to the same cluster. Once we obtain an observation that falls below u , the cluster is deemed to have terminated. The next exceedance of u then initiates the next cluster. The problem is that under this scheme, separate clusters are allowed to be separated by a single observation, in which case the argument for independence across cluster maxima is flimsy. Furthermore, the separation of extreme events into clusters could be sensitive to threshold choice.

To overcome these deficiencies, it is common to introduce a ‘declustering parameter’, κ , which in practice is a separation interval between neighbouring clusters. Now, a cluster is deemed active until at least κ consecutive observations fall below the threshold. The choice of κ requires care: if κ is too small, the assumption of independence for nearby clusters will be unrealistic; however, if κ is chosen too large, nearby clusters that could reasonably have been considered independent will be merged, leading to a loss of valuable data once the maxima are extracted. Hopefully, however, inferences will be reasonably robust to the precise choice of κ . Section 2.5 describes a procedure for the joint selection of threshold u and declustering parameter κ , though such graphical procedures also rely on common-sense judgement. The particular declustering scheme described here is often referred to as ‘runs’ declustering, though ‘blocks’ declustering (see section 5.2.1) may also be used. Runs declustering is preferred for the wind speed data, however, since it allows both the duration of storms – and the duration of intervals between them – to vary according to the data, reflecting the inherent natural variability of these quantities.

2.5 Threshold selection

Now that a stationary series of hourly gust maxima has been created in the upper tail (figure 2.6), we are in a position to experiment with various choices of threshold. A constant threshold for the prewhitened series is proposed, since we can now assume the region of the data to be treated as extreme will constitute the same upper quantile for all seasons. Applying the inverses of the prewhitening transformations to this threshold will then provide us with a set of monthly varying thresholds to use in the final model.

Since exceedances of these thresholds will occur in clusters, and both the threshold and declustering parameter interact in the manner in which they determine the set of cluster peak exceedances, a method which determines these quantities in combination has obvious appeal. The traditional approach for threshold selection involves the examination of a *mean residual life plot*, which plots the mean number of excesses of u against u (for various u); a suitable threshold is then chosen at the minimum level such that the plot appears approximately linear. In theory, it would be possible (though rather cumbersome) to examine a large number of such plots with a variety of declustering parameters. Instead, we adopt the route of Walshaw (1994), which advocates the use of *reclustered excess plots*. Here, given a declustering parameter κ , the mean excess above threshold u is plotted against u , with the identification of cluster peak exceedances being carried out separately for each threshold u . Linearity in such a plot above a certain threshold $\tilde{u}$ suggests a good fit of the GPD to cluster peak excesses above $\tilde{u}$ selected using declustering parameter κ , as well as a robustness of the mean cluster peak excess to the threshold at which declustering is carried out. Producing a range of reclustered excess plots for different κ , we can identify the smallest declustering parameter which allows us to safely assume independence of cluster peak excesses. This will be the smallest value giving a plot which straightens out above a certain threshold $\tilde{u}$. Once the optimal pair $(\tilde{u}, \kappa)$ have been chosen, a conventional mean residual life plot can then be constructed to verify the validity of the choice.

2.5.1 Implementation for the Bradfield data

Figure 2.7 shows reclustered excess plots produced for declustering parameters $\kappa = 0$ (all excesses), 6 hours, 15 hours, 30 hours, 60 hours and 100 hours. Linearity in the plots is achieved (arguably) at a separation interval of 30 hours with a threshold of about 2.69. Notice the larger declustering parameters also produce reasonably linear plots; however, we would prefer to have as small an interval as possible to allow the inclusion of a greater number of cluster peaks. With this in mind, we select the pair $(\tilde{u} = 2.69, \kappa = 30)$.

The threshold $\tilde{u} = 2.69$ lies at the 0.923 quantile in the empirical distribution of the transformed hourly maxima. If all threshold excesses are to be incorporated into the analysis (i.e. we choose to ignore dependence), a threshold of 2.2 is chosen (from the first plot with $\kappa = 0$). Figure 2.8 shows conventional mean residual life plots for (a) using all excesses, and (b) using cluster peak excesses with $\kappa = 30$ hours, plotted in the range $u > \tilde{u}$. Both show a reasonably linear plot above $\tilde{u}$, suggesting that the choice of threshold is adequate. The far right-hand-side of these plots have been omitted; here, variability is high due to the limited amount of data above such high thresholds, and so information in the plots is unreliable.

Figure 2.7: Reclustered excess plots for the transformed series

Figure 2.8: Mean residual life plots

2.6 Model fitting and verification

2.6.1 Maximum likelihood

Using the thresholds and declustering parameter obtained from the reclustered excess plots, we can move directly to a separate seasons model for the raw (untransformed) threshold excesses. Under the appropriate model, the excesses in month m over a segmented monthly varying threshold (found by applying the inverses of the prewhitening transformations to a threshold of $\tilde{u} = 2.69$ for cluster peak excesses and 2.2 for all excesses) are independent GPD with associated distribution functions

$$G_m(y; \sigma_m, \xi_m) = 1 - \left(1 + \frac{\xi_m y}{\sigma_m}\right)_+^{-1/\xi_m}, \quad (2.4)$$

where σ_m and ξ_m are scale and shape parameters (respectively) for month m . We use maximum likelihood to estimate these parameters, applying the inverses of the prewhitening transformations to parameters found for the transformed series to obtain preliminary estimates for σ_m and ξ_m . The log-likelihoods for each month (with n_m excesses) are derived from (2.4) as

$$\ell(\sigma_m, \xi_m) = -n_m \log \sigma_m - (1 + 1/\xi_m) \sum_{j=1}^{n_m} \log(1 + \xi_m y_j / \sigma_m), \quad (2.5)$$

provided $(1 + \sigma_m^{-1} \xi_m y_j) > 0$ for $i = 1, \dots, n_m$; otherwise, $\ell(\sigma_m, \xi_m) = -\infty$. In the case $\xi = 0$, the log-likelihood is

$$\ell(\sigma_m) = -n_m \log \sigma_m - \sigma_m^{-1} \sum_{j=1}^{n_m} y_j. \quad (2.6)$$

Analytical maximisation of the log-likelihood is not possible, so numerical techniques are employed, taking care to avoid numerical instabilities when $\xi \approx 0$ (in 2.5), and ensuring that the algorithm does not fail due to evaluation outside of the allowable parameter space. Standard errors are obtained in the usual way from standard likelihood theory.

2.6.2 Graphical evaluation

The fit of the separate seasons model can be assessed for all months simultaneously by using the fitted parameter values to transform each month of threshold excesses to a common margin (unit exponential, for example). The overall fit of the model can then be verified via probability plots or quantile-quantile plots (plots of fitted distribution function versus empirical distribution function, or fitted quantile versus empirical quantile; the plotting points being defined by all threshold exceedances or cluster peak exceedances, depending on the approach taken).

2.6.3 Results for the Bradfield data

Tables 2.2 and 2.3 show the results of the separate seasons model fitted to all threshold excesses and cluster peak excesses (respectively). The tables contain the monthly varying thresholds u_m , the number of excesses included in the analysis (n_m), and maximum likelihood estimates for σ_m and ξ_m , with associated standard errors in parentheses. 9

Month (m)	u_m	n_m	$\hat{\sigma}_m$	$\hat{\xi}_m$
1	59.777	326	8.149 (0.633)	-0.102 (0.055)
2	45.650	268	8.145 (0.673)	-0.104 (0.056)
3	51.269	256	9.009 (0.691)	-0.359 (0.050)
4	43.396	276	8.413 (0.655)	-0.271 (0.052)
5	35.395	271	6.870 (0.593)	-0.120 (0.062)
6	37.888	288	3.959 (0.358)	0.058 (0.069)
7	35.142	247	7.519 (0.571)	-0.414 (0.049)
8	35.147	263	6.972 (0.613)	-0.100 (0.063)
9	35.128	939	9.898 (0.400)	-0.137 (0.024)
10	48.396	290	7.758 (0.642)	-0.231 (0.060)
11	52.396	268	7.385 (0.583)	-0.061 (0.050)
12	53.781	292	8.411 (0.750)	-0.121 (0.068)

Table 2.2: Results when the separate months model is fitted to all threshold excesses

Month (m)	u_m	n_m	$\hat{\sigma}_m$	$\hat{\xi}_m$
1	55.341	28	21.373 (5.358)	-0.420 (0.183)
2	41.531	24	15.130 (4.635)	-0.226 (0.233)
3	48.100	29	23.277 (6.316)	-0.894 (0.259)
4	39.910	29	14.853 (4.448)	-0.440 (0.249)
5	31.943	46	9.456 (1.990)	-0.158 (0.147)
6	35.670	35	12.329 (2.592)	-0.409 (0.143)
7	32.290	36	12.517 (2.609)	-0.605 (0.161)
8	32.639	34	10.199 (2.361)	-0.203 (0.159)
9	33.232	49	18.772 (3.668)	-0.255 (0.138)
10	44.914	34	11.669 (3.533)	-0.274 (0.254)
11	48.394	33	14.991 (3.381)	-0.225 (0.149)
12	49.341	35	18.681 (4.229)	-0.416 (0.166)

Table 2.3: Results when the separate months model is fitted to cluster peak excesses

Figure 2.9 shows probability plots and quantile–quantile plots to assess the fit of the model using both cluster peak excesses (removing dependence) and all threshold excesses (ignoring dependence).

Figure 2.9: Probability plots and quantile–quantile plots

Comparing tables 2.2 and 2.3, we can see marked differences in the monthly MLEs for both parameters when using all threshold excesses and just cluster peak excesses. Estimated values of σ are significantly higher, and estimates of ξ considerably lower, when fitting to cluster peak excesses only. The standard errors shown are those obtained in the usual way from standard likelihood theory – i.e. the (positive) square root of the elements in the leading diagonal of the inverse of Fisher’s information matrix. As has already been discussed, this method is incorrect for the model which fits to all threshold excesses since there is strong serial correlation present in the extremes. Thus, these standard errors under–estimate the true standard errors and the method suggested by Smith (1990), outlined in section 2.4.1, can be used to inflate these accordingly. Table 2.4 compares the standard errors for σ and ξ shown in table 2.2 with those obtained from Smith’s method which accounts for serial dependence. Even when these standard errors have been adjusted to account for dependence, we can see that the approach which fits to

cluster peak excesses only still has larger standard errors attached to the MLEs. This is a consequence of including fewer observations in the cluster peaks approach, reducing the precision of our analysis. The total number of threshold exceedances in the first approach is 3984, approximately 5% of the total number of observations available; this compares to just 412 in the cluster peaks approach – less than 0.5% of the total. Fitting to each month separately, this gives an average of about 34 observations available for analysis in each month, with February having only 24 cluster peak excesses. The plots in figure 2.9 support the overall adequacy of both models, with the model which uses all threshold excesses appearing to give the best fit. Note the (comparatively) large number of excesses for September in table 2.2 (n_9). As we shall see in chapter 4 (table 4.1) of this thesis, successive extreme wind speeds in September are more strongly correlated than in any other month; thus, there tend to be longer runs of extremes in this month, and so (possibly) a larger number of exceedances over a suitable threshold.

month (m)	$\hat{\sigma}$		$\hat{\xi}$	
	s.e.	s.e. _{adj}	s.e.	s.e. _{adj}
1	0.633	2.482	0.055	0.089
2	0.673	2.542	0.056	0.097
3	0.691	2.553	0.050	0.102
4	0.655	2.519	0.052	0.101
5	0.593	1.437	0.062	0.082
6	0.358	1.553	0.069	0.097
7	0.571	1.874	0.049	0.104
8	0.613	1.650	0.063	0.109
9	0.400	1.501	0.024	0.081
10	0.642	2.387	0.060	0.113
11	0.583	2.390	0.050	0.096
12	0.750	2.491	0.068	0.099

Table 2.4: Standard errors found using the ‘traditional’ method (assuming independence) compared to those adjusted for dependence (using Smith’s method)

2.7 Return level estimation

2.7.1 Methods of estimation

It is usually more convenient (and practically important) to interpret extreme value models in terms of quantiles or return levels, instead of individual parameter values. Suppose

that a GPD with parameters σ and ξ is a suitable model for exceedances of a threshold u by a variable X , i.e. for $x > u$,

$$\Pr(X > x | X > u) = \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]_+^{-1/\xi}. \quad (2.7)$$

Then

$$\Pr(X > x) = \lambda_u \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]_+^{-1/\xi}, \quad (2.8)$$

where $\lambda_u = \Pr(X > u)$. Hence, the level x_t that is exceeded on average once every t observations is the solution of

$$\lambda_u \left[1 + \xi \left(\frac{x_t - u}{\sigma} \right) \right]_+^{-1/\xi} = \frac{1}{t}. \quad (2.9)$$

Rearranging, we get

$$x_t = u + \frac{\sigma}{\xi} [(t\lambda_u)^\xi - 1], \quad (2.10)$$

provided t is sufficiently large to ensure that $x_t > u$, and $\xi \neq 0$. If $\xi = 0$, we have the exponential case, and so

$$x_t = u + \log(t\lambda_u), \quad (2.11)$$

again provided t is sufficiently large. By construction, x_t is the t -observation return level; however, it is often more convenient to give return levels on an annual scale, so that the r -year return level is the level expected to be exceeded once every r years. If there are n_y observations per year, this corresponds to the t -observation return level with $t = r \times n_y$. Hence, the r -year return level q_r is defined by

$$q_r = u + \frac{\sigma}{\xi} [(rn_y\lambda_u)^\xi - 1], \quad (2.12)$$

unless $\xi = 0$, in which case

$$q_r = u + \sigma \log(rn_y\lambda_u). \quad (2.13)$$

Since we have seasonally varying data, and a separate GPD being fitted to each month, the r -year return level q_r is now obtained as the solution of

$$\sum_{i=1}^{12} \lambda_{u_m} \left[1 + \xi_m \left(\frac{q_r - u_m}{\sigma_m} \right) \right]_+^{-1/\xi_m} = \frac{1}{r}, \quad (2.14)$$

which cannot be found analytically. A point estimate $\hat{q}_r$ is obtained by substituting the MLEs $\hat{\lambda}_{u_m}$, $\hat{\sigma}_m$ and $\hat{\xi}_m$ and thresholds u_m into (2.14) and solving numerically for q_r . The

MLEs for λ_{u_m} are simply the mean number of exceedances per year in month m , denoted $\bar{x}_m$, and have associated standard errors estimated by $\sqrt{\bar{x}_m/10}$. Equation (2.14) can be solved using a program containing the NAG routine C05AJF (Numerical Algorithms Group, 1999), which attempts to locate a zero of a continuous function using numerical search methods.

2.7.2 Standard errors

Standard errors for q_r can be derived by the delta method (Rao, 1973, p388), with the uncertainty in estimates of λ_{u_m} being included in the calculation. Since the number of exceedances of u_m follows a binomial $\text{Bin}(n_m, \lambda_{u_m})$ distribution, $\text{Var}(\hat{\lambda}_{u_m}) \approx \hat{\lambda}_{u_m}(1 - \hat{\lambda}_{u_m})/n_m$, so the complete variance–covariance matrix for $(\hat{\lambda}_{u_m}, \hat{\sigma}_m, \hat{\xi}_m)$ is approximately

$$V = \begin{pmatrix} \frac{\hat{\lambda}_{u_1}(1-\hat{\lambda}_{u_1})}{n_1} & 0 & \dots & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \ddots & \frac{\hat{\lambda}_{u_{12}}(1-\hat{\lambda}_{u_{12}})}{n_{12}} & 0 & \dots & 0 \\ 0 & \dots & 0 & v_{1,1} & \dots & v_{1,24} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & v_{24,1} & \dots & v_{24,24} \end{pmatrix}, \quad (2.15)$$

where $v_{i,j}$ denotes the (i,j) –th term of the variance–covariance matrix of $\hat{\sigma}_m$ and $\hat{\xi}_m$, $m = 1, \dots, 12$. Hence, by the delta method,

$$\text{Var}(\hat{q}_r) \approx \nabla q_r^T V \nabla q_r, \quad (2.16)$$

where

$$\nabla q_r^T = \left[\frac{\partial q_r}{\partial \lambda_{u_1}}, \dots, \frac{\partial q_r}{\partial \lambda_{u_{12}}}, \frac{\partial q_r}{\partial \sigma_1}, \dots, \frac{\partial q_r}{\partial \sigma_{12}}, \frac{\partial q_r}{\partial \xi_1}, \dots, \frac{\partial q_r}{\partial \xi_{12}} \right], \quad (2.17)$$

evaluated at $(\hat{\lambda}_{u_1}, \hat{\sigma}_1, \hat{\xi}_1, \dots, \hat{\lambda}_{u_{12}}, \hat{\sigma}_{12}, \hat{\xi}_{12})$.

2.7.3 Implementation for the Bradfield data

	Return period (years)			
	10	50	200	1000
All exceedances	96.785 (3.534)	103.180 (6.045)	108.085 (10.702)	113.249 (15.678)
Cluster peaks	94.916 (3.970)	100.469 (15.971)	104.950 (23.793)	114.954 (44.865)

Table 2.5: Return level estimates with associated standard errors

Table 2.5 gives point estimates and standard errors (in parentheses) for the 10, 50, 200 and 1000-year return levels at Bradfield, using all excesses and cluster peak excesses. As can be seen, there is a discrepancy between results based on using all threshold excesses and just cluster peak excesses, the latter approach giving smaller estimates for the 10, 50 and 200-year return periods (results for the 1000-year return period are more similar); these differences, however, are insignificant when the standard errors are taken into account. As with estimates of the GPD scale and shape parameters, without knowing the *true* r -year return level for the Bradfield wind speed data, we are not in a position to comment on which approach (*ignoring* dependence or *removing* dependence) gives the more accurate estimates. As we might expect, estimates based on using just cluster peak excesses also have larger standard errors attached to them as a result of far fewer data points being used in the analysis. In fact, the standard errors quoted in table 2.5, for *both* approaches, are so large (particularly for the longer return periods) that the Wald 95% confidence interval (parameter estimate $\pm 1.96 \times$ estimated standard error) is unlikely to produce good results. Therefore, we have also constructed interval estimates based on the profile-likelihood of the return level (see table 2.6). Rather than use the limiting quadratic form of the likelihood surface, profile-likelihood makes use of its actual shape for the data in question. The severe asymmetry of the surface often encountered when it is calculated for return levels also suggests that conventional symmetrical confidence intervals can be misleading.

To obtain intervals based on the profile-likelihood, we maximise the log-likelihood with respect to the model parameters subject to the constraint equation (2.14), for each of a range of possible values of q_r . This can be achieved by making one of the parameters the subject of equation (2.14). If, for example, σ_1 is chosen, then we have

$$\sigma_1 = \xi_1(q_r - u_1) / \left\{ [(r^{-1} - C)/\lambda_{u_1}]^{-\xi_1} - 1 \right\}, \quad (2.18)$$

where

$$C = \sum_{m=2}^{12} \lambda_{u_m} \left[1 + \xi_m \left(\frac{q_r - u_m}{\sigma_m} \right) \right]_+^{-1/\xi_m}. \quad (2.19)$$

We fix the return level q_r at the desired level and then maximise the log-likelihood $\ell \equiv \ell(q_r)$ with respect to all parameters except σ_1 . We calculate σ_1 numerically at each iteration in the maximisation (from equation (2.18)), and ℓ is obtained in the following way: suppose the excesses being modelled occur over a period of l years, and the number of excesses in month m and year j is $n_{m,j}$. Let $n_m = \sum_{j=1}^l n_{m,j}$, and denote the excesses

$y_{m,i}$, $i = 1, \dots, n_m$. Then

$$\begin{aligned} \ell(\sigma_m, \xi_m, \lambda_{u_m}) = & \sum_{m=1}^{12} \left[-n_m \log \sigma_m - \left(1 + \frac{1}{\xi_m} \right) \sum_{i=1}^{n_m} \log \left(1 + \frac{\xi_m y_{m,i}}{\sigma_m} \right) \right] \\ & - l \sum_{m=1}^{12} \lambda_{u_m} + \sum_{m=1}^{12} n_m \log \lambda_{u_m} - \sum_{m=1}^{12} \sum_{j=1}^l \log(n_{m,j}!). \end{aligned} \quad (2.20)$$

We can then form a confidence interval for q_r by inversion of a likelihood ratio test, i.e. the set of values q_0 for which $2[L(\hat{q}_r) - L(\hat{q}_0)]$ is not significant when compared with a χ^2 distribution on one degree of freedom, where $\hat{q}_r$ is the MLE for q_r . Figure 2.10 shows the profile likelihood for q_{50} (using *all* threshold excesses), illustrating asymmetry in the surface. The solid vertical line is plotted through the maximum likelihood estimate $\hat{q}_{50} = 103.18$ knots. The horizontal line lies at a level $0.5 \times \chi_1^2(0.95)$ below the maximised log-likelihood; the intersections with the surface are given by the dashed vertical lines, and provide the bounds for the 95% confidence interval for $\hat{q}_{50}$. Table 2.6 gives the 95% profile-likelihood confidence intervals for return periods 10, 50, 200 and 1000 years, as well as the 95% Wald confidence intervals using the standard errors from table 2.5 (for comparison).

Just as the standard errors for maximum likelihood estimates were under-estimated when fitting the GPD to all excesses (and therefore ignoring serial dependence – see section 2.6.3), the profile-likelihood intervals for all exceedances in table 2.6 must surely be too narrow. Though these intervals take into account the (possible) asymmetry of the likelihood surface for return levels, the failure to filter out (or successfully handle) dependence between successive threshold excesses will result in confidence intervals which are tighter than they should be. Though not attempted in this thesis, it may be possible to formulate a profile-likelihood procedure for using *all* threshold excesses which takes the inevitable temporal dependence into account, via a method similar to that proposed by Smith (1990) used in 2.6.3 (results from which are given in table 2.4).

2.8 Discussion

This chapter has demonstrated how threshold methods can be used to model extreme wind speeds. The issues of temporal dependence and seasonal variation have been addressed from a purely statistical viewpoint, modelling changes in the marginal behaviour of extremes rather than appealing to any additional asymptotic theory. Indeed, because of the generality with which such effects of non-stationarity may arise in practice, there is little point in the theoretical study of such extremes.

Figure 2.10: Profile log-likelihood for the fifty year return level q_{50}

	Return period (years)			
	10	50	200	1000
All exceedances	(93.70, 100.35)	(99.77, 107.91)	(103.88, 113.10)	(108.52, 118.79)
	(89.86, 103.71)	(91.33, 115.03)	(87.11, 129.06)	(82.52, 143.98)
Cluster peaks	(90.02, 100.15)	(77.82, 126.40)	(75.43, 138.55)	(74.75, 167.53)
	(87.13, 102.70)	(69.17, 131.77)	(58.32, 151.58)	(27.02, 202.89)

Table 2.6: 95% profile-likelihood confidence intervals (Wald intervals given underneath for comparison)

The issue of seasonal variability was easily overcome by fitting a GPD to each month of excesses separately. However, the problem of temporal dependence has been investigated a little more thoroughly. The inferential consequences of both *removing* dependence and *ignoring* dependence have been compared, and the results of fitting a GPD to all threshold excesses and to cluster peak excesses only were substantially different. Not only were estimates of the GPD scale and shape parameters significantly different under the two approaches, but discrepancies were also carried through to return level inference. For the 10, 50 and 200-year return levels, the method which used only cluster peak excesses appeared to under-estimate relative to the method using all excesses. Results for the 1000-year return period were more consistent. Most practitioners tend to adopt the method of filtering out a set of independent cluster maxima to avoid the problem of serial correlation evident in many environmental time series; however, if this approach is systematically under-estimating return levels, design strengths of buildings, for example, might not be adequate to withstand the strongest gusts which are likely to occur. Of course, the problem might lie with the approach which uses all excesses. It could be that the method which uses the independent cluster peak excesses estimates return levels accurately, whilst using all threshold excesses results in over-estimated return levels. Further, the (possible) discrepancies found under the two approaches could be due to sampling error. Indeed, though large, the standard errors suggest the two methods yield return level estimates which are not *significantly* different. Though there's nothing (necessarily) interesting about that, this does advocate the use of all excesses over cluster peak excesses, simply because of the inclusion of more data and so increased precision. Without knowing the true return levels, we can't be sure which approach provides the more accurate return level estimates.

From this preliminary study, the arguments in favour of using all excesses in any threshold-based extreme value analysis seem compelling; though ignoring the serial correlation in extreme gusts had the effect of under-estimating the standard errors of the GPD scale and shape parameters, the method proposed by Smith (1990) provides a solution by inflating these accordingly, and the gain in efficiency by incorporating all extreme values into the analysis is appealing. The decision over which method should be used to decluster the series, as well the (arguably subjective) choice of declustering parameter, is no longer an issue if we decide to use all threshold excesses, and a simple mean residual life plot can be used to aid the selection of an appropriate threshold without having to worry about the interaction between threshold level and declustering parameter. Further, the probability plots and quantile-quantile plots for the approach which used all excesses seemed to indicate a slightly better fit anyway, the points in the plots for cluster peak exceedances

showing a (slight) departure from the line. In two instances, the method which used just cluster peak excesses also gave estimates of shape parameter less than -0.5 (March: $\hat{\xi}_3 = -0.894$ and July: $\hat{\xi}_7 = -0.605$), for which standard asymptotic properties are questionable (see Coles, 2001).

To investigate the inferential effects of declustering, and possibly ascertain the ‘best’ solution to deal with temporal dependence, chapter 3 examines the effects of both ignoring dependence (using all excesses) and removing dependence (declustering and using cluster peak excesses) by using simulated data, where the GPD scale and shape parameters, as well as the actual return levels, are *known*. The sensitivity of the analysis to different choices of declustering parameter, and different declustering schemes, will also be investigated, with the main focus of attention given to the effects on return level inference.

Chapter 3

Simulation study

3.1 Introduction

Chapter 2 discussed the importance of accurately and reliably estimating extreme quantiles, or return levels, in any analysis of extreme wind speeds. In the U.K., and indeed many other countries throughout the world (e.g. Australia and Canada), codes of practice exist which supply wind-loading estimates, and regulate structural design standards accordingly. For the British Isles, the British Standards Institution produces contour maps based on 50-year return level gusts. However, without strong justification, these maps are produced as the result of very simplistic extreme value analyses based on fitting the Gumbel distribution to annual gust maxima. As has been discussed, the traditional ‘annual maxima’ method is deficient; by throwing away all but the annual maximum gusts recorded at each site, the method is extremely wasteful of data. The efficiency and flexibility of the analyses employed to generate the estimates used by the British Standards Institute are further limited by using the Gumbel distribution to model extreme wind speeds; this distribution is one of three ‘types’ of limiting extreme value distribution (see section 1.2.1) – the assumption that this is the correct ‘type’ for annual maximum gust speeds has no theoretical justification, and as Walshaw (1991) shows, can lead to misleading results being obtained.

The previous chapter also discusses how we can deal with the problem of seasonal variability in our data, by either transforming the series to stationarity or adopting a ‘separate seasons’ approach. Two solutions to the problem of temporal dependence are also discussed and demonstrated through the Bradfield wind speed data – ignoring dependence (fitting to all threshold excesses) and removing dependence (fitting to a filtered set of cluster peak excesses). It is the issue of temporal dependence which is of interest here; practically, most practitioners adopt the latter approach, and proceed with filtering the se-

ries using a pre-determined declustering scheme, usually fitting a GPD to the set of cluster maxima, without considering the sensitivity of the analysis to the choice of declustering scheme. In fact, as the analysis in chapter 2 shows, there is a potential disagreement in return level estimation between the two approaches, and without further investigation we are surely unclear as to which of the approaches estimates return levels more accurately. The major pitfall of an analysis which uses declustered excesses is similar to that of the ‘annual maxima’ approach – in throwing away all but the cluster peaks, we are wasting valuable data, data which can also be considered to be extreme since they exceed the chosen threshold. We also need to consider the interaction between declustering parameter and threshold level, and need to examine a series of ‘reclustered’ excess plots in order to provide a good fit of the GPD over this threshold using a suitable declustering parameter (see section 2.5.1). The diagnostic plots (probability plots and quantile–quantile plots) shown in figure 2.9 also indicate a better fit to all excesses than to just cluster peak excesses. The potential disagreement in return level estimation between the two approaches, however, is the main cause for concern, and is the focus of our attention in the present chapter.

The main aim of this chapter is to ascertain the effects of *ignoring* temporal dependence, and *removing* temporal dependence, on return level estimation; the primary goal being able to decide which approach yields the most accurate inference. To this end, the GPD will be fitted to simulated data, first to all threshold excesses and then to cluster peak excesses, and the results of the analyses compared to the true parameter values used to simulate the data. Various criteria will be used to determine which of the approaches give the ‘best’ estimators, such as bias and efficiency, and different methods of declustering the series will be used to identify the sensitivity of the analysis to any particular declustering scheme.

3.2 The simulated data

Here, we aim to simulate a time series which somehow reflects the temporal dependence inherent in the real-life wind speed data observed at Bradfield. To achieve this, we draw upon results from bivariate extreme value theory. Let (X, Y) be a random vector and let $S(x, y) = \Pr(X > x, Y > y)$. We say that (X, Y) follows an extreme value distribution (with unit exponential margins) if and only if

$$\Pr(X > x) = e^{-x} \quad \text{and} \quad (3.1)$$

$$\Pr(Y > y) = e^{-y}, \quad (3.2)$$

for $x, y > 0$, and for any $n \geq 1$

$$S^n(x, y) = S(nx, ny) \quad (3.3)$$

again for $x, y > 0$. Pickands (1981) shows that (X, Y) follows such a distribution if and only if the joint survivor function can be written in the form

$$S(x, y) = \exp \left\{ -(x + y)A \left(\frac{y}{x + y} \right) \right\}, \quad (3.4)$$

$x, y > 0$, where

$$A(w) = \int_0^1 \max \{ (1 - w)q, w(1 - q) \} dH(q). \quad (3.5)$$

In order that (3.1) and (3.2) are satisfied we need

$$1 = \int_0^1 q dH(q) \quad (3.6)$$

$$= \int_0^1 (1 - q) dH(q); \quad (3.7)$$

for more details, see Tawn (1988). Here, H is a positive finite measure on $[0, 1]$. Following Pickands (1981) we call $A(\cdot)$ the dependence function of (X, Y) . An alternative result obtained by Pickands (1981) is that (X, Y) satisfy (3.1), (3.2) and (3.3) if and only if $W = \min(aX, bY)$ is exponential for all $a, b > 0$.

3.2.1 Parametric models for the dependence function

Parametric models for the dependence function have been known since Gumbel (1960), and have received much attention in Tiago de Oliveira's work (1984). Two classes of models exist: those that are differentiable, and those that are not everywhere differentiable. The differentiable models are those for which (X, Y) have a density, whereas non-differentiable models are singular. Here, we outline a procedure for simulating from two differentiable models.

The Gumbel model

Gumbel's bivariate model (Gumbel, 1960) has unit exponential marginals with (joint) density function

$$f(x, y; \alpha) = [(1 + \alpha x)(1 + \alpha y) - \alpha] e^{-x-y-\alpha xy}, \quad (3.8)$$

where α is a quantity (dependence measure) related to the correlation between X and Y . The conditional density function of $Y|X$ is found to be

$$f(y|x) = ([\alpha(x-1) + 1] + \alpha[1 + \alpha x]y)e^{-(1+\alpha x)y} \quad (3.9)$$

$$= (a + by)e^{-cy}, \quad (3.10)$$

where

$$a = \alpha(x-1) + 1, \quad (3.11)$$

$$c = 1 + \alpha x \quad \text{and} \quad (3.12)$$

$$b = \alpha c. \quad (3.13)$$

Replacing X and Y with X_{t-1} and X_t (respectively) gives the desired joint density function for successive pairs of observations in the simulated series. To simulate from the dependent series we can then use the following scheme:

1. Simulate the first observation, X_1 , from the unit exponential distribution;
2. Use the simulated observation, X_1 , to form the conditional distribution of $X_2|X_1$, i.e. $f(x_2|x_1) = (a + bx_2)e^{-cx_2}$. Simulate X_2 from this distribution;
3. Use the simulated observation, X_2 , to form the conditional distribution of $X_3|X_2$, i.e. $f(x_3|x_2) = (a + bx_3)e^{-cx_3}$. Simulate X_3 from this distribution;
4. Continue forming and sampling from new conditionals until you have generated the required number of observations.

The conditional distribution given in (3.10) is not invertible, and so a simple inverse transformation method to generate an observation from this conditional is not workable. Instead, we employ an envelope rejection scheme: suppose we wish to simulate X with probability density function $f(\cdot)$, and we can easily simulate values of Y (with the same support as X), which has probability density function $g(\cdot)$. Suppose further that there exists some constant K such that

$$f(x) = Kg(x), \quad \forall x; \quad (3.14)$$

that is, K is an upper bound for $f(x)/g(x)$. We draw $Y = y$ from $g(\cdot)$, and then $U = u \sim U[0, Kg(y)]$. We then accept y as a simulated value of X if $u < f(y)$; otherwise we reject and try again. This works because it distributes points uniformly over a region covering $f(x)$, and then only keeps points in the required region (under $f(x)$). So, for the conditional density $f(x_t|x_{t-1}) = (a + bx_t)e^{-cx_t}$, we can use

$$g(x_t) = Ke^{-x_t} \quad (3.15)$$

as the bounding function. To ensure this always bounds $f(x_t)$, we need to find a suitable constant K . We need

$$f(x_t) \leq g(x_t), \quad (3.16)$$

i.e.

$$K \geq (a + bx_t)e^{(1-c)x_t}. \quad (3.17)$$

Maximising the right-hand-side gives

$$K = \left(\frac{b}{c-1} \right) \exp \left\{ - \left(\frac{b - a(c-1)}{b} \right) \right\}. \quad (3.18)$$

Using this value for K ensures the enveloping function always blankets the conditional we want to simulate from, but keeps it as close to this conditional as possible (to obtain a high overall acceptance probability).

The correlation coefficient ρ is given, in terms of the dependence parameter α , as

$$\rho = -\frac{e^{1/\alpha}}{\alpha} Ei(-\alpha^{-1}) - 1, \quad (3.19)$$

where Ei is the integral logarithm. The correlation coefficient is zero for $\alpha = 0$, and decreases for increasing values of α up to $\rho = -0.40365$, valid for $\alpha = 1$.

The (symmetric) logistic model

Since, in most environmental applications, temporal dependencies tend to be positive, the logistic model provides a more appropriate model for the distribution of consecutive wind speed observations. The logistic model has dependence function

$$A(w) = \{(1-w)^{1/\alpha} + w^{1/\alpha}\}^\alpha, \quad (3.20)$$

for $0 < \alpha \leq 1$, with corresponding (joint) survival function

$$S(x, y; \alpha) = \exp \left\{ - \left(x^{\frac{1}{\alpha}} + y^{\frac{1}{\alpha}} \right)^\alpha \right\}. \quad (3.21)$$

Independence and complete dependence correspond to $\alpha = 1$ and $\alpha = 0$ respectively; the popularity of this model owes much to this flexibility. The variables are exchangeable and have correlation coefficient

$$\rho = \alpha \{\Gamma(2\alpha)\}^{-1} \{\Gamma(\alpha)\}^2 - 1. \quad (3.22)$$

As with the Gumbel model, the marginal distributions are both unit exponential. Simulating a series with successive pairs being drawn from the distribution given by (3.21) can be achieved using the same techniques as before, i.e. using an envelope rejection scheme.

3.2.2 GPD marginals

Both parametric models (Gumbel and logistic) described in 3.2.1 can be used to simulate a first-order Markov chain with unit exponential marginals. Since a GPD with $\xi = 0$ is taken to be the limit as $\xi \rightarrow 0$, and is the exponential distribution, this corresponds to a series with GPD(1,0) marginals. However, as can be seen from the results for the Bradfield data (tables 2.2 and 2.3), in real-life applications it is often the case that $\xi \neq 0$. A simple transformation can, however, be used on the simulated data with unit exponential marginals to create a series with GPD marginals for any specified σ and ξ . If $x_1, x_2, \dots, x_n$ is a series of unit exponential variates with distribution function $F(x)$, and $y_1, y_2, \dots, y_n$ a series of GPD variates with distribution function $G(y)$, then equating these gives

$$\begin{aligned} G(y) &= F(x) \\ 1 - \left[1 + \xi \frac{y_i}{\sigma}\right]_+^{-1/\xi} &= 1 - e^{-x_i}. \end{aligned} \quad (3.23)$$

Solving for y_i , we get

$$y_i = \frac{\sigma}{\xi} (e^{\xi x_i} - 1), \quad (3.24)$$

where equation (3.24) provides the transformation from unit exponential to GPD(σ, ξ). Figure 3.1 shows a simulated time series with unit exponential marginals using Gumbel's model (with $\alpha = 1$ and so $\rho = -0.40365$; in figure 3.1, the (lag 1) autocorrelation is -0.39892). Figure 3.2 shows a simulated series, again with unit exponential marginals but this time using the logistic model ($\alpha = 0.2$ – strong dependence). Figure 3.3 shows a time series plot of simulated data with GPD marginals, transformed from the series in figure 3.2 using (3.24) ($\sigma = 1$, $\xi = -0.4$). Notice the dependence structure of the series in figures 3.1 and 3.2; recall that, under the more restrictive Gumbel model, serial correlation can range between -0.40365 and 0 only. Also, the series in figure 3.3 has stronger serial correlation than the series from which it was derived (it has a lag 1 sample autocorrelation coefficient of 0.938 compared to 0.904 for the series in figure 3.2). Though we might expect the transformation through (3.24) to change the dependence structure, other work has shown that such changes are small. Indeed, for the purposes of this simulation study, we are not interested in the precise strength of temporal dependence, just whether this is ‘strong’, ‘moderate’ or ‘weak’ (see section 3.3 below).

3.3 Simulation study details

The aim of the simulation study is to investigate the consequences of declustering on return level inference, and to verify whether or not we should decluster at all. We use the logistic

Figure 3.1: Simulated time series with unit exponential marginals (using Gumbel's model)

Figure 3.2: Simulated time series with unit exponential marginals (using the logistic model)

Figure 3.3: Simulated time series with GPD marginals (transformed from figure 3.2)

model to simulate our data since with this model all positive levels of dependence, ranging from complete independence ($\alpha = 1$) to perfect dependence ($\alpha \rightarrow 0$), can be incorporated into the data. Data are simulated using three levels of dependence: ‘weak’ dependence ($\alpha = 0.8$), ‘moderate’ dependence ($\alpha = 0.5$) and ‘strong’ dependence ($\alpha = 0.2$). We also choose five different values of shape parameter ξ : $-0.4, -0.1, 0, 0.3$ and 0.8 to reflect the various tails which might be observed in real-life data. Thresholds u are chosen such that approximately 5% of the simulated values exceed them; these thresholds are used to observe clustering of extremes. This is achieved by setting the GPD survival function $S(u)$ equal to 0.05 and then rearranging to find u . Thus, we find

$$u = \frac{\sigma}{\xi} (0.05^{-\xi} - 1); \quad (3.25)$$

so for each of the five values of ξ we have five corresponding thresholds u which will retain approximately 5% of the data above them. ‘Runs’ declustering (2.4.2) will be used to filter a set of ‘independent’ cluster maxima, with various declustering parameters used to identify any sensitivity to the choice of cluster termination interval. The scale parameter σ will remain unit constant; however, σ is threshold-dependent and will vary according to the threshold we use to select exceedances. Further, since all simulated values (and not just those greater than u) have a $\text{GPD}(1, \xi)$ marginal distribution (by the way we have constructed the series), and we are only incorporating those observations which are greater than u , we are essentially modelling excesses over a threshold u given that these excesses already follow a GPD with threshold 0. This leads us to the threshold stability property, mentioned in section 1.3 (and shown below).

Result: Distribution of $X > x + u_2 | X > u_2$ given that $X > x + u_1 | X > u_1$ is $\text{GPD}(\sigma, \xi)$, and $u_2 > u_1$.

Given

$$\Pr(X > x + u_1 | X > u_1) = \left[1 + \frac{\xi x}{\sigma} \right]_+^{-1/\xi}, \quad (3.26)$$

we want to find

$$\Pr(X > x + u_2 | X > u_2). \quad (3.27)$$

It is easily seen that

$$\Pr(X > x + u_2 | X > u_2) = \frac{\Pr(X > x + u_2 | X > u_1)}{\Pr(X > u_2 | X > u_1)}, \quad (3.28)$$

for $u_2 > u_1$. This is equal to

$$\frac{\Pr(X > (x + u_2 - u_1) + u_1 | X > u_1)}{\Pr(X > (u_2 - u_1) + u_1 | X > u_1)} = \frac{\left[1 + \frac{\xi(x+u_2-u_1)}{\sigma}\right]_+^{-1/\xi}}{\left[1 + \frac{\xi(u_2-u_1)}{\sigma}\right]_+^{-1/\xi}} \quad (3.29)$$

$$= \left[1 + \frac{\xi x}{\sigma + \xi(u_2 - u_1)}\right]_+^{-1/\xi} \quad (3.30)$$

$$= \left[1 + \frac{\xi^* x}{\sigma^*}\right]_+^{-1/\xi^*}, \quad (3.31)$$

for all $x > 0$. This is the survival function of the $\text{GPD}(\sigma^*, \xi^*)$, with $\xi^* = \xi$ and $\sigma^* = \sigma + \xi(u_2 - u_1)$. In the simulation study, $u_2 = u$ and $u_1 = 0$, so the excesses above u can be modelled by a $\text{GPD}(\sigma^*, \xi)$, where $\sigma^* = \sigma + \xi u$.

Table 3.1 shows a layout of the study, which is split up into 15 components. In each component we simulate a series with GPD marginals; we then fit the GPD to all threshold excesses and obtain MLEs for σ^* and ξ . We then decluster the series using ‘runs’ declustering (having specified some declustering parameter κ), and proceed to fit the GPD to the set of ‘independent’ cluster peak excesses – again obtaining MLEs for σ_p^* and ξ_p , the subscript p denoting the use of cluster *peak* excesses rather than *all* excesses. Once the GPD parameters have been estimated, we can estimate return levels for both cases – using all excesses and using cluster peak excesses. These estimates can then be compared to the *true* return levels obtained from equation (3.25), replacing 0.05 with $1/t$, where t is the return period. Each simulated series has length 20,000, leaving about 1000 observations available for analysis after declustering. Longer datasets could be simulated, though this would increase computing time. Though the standard deviations of the sampling distributions will depend on the number of exceedances used in the analysis, this is unimportant: it is the effect of declustering on parameter estimation we are interested in, and since the threshold is always set at the 95% quantile, the ratio of cluster peak exceedances to all threshold exceedances will be approximately the same no matter how large the simulated datasets are. Each series, unlike the Bradfield wind speed data, is also stationary, with no ‘seasonal’ cycles; this avoids the need to ‘prewhiten’ the series prior to analysis or fit a separate model to each ‘month’ of data, and makes the calculation of return levels much more straightforward (using equation (2.9) instead of (2.14)). The simulation–estimation procedure is repeated 1000 times to give sampling distributions of length 1000 for each of the GPD parameters (and return levels).

Standard errors for the GPD parameters and return levels can be estimated by calculating standard deviations of the sampling distributions for these parameters. Similarly,

	$\xi = -0.4$ $\sigma^* = 0.302$	$\xi = -0.1$ $\sigma^* = 0.741$	$\xi = 0.0$ $\sigma^* = 1.0$	$\xi = 0.3$ $\sigma^* = 2.456$	$\xi = 0.8$ $\sigma^* = 10.99$
$\alpha = 0.2$	1	4	7	10	13
$\alpha = 0.5$	2	5	8	11	14
$\alpha = 0.8$	3	6	9	12	15

Table 3.1: Layout of the simulation study (cell shows component number)

confidence intervals can also be obtained directly from the sampling distributions, thus avoiding the need to use the delta method (in the case of return levels) or profile-likelihood methods. However, for comparison, for each parameter estimate, standard likelihood theory is used to estimate associated standard errors, thus obtaining a sampling distribution for the standard errors also. The standard deviations of the sampling distributions for each parameter can then be compared to the mean of the sampling distribution for the estimated standard error for each parameter.

3.4 Results

A full set of results of the simulation study is given in appendix A.1; a discussion, with selected summaries and plots of the most interesting points, is given here. The effects of declustering on estimates of the GPD parameters σ^* and ξ are discussed first, with consequences for return level inference then examined. At first, a cluster termination interval of 10 observations was used to determine clusters of exceedances, and the set of cluster peaks used to fit the GPD. However, various extensions to the study were then undertaken, with different methods of declustering employed to detect any sensitivity to the declustering scheme used.

3.4.1 Estimates of σ^* and ξ

The results in appendix A.1 show that, in most cases, both σ^* and ξ were more accurately estimated when all threshold excesses were used in the analysis. For all 15 components of the study, the means of the sampling distributions for both GPD parameters were closer to the true values than their ‘cluster peaks’ counterparts. For example, in study 1, the target values were $\sigma^* = 0.302$ and $\xi = -0.4$; using all excesses gave $\hat{\sigma}^* = 0.305$ (0.037) and $\hat{\xi} = -0.424$ (0.068), compared to $\hat{\sigma}_p^* = 0.492$ (0.091) and $\hat{\xi}_p = -0.726$ (0.164) using cluster peak excesses (using sampling distribution means (and standard deviations in parentheses)). Similar discrepancies were observed elsewhere in the simulation study, with σ^* being consistently over-estimated, and ξ under-estimated, when just cluster peak

excesses were modelled. In many instances the 95% (and sometimes the 99%) confidence intervals for estimates based on using cluster peak excesses failed to include the actual parameter values, whereas all intervals for estimates which used all threshold excesses successfully covered these. The accuracy of estimates which used cluster peak excesses also seems to deteriorate as the serial dependence in the series is made stronger; however, even when $\alpha = 0.8$ and we have (reasonably) weak dependence, using cluster peak excesses still fails to provide estimates as accurate as those which have used all threshold excesses. Figure 3.4 shows the sampling distributions for σ^* and ξ in study 1. The heavy line indicates the use of all threshold excesses, the thin line just cluster peak excesses. The vertical dashed lines indicate the position of the actual parameter values. The plots clearly show that the estimates based on all threshold excesses estimate the parameters more accurately (and with less error). Similar results were obtained for the other fourteen studies, with density plots similar to those in figure 3.4 (see appendix A.1).

Figure 3.4: Sampling distributions of $\hat{\sigma}^*$ and $\hat{\xi}$ (study 1) using all threshold excesses (heavy line) and cluster peak excesses (thin line)

3.4.2 Return level estimates

As has already been discussed, it is usually more convenient (and practically important) to interpret extreme value models in terms of extreme quantiles, or return levels. Since we have simulated stationary Markov chains, equation (2.10) (and (2.11) for the case $\xi = 0$) can be used to estimate the t -observation return level, with σ and ξ being replaced with $\hat{\sigma}^*$ and $\hat{\xi}$ (or $\hat{\sigma}_p^*$ and $\hat{\xi}_p$ in the case which uses cluster peak excesses, the estimates being the means of the sampling distributions), and λ_u being estimated by the mean exceedance rate over threshold u . As with the GPD parameters, standard errors for return levels can be estimated by calculating the standard deviation of the sampling distribution of $\hat{q}_t$; however, we can also compute standard errors for each return level estimate, and look at the sampling distribution of these standard errors.

The results for return level estimation show that, when $\alpha = 0.2$, and the simulated series displays strong serial correlation, the 1000-observation return level q_{1000} is systematically under-estimated when just cluster peak excesses are used to fit the GPD. This under-estimation is consistent for all studies whose simulated Markov chain displays strong temporal dependence (studies 1, 4, 7, 10 and 13 – see table 3.1). In these studies, using all threshold exceedances in the analysis (and so ignoring dependence) gave more accurate return level estimates. In other studies, where this temporal dependence is less obvious ($\alpha = 0.5$ and $\alpha = 0.8$), the sampling distributions of q_{1000} were very similar when using all excesses and just cluster peak excesses, especially when $\alpha = 0.8$. However, the case $\alpha = 0.2$ provides the most interesting scenario, since this level of dependence reflects more realistically the sort of temporal dependence which might be observed in real-life environmental data. Figure 3.5 shows sampling distributions of q_{1000} for the three levels of dependence used (for $\xi = -0.4$); Similar results were obtained for other components in the study. Table 3.2 summarises some results of the simulation study for return level inference.

Both figure 3.5 and table 3.2 illustrate the effect of declustering on return level estimation. Declustering has had virtually no effect on return level inference when $\alpha = 0.5$ and $\alpha = 0.8$; figure 3.5 shows the sampling distributions under both scenarios to be barely indistinguishable, and the means of these sampling distributions (see first two columns of table 3.2), as in other studies where $\alpha = 0.5$ or 0.8 (see columns 4 and 5 in table 3.2), are barely unchanged when we use the declustered series. However, clear differences between the two approaches are noticeable when $\alpha = 0.2$. As the far-right plot in figure 3.5 shows, there is a clear under-estimation when cluster peak excesses are modelled. Table 3.2 supports this finding, with the means of the sampling distributions using cluster peak excesses being significantly lower than both the true value and the estimates based on all threshold

excesses. For example, in the first study, the actual return level value is $q_{1000} = 2.342$; When $\alpha = 0.2$, the mean of the sampling distribution for the 1000-observation return level (using all threshold excesses) is 2.325, whereas using cluster peak excesses gave a mean value of 2.188 (see column 3 in table 3.2). In the same example, the 95% confidence interval from the sampling distribution based on using cluster peak excesses also failed to include the actual parameter value. Similar results were found elsewhere in the simulation study, as the last column of table 3.2 illustrates.

Figure 3.5: Sampling distributions of $\hat{q}_{1000}$ for all excesses (heavy line) and cluster peak excesses (thin line) for three strengths of temporal dependence. The vertical lines represent the actual 1000-observation return level

	$q_{1000} = 2.342$			$q_{1000} = 23.144$		
	$\alpha = 0.8$	$\alpha = 0.5$	$\alpha = 0.2$	$\alpha = 0.8$	$\alpha = 0.5$	$\alpha = 0.2$
Mean (All excesses)	2.338	2.334	2.325	23.202	23.012	22.864
Mean (cluster peaks)	2.351	2.343	2.188	23.261	22.367	17.518
95% CI (All excesses)	(2.293, 2.373)	(2.286, 2.375)	(2.249, 2.395)	(17.845, 29.248)	(18.103, 30.134)	(15.259, 37.714)
95% CI (Cluster peaks)	(2.304, 2.390)	(2.296, 2.381)	(2.114, 2.302)	(18.310, 29.900)	(17.528, 28.643)	(13.348, 23.934)

Table 3.2: Comparison of some return level estimates based on (a) all threshold exceedances and (b) cluster peak exceedances. The table quotes the sampling distribution means and associated 95% confidence intervals for the case $\xi = -0.4$ and $\xi = 0.3$ (components 1–3 and 10–12)

As section 2.1 discusses, accurate estimation of return levels is extremely important; for example, the British Standards Institution calculates the strength requirements of buildings using estimates of 50-year return level gust speeds, and the governmental coastal flood defence division uses return level estimates to calculate the required height of sea walls to prevent flooding. If methods have been employed which result in these return levels being under-estimated, human life could be at risk if, for example, a building cannot withstand a series of extreme gusts or a sea wall has not been built high enough to defend against extremely high sea-levels during a storm. Indeed, this study suggests that one particular method which is often used in practice – ‘runs’ declustering – could lead to (significantly) under-estimated return levels.

3.4.3 Performance of estimators

The sampling distributions of σ^* , ξ and q_{1000} have been examined further to provide insight into which approach gives the ‘best’ estimates. The under-estimation of the shape parameter ξ , and the over-estimation of the scale parameter σ^* , using cluster peak excesses, was illustrated in section 3.4.1. More importantly, section 3.4.2 showed the under-estimation of the 1000-observation return level q_{1000} when cluster peak excesses were used in series with strong temporal dependence. This section investigates the performance of the estimators by looking at their bias, efficiency and mean squared error (MSE). Obviously, we’d like our estimators to be unbiased, i.e. $E[\hat{\theta}] = \theta$; efficient, i.e. have small variance; and have as small a MSE as possible, where

$$\text{MSE} = E[(\hat{\theta} - \theta)^2], \quad (3.32)$$

θ being a generic parameter. The MSE for each estimate $\hat{\sigma}^*$, $\hat{\xi}$ and $\hat{q}_{1000}$ is thus found by subtracting the actual parameter value from each parameter estimate in the sampling distribution, squaring these differences and then finding the mean of these differences. The bias of each parameter can be estimated by subtracting the actual parameter value from the mean of the sampling distribution. The relative efficiency of one parameter estimate to another, e.g. $\hat{\sigma}^*$ to $\hat{\sigma}_p^*$, can be found by calculating

$$\text{eff}(\hat{\sigma}^*) = \frac{\text{var}(\hat{\sigma}^*)}{\text{var}(\hat{\sigma}_p^*)}. \quad (3.33)$$

Appendix A.1 gives the results for all components of the simulation study; table 3.3 shows the results for studies 1, 2 and 3 – i.e. the case where $\xi = -0.4$ (for all three levels of dependence).

As table 3.3 shows, the results are, generally, in favour of using all excesses rather than

		$\hat{\sigma}^*$	$\hat{\sigma}_p^*$	$\hat{\xi}$	$\hat{\xi}_p$	$\hat{q}_{1000}$	$\hat{q}_{p,1000}$
Estimated	$\alpha = 0.8$	0.008	0.126	-0.024	-0.218	-0.004	0.009
bias	$\alpha = 0.5$	0.006	0.145	-0.021	-0.250	-0.008	0.001
	$\alpha = 0.2$	0.003	0.190	-0.024	-0.326	-0.017	-0.154
	$\alpha = 0.8$	0.001	0.017	0.003	0.053	0.000	0.001
MSE	$\alpha = 0.5$	0.001	0.023	0.004	0.070	0.001	0.001
	$\alpha = 0.2$	0.001	0.044	0.005	0.133	0.002	0.004
%	$\alpha = 0.8$	100	39.748	100	49.152	100	85.543
Relative	$\alpha = 0.5$	100	30.532	100	40.576	100	91.476
efficiency	$\alpha = 0.2$	100	16.773	100	17.130	100	112.825

Table 3.3: Estimated bias, mean squared errors and relative efficiencies of parameter estimates

just cluster peak excesses. For estimates of σ^* and ξ , the estimated bias is always closer to zero when using all threshold excesses. In the case which is of most interest – i.e. strong dependence – the estimated bias for $\hat{q}_{1000}$ is closer to zero when using all excesses. Again, similar results were obtained for other components of the study, suggesting estimates of the 1000-observation return level are more biased when using cluster peak excesses. The mean squared errors for all estimators were consistently smaller when all threshold excesses were used in the analysis, again suggesting analyses based on all threshold exceedances provide the better estimates (though it should be noted that the cost in MSE in terms of not using all excesses is less for return levels than it is for the GPD parameters themselves). Estimates using all threshold excesses were also more efficient than their cluster peaks counterparts, except for $\hat{q}_{1000}$ when $\alpha = 0.2$. Here, the efficiency of $\hat{q}_{p,1000}$ was 112.825% relative to the estimate which used all threshold excesses. However, though the sampling distribution of $\hat{q}_{1000}$ is more variable than that of $\hat{q}_{p,1000}$, the estimate found using all threshold excesses is more accurate; as table 3.2 shows, the 95% confidence interval from the sampling distribution of $\hat{q}_{p,1000}$ fails to include the actual parameter value q_{1000} .

3.4.4 Standard errors

Each time the routine maximised the log-likelihood to find MLEs for σ^* and ξ , the estimated standard error for each was obtained by inversion of the observed information matrix. Thus, as well as obtaining sampling distributions for σ^* and ξ , sampling distributions were also obtained for the standard errors of these estimates. Calculating the mean of these distributions should give an idea of the true standard errors of the param-

		$\hat{\sigma}^*$	$\hat{\sigma}_p^*$	$\hat{\xi}$	$\hat{\xi}_p$	$\hat{q}_{1000}$	$\hat{q}_{p,1000}$
$\alpha = 0.8$	St. dev.	0.024	0.039	0.051	0.074	0.021	0.022
	mean(e.s.e.)	0.023	0.039	0.050	0.064	0.019	0.020
$\alpha = 0.5$	St. dev.	0.026	0.047	0.056	0.090	0.025	0.025
	mean(e.s.e.)	0.024	0.046	0.051	0.075	0.019	0.019
$\alpha = 0.2$	St. dev.	0.039	0.095	0.075	0.174	0.041	0.040
	mean(e.s.e.)	0.024	0.078	0.053	0.133	0.019	0.033

Table 3.4: Comparison of estimated standard errors with standard deviations of estimators

eters; however, as section 2.4.1 details, in the situation where we have serial dependence these estimated standard errors will under-estimate the true standard errors due to the assumption of independence. A more accurate estimate of the standard errors of σ^* and ξ can be found by computing the standard deviations of the sampling distributions for these parameters. Table 3.4 compares the estimated standard errors with the standard deviations of the estimators (for studies 1, 2 and 3).

The under-estimation of the true standard error of each parameter (if we assume independence) is obvious in table 3.4. We see, for the case $\alpha = 0.8$ (weak dependence), only a slight under-estimation given by the estimated standard errors (in some cases this under-estimation is so small that when taken to three decimal places both the standard deviation of the parameter sampling distribution, and the mean of the sampling distribution of the estimated standard errors, appear identical). However, as the strength of dependence increases, we see an increase in this under-estimation, most noticeable when $\alpha = 0.2$ (strong dependence). This confirms the suspicion that simply assuming independence and using all threshold excesses in the analysis results in under-estimated standard errors of the parameters. Even when declustering has been used to filter out a set of ‘independent’ threshold excesses, the mean of the distributions of estimated standard errors slightly under-estimates the standard error – possibly implying that the declustering parameter $\kappa = 10$ was not set high enough to ensure neighbouring clusters were independent. To investigate this further, section 3.5.1 considers the effect of varying this declustering interval.

3.4.5 Results at other levels of dependence

So far, we have compared estimates of the GPD scale and shape parameters and the 1000-observation return level when incorporating all threshold excesses, and cluster peak excesses only, in the analysis. These comparisons have been made when series are charac-

terised by ‘strong’, ‘medium’ and ‘weak’ temporal dependence, with corresponding values of the logistic dependence parameter α of 0.2, 0.5 and 0.8 (respectively). We have seen a substantial under-estimation of q_{1000} when processes were simulated with $\alpha = 0.2$, though at the weaker levels of temporal dependence this under-estimation was not apparent. In fact, careful inspection of table 3.2 suggests that, particularly at the weakest level of dependence ($\alpha = 0.8$), using cluster peak excesses might result in slightly *over-estimated* return levels. We now investigate the estimation of return levels for a whole range of strengths of temporal dependence, to see if there is any association between α and the bias for q_{1000} (and for the GPD scale and shape parameters also).

Figure 3.6 shows plots of bias in estimates of σ^* , ξ and q_{1000} against logistic dependence parameter α , for various values of α between 0.01 and 1. First-order Markov chains (of length 50,000) are simulated with GPD marginals (as discussed in section 3.2; here, we use $\sigma = 1$ (and so $\sigma^* = 0.30171$) and $\xi = -0.4$) for each value of α , and the GPD fitted to the set of all threshold excesses, and cluster peak excesses, separately. From the estimates of σ^* and ξ , estimates of q_{1000} are also obtained from equation (2.10). Thus, sampling distributions are obtained for σ^* , ξ and q_{1000} , and the bias for each parameter obtained for each parameter estimate. The mean bias is then used as the estimated bias for each parameter. As before, the heavy line indicates the use of all threshold excesses, the thin line cluster peak excesses. The jagged traces are simply an artefact of sampling variation.

Looking at the plot for return levels first, we see the under-estimation when using cluster peak excesses for $\alpha = 0.2$ (as found in the simulation study). However, from figure 3.6 we also learn that, as the strength of dependence is increased further, this under-estimation is even more profound; for example, when $\alpha = 0.01$, the estimated bias for q_{1000} is just less than -0.5 . Further, figure 3.6 illustrates the *over-estimation* of q_{1000} at weaker levels of dependence – particularly for values of α larger than about 0.4. Return level estimates using all threshold excesses are better, with relatively small bias at very strong levels of dependence (α less than about 0.1) and virtually zero bias elsewhere. In comparison, both approaches give little bias in σ^* and ξ at very strong levels of dependence, with rapidly increasing bias as α approaches 0.2 from below for the cluster peaks method; however, as α increases beyond 0.2, the estimated bias for σ^* and ξ (under the cluster peaks approach) decreases again. For the approach which uses all threshold excesses, the estimated bias for σ^* and ξ is virtually insignificant for all α greater than about 0.2. Generally, the distribution of cluster peak excesses appears to have a shorter tail and a larger spread than that of all threshold excesses. In the limit, the plots for cluster peak excesses should converge to the dashed line (which indicates zero bias) since the limiting

Figure 3.6: Plots of bias versus level of temporal dependence (characterised by the logistic dependence parameter α) for σ^* , ξ and q_{1000}

distribution of cluster peak excesses is exactly the same as that for all excesses; indeed, in further studies we have found that, if the length of the simulated series is increased and the threshold raised, plots of the type shown in figure 3.6 show less bias under the cluster peaks approach than is shown here.

One striking feature of figure 3.6 is the under-estimation of return levels in the cluster peaks analysis at the strongest levels of dependence, but comparatively good corresponding estimates of σ^* and ξ . One possible explanation for this is under-estimation of the threshold exceedance rate λ_u – in both approaches, but particularly in the cluster peaks analysis. When α is set at 0.01, for example, clusters of extremes tend to be large; in a series of length 50,000, runs declustering with $\kappa = 10$ tends to identify, typically, between twenty and thirty such clusters. Thus, the threshold exceedance rate for a cluster peaks analysis is often as small as 0.0004, and sometimes even smaller. Even with relatively small bias in estimates of σ^* and ξ , such an underestimation of λ_u can result in huge under-estimation of q_{1000} obtained through equation (2.10).

3.5 Extensions

The sensitivity of the results to the choice of declustering scheme has yet to be investigated; in all situations where the GPD was fitted to a set of ‘independent’ cluster maxima, ‘runs’ declustering (section 2.4.2) was used to identify clusters, and the maximum value in each cluster selected to filter out a set of cluster peaks. The declustering parameter, κ , was chosen (arbitrarily) to be 10; however, this may not have been large enough (especially where the simulated series displayed strong serial dependence) to ensure independence amongst members of the filtered set of exceedances. Further, the choice of cluster maximum instead of, say, the first cluster exceedance, or indeed a randomly chosen cluster observation, may have influenced the results (and had an adverse effect on return level inference). The following extensions to the simulation study attempt to identify any such sensitivity by investigating the effects of other declustering methods.

3.5.1 Varied declustering parameter κ

Here, the entire simulation study is repeated, but with different declustering parameters used to identify clusters in the ‘runs’ method of declustering. We have already used $\kappa = 10$; Now $\kappa = 1, 5, 20$ and 50 are used. Though a cluster termination interval of $\kappa = 1$ will probably be insufficient to give an independent set of filtered exceedances (especially when $\alpha = 0.2$), and $\kappa = 50$ unnecessarily large (even when $\alpha = 0.2$), these two values are used to find any ‘pattern’ in return level estimation as the termination interval is increased.

Figure 3.7: Sampling distributions of $\hat{q}_{1000}$ for all excesses (heavy line) and cluster peak excesses (thin line) for varied declustering parameters: $\kappa = 1, 5, 20$ and 50

Figure 3.7 shows the sampling distributions of the 1000-observation return level for analyses using all threshold excesses and cluster peak excesses only (for studies 1, 2 and 3 – i.e. $\xi = -0.4$ for $\alpha = 0.2, 0.5$ and 0.8 respectively); similar results were obtained in other studies. As these plots suggest, changing the declustering parameter κ has had little, if any, effect on return level inference: though there is little difference between using cluster peak excesses and all threshold excesses when $\alpha = 0.8$ and $\alpha = 0.5$, again, as in the original study, we see a clear under-estimation of q_{1000} using cluster peak excesses – no matter how wide the cluster termination interval κ .

3.5.2 “Automatic” declustering

Section 3.5.1 investigated the effect of changing the declustering parameter κ when identifying clusters using the ‘runs’ method. Though we concluded that the different choices of κ had no effect on return level inference in strongly dependent series (q_{1000} was still systematically under-estimated), choices of κ were still made arbitrarily with no justification or support from information in the data. Further, only a few values were tried – namely $\kappa = 1, 5, 20$ and 50 . Ferro and Segers (2003) propose an “automatic” method of declustering extreme values that is justified by an asymptotic result for the arrival times between threshold exceedances. The scheme relies on the extremal index, θ (see section 1.4); in the past, θ has been interpreted as the reciprocal of the (limiting) mean cluster size (Leadbetter, 1983), and so in practice estimation of θ has relied upon the identification of clusters, usually using the ‘runs’ or ‘blocks’ declustering methodology. However, Ferro and Segers present a method of estimating θ without declustering. Suppose N exceedance times are observed: $S_1 < S_2 < \dots < S_N$, and $T_i = S_{i+1} - S_i$, for $i = 1, \dots, N - 1$ are the inter-arrival times. As we shall see (in section 6.3), the distribution of the T_i is given by a mixture distribution (see equation 6.16); the second moment of this distribution can be shown to be $E(T^2) = 2/\theta$ (see Ferro and Segers, 2003). Equating this to the empirical second moment, replacing θ with $\hat{\theta}$ and solving for $\hat{\theta}$ gives Ferro and Segers’ ‘intervals’ estimator as

$$\hat{\theta} = \frac{2 \left\{ \sum_{i=1}^{N-1} (T_i - 1) \right\}^2}{(N-1) \sum_{i=1}^{N-1} (T_i - 1)(T_i - 2)}. \quad (3.34)$$

Typically, some large inter-arrival times are observed from series which display strong temporal dependence (since, after a threshold downcrossing, there tends to be many successive observations sub-threshold before another threshold upcrossing); in equation (3.34), this gives rise to small values of $\hat{\theta}$. On the other hand, most inter-arrival times from near-independent series are quite small, giving relatively large values of $\hat{\theta}$ through equation (3.34).

Based on this estimator, Ferro and Segers propose a method of declustering which avoids the need to make an arbitrary choice of declustering parameter, allowing the inter-arrival times between threshold exceedances to identify clusters in a more ‘natural’ way. They categorise the inter-arrival times of threshold excesses into two types: independent inter-cluster times and independent sets of intra-cluster times. Given the set of inter-arrival times $T_i = S_{i+1} - S_i$, $i = 1, \dots, N - 1$, we can assume that the largest $C - 1 = \lfloor \theta N \rfloor$

inter-arrival times are approximately independent inter-cluster times that divide the remainder into approximately independent sets of intra-cluster times. More precisely, if $T_{(C)}$ is the C -th largest inter-arrival time and T_{i_j} is the j -th inter-arrival time to exceed $T_{(C)}$, then $\{T_{i_j}\}_{j=1}^{C-1}$ is a set of approximately independent inter-cluster times. Let $\mathcal{T}_j = \{T_{i_{j-1}+1}, \dots, T_{i_j-1}\}$, where $i_0 = 0$, $i_C = N$ and $\mathcal{T}_j = \emptyset$ if $i_j = i_{j-1} + 1$. Then $\{\mathcal{T}_j\}_{j=1}^C$ is a collection of approximately independent sets of intra-cluster times. Essentially, this approach is equivalent to ‘runs’ declustering with $\kappa = T_{(C)}$, where $C = \lfloor \theta N \rfloor + 1$. In practice, θ is replaced with the intervals estimate given in (3.34).

This “automatic” method of declustering was applied to the simulated data. The resulting sampling distributions for the 1000-observation return level are shown in figure 3.8 (again for studies 1, 2 and 3). Even when declustering has been carried out using Ferro and Segers’ “automatic” method, avoiding the need to make an arbitrary choice of declustering parameter, figure 3.8 shows the 1000-observation return level still being under-estimated relative to the analysis which uses all threshold excesses. From the investigations in section 3.5.1 and the present section, we can therefore conclude that the systematic under-estimation of the 1000-observation return level when using cluster peak excesses is probably not a result of any arbitrary or incorrect choice of cluster termination interval used in the process of declustering.

Figure 3.8: Sampling distributions of $\hat{q}_{1000}$ for all excesses (heavy line) and cluster peak excesses (thin line) chosen via Ferro and Segers’ “automatic” declustering method

3.5.3 Choice of cluster observation

Another aspect of the declustering scheme used in the simulation study which could be altered is the choice of cluster observation used to create the set of ‘independent’ threshold excesses. The most common approach taken in practice is to select the set of cluster maxima; however, the cluster peaks are obviously not representative of the cluster to which they belong – they do, after all, represent the highest observation within each cluster. Though, in the limit, it can be shown that a randomly chosen exceedance has exactly the same distribution as that of a cluster peak (this apparent paradox is a consequence of length-biased sampling – see Anderson (1990) and section 3.6 for more details), it would be unsurprising to find that, in practice, both approaches return different GPD parameter estimates. To address this issue, this section repeats the simulation study, but looks at the effects of three different methods of selecting an observation from within each identified cluster: the first uses the first observed cluster value, the second uses a randomly selected cluster observation, and the third looks at modelling the set of mean cluster excesses. In all cases, Ferro and Segers’ “automatic” declustering has been used to identify clusters.

First observed cluster exceedances

Here, instead of selecting the maximum exceedance within each cluster, the first exceedance is chosen. The idea that exceedances other than the cluster maximum may be incorporated into the analysis under this scheme has some appeal, though as figure 3.9 shows, little improvement is made upon return level estimation – in fact, the 1000-observation return level appears to be even more greatly under-estimated when $\alpha = 0.2$, and we see an under-estimation when declustering was used at the lower levels of dependence which was not previously observed. In many instances, particularly when the series displayed strong serial dependence ($\alpha = 0.2$), this method of declustering was equivalent to selecting the cluster minimum, with the first observed cluster exceedance being just over the chosen threshold. Thus, whereas in the original scheme we restricted the filtered set of cluster exceedances to be the maximum observed exceedances from each cluster, in many cases this scheme had similar restrictions, with the set of cluster minima being obtained instead. Also, the process of selecting the first observed cluster exceedance is an arbitrary method; indeed, any cluster exceedance may have been chosen, such as the last observed cluster value. The desire to avoid such an arbitrary choice motivates the next filtering scheme.

Randomly chosen cluster exceedances

Under this filtering scheme, clusters of extreme values are identified in the usual way; however, rather than choosing a specific cluster inhabitant (such as the cluster maximum or the first cluster value), an observation from within each cluster is selected at random. This approach avoids the need to pre-determine which of the cluster observations is to join the set of filtered exceedances, thus removing any choices which might be considered arbitrary. The second row of figure 3.9 shows the result of this scheme on return level inference – yet again, the under-estimation of the 1000-observation return level when $\alpha = 0.2$ is clearly evident. Further, we see an under-estimation at the $\alpha = 0.5$ and $\alpha = 0.8$ levels of dependence, which did not occur when using cluster peak excesses.

Mean cluster excesses

This scheme attempts to filter out a set of values which are ‘representative’ of the cluster to which they belong. Again, “automatic” declustering is used to identify clusters of extreme values; however, instead of selecting one occupant from within each cluster, information from all the cluster occupants is pooled by calculating the mean exceedance. The GPD is then fitted to the set of mean excesses. The bottom row of figure 3.9 shows the effects on return level inference for this scheme – again, we see the 1000-observation return level is clearly under-estimated when the series displays strong dependence ($\alpha = 0.2$). In fact, this method of filtering seems worse again, with the under-estimation due to declustering exacerbated, even at the level of weak dependence ($\alpha = 0.8$).

Figure 3.9: Sampling distributions of $\hat{q}_{1000}$ for all excesses (heavy line) and cluster peak excesses (thin line) for different declustering procedures. Top row: first cluster exceedance; middle row: randomly chosen cluster exceedance; bottom row: mean cluster exceedance

3.6 A model for cluster peak excesses

In the limit, the distribution of excess for a randomly chosen threshold exceedance *has exactly the same distribution as that of a cluster peak*. This apparent paradox is a consequence of length-biased sampling: a randomly chosen exceedance has a disproportionate chance of coming from a large cluster, and in large clusters there tend to be large excesses. The special structure of the limiting point process ensures that the two effects work together to give exact equality in the limit (see Anderson (1990) for more detail).

However, as both the present chapter and chapter 2 have shown, when working with finite samples the distributions of *cluster peak excesses* and *all excesses* are different. Though both seem to be well represented by a generalised Pareto distribution, the parameters of the GPD are different under each approach. Similarly, we have shown that return level estimates are different under the two approaches, and when analysing simulated data where the *true* return levels were *known*, the approach which fitted to cluster peak excesses appeared to substantially under-estimate these return levels when the process was characterised by strong temporal dependence (and slightly over-estimate at weaker levels of dependence).

In section A.2 we obtain an expression for the distribution of cluster peak excesses at sub-asymptotic levels. This result assumes independence between consecutive observations; of course, in practice, if we knew that successive observations *were* independent, there would be no need to decluster and the GPD could be used to model all excesses. However, any results which shed some light on the effects of declustering on parameter estimation in the independence case might lead to an understanding of where the declustering procedure fails at other levels of dependence. To simplify the calculations, we define an observation to be a *cluster peak* above a threshold u if, and only if, the observation exceeds u and is larger than the κ observations either side of it. We call this method of identifying cluster peaks “top-down” declustering. It is found that, under this declustering regime, the distribution of cluster peak excesses is given exactly by

$$\begin{aligned} & \Pr(X_i > u + x | X_i \text{ is a cluster peak above } u) \\ &= \frac{\lambda_{u+x}}{\lambda_u} \sum_{T=1}^{2\kappa+1} T \left\{ \sum_{n=0}^{T-1} \frac{\binom{T-1}{n}}{n+1} \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^{T-n-1} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^n \right\} \binom{2\kappa}{T-1} \lambda_u^{T-1} (1 - \lambda_u)^{2\kappa+1-T}, \end{aligned} \quad (3.35)$$

where $\lambda_{u+x} < \lambda_u$ and $\frac{\lambda_{u+x}}{\lambda_u}$ is the distribution of a randomly chosen threshold exceedance given by the generalised Pareto distribution. For full details, see section A.2.

Figure 3.10: Sampling distributions of $\hat{q}_{1000}$ for all excesses (heavy line), cluster peak excesses found using runs declustering (thin line) and cluster peak excesses found using “top-down” declustering (dashed line), for $\alpha = 0.2$. The vertical line represent the actual 1000-observation return level

The rather crude “top-down” declustering scheme also results in under/over-estimated return levels in processes with strong/weak temporal dependence (as did using cluster peak excesses obtained under runs declustering – see section 3.4.5). Figure 3.10 shows sampling distributions for $q_{1000} = 2.342$ (as in studies 1, 2 and 3 of the simulation study), based on using (a) all excesses (bold line), (b) cluster peak excesses obtained using runs declustering (thin line) and (c) cluster peak excesses obtained using “top-down” declustering (dashed line), for a simulated dataset with logistic dependence parameter $\alpha = 0.2$ characterising the temporal dependence between successive pairs of observations. As with runs declustering, using cluster peak excesses obtained under “top-down” declustering results in under-estimated return levels. Estimated return levels using cluster peak excesses obtained under runs and “top-down” declustering schemes were found to be similar for series displaying other strengths of serial correlation. Thus, it appears that the under/over-estimation of return levels is not an artefact of the type of declustering scheme used (though other schemes, such as ‘blocks’ declustering (see section 5.2.1), have not been considered here).

The expression in (3.35) shows how the distributions of cluster peak excesses and all threshold excesses differ when cluster peaks are identified using “top-down” declustering, and successive observations in the series are independent. Figure 3.11 shows the

Figure 3.11: Solid line: Probabilities of exceeding $u + x$ for cluster peak excesses, obtained from equation (3.35). Dashed line: Corresponding empirical probabilities obtained from simulated data

distribution of cluster peak excesses identified in this way for various declustering parameters κ (calculated from equation (3.35)). This ‘theoretical’ distribution is shown by the solid curve in the plot. Also shown is the empirical distribution of excess for cluster peaks obtained from simulated data. Independent data were generated, and “top–down” declustering was used to identify a set of cluster peaks; the probability that an observation exceeds the threshold $u + x$ given that this observation is a cluster peak was estimated as the proportion of cluster peaks which exceeded $u + x$. The thresholds u and $u + x$ were set at the 95% and 99% quantiles (respectively). We can see from figure 3.11 that the model for cluster peak excesses given by the expression in (3.35) corresponds closely to the distribution of cluster peak excesses found empirically, supporting the validity of (3.35).

Figure 3.11 shows the distribution of cluster peak excesses obtained from a set of independent threshold exceedances; obviously, if we knew we had independent exceedances, concern about the effects of declustering on parameter estimation would not be an issue since we could model all threshold excesses anyway. However, it is at the level of independence where we might most easily be able to determine (and explain) the effects of declustering on parameter estimation. Recall from the plots in figure 3.6 that, in analyses using cluster peaks obtained from independent threshold exceedances, return levels were *over-estimated*. Here, the 1000-observation return level q_{1000} was estimated through equation (2.10), replacing σ and ξ with the maximum likelihood estimates $\hat{\sigma}_p^*$ and $\hat{\xi}_p$ (the subscript p denoting the use of cluster peaks), and λ_u with the threshold exceedance rate. Estimation may be improved by assuming the model in (3.35) instead of the generalised Pareto distribution for all threshold excesses (as given in (2.8)). Now, just as equation (2.7) could be rewritten as equation (2.8), the expression in (3.35) can be written as

$$\begin{aligned} & \Pr(X_i > u + x) \\ &= \Pr(X_i \text{ is a cluster peak above } u) \sum_{T=1}^{2\kappa+1} T \left\{ \sum_{n=0}^{T-1} \frac{\binom{T-1}{n}}{n+1} \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^{T-n-1} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^n \right\} \\ & \quad \times \binom{2\kappa}{T-1} \lambda_u^{T-1} (1 - \lambda_u)^{2\kappa+1-T}, \end{aligned} \quad (3.36)$$

where

$$\frac{\lambda_{u+x}}{\lambda_u} = S(x_i; \sigma^*, \xi) \quad (3.37)$$

$$= \left[1 + \frac{\xi x_i}{\sigma^*} \right]_+^{-1/\xi}. \quad (3.38)$$

By setting (3.36) equal to $1/t$, replacing x_i with q_t and solving for q_t , we should achieve more accurate estimates of the t -observation return level when using cluster peak excesses; this is because we are using the correct distribution for cluster peak excesses (equation (3.35) instead of (2.7)). The probability $\Pr(X_i \text{ is a cluster peak above } u)$ is estimated as the sample proportion of cluster peaks in the series.

Table 3.5 shows estimates of the 1000-observation return level, using cluster peak excesses, obtained via numerical solution of (3.36). Again, simulated data were used where the *true* 1000-observation return level was known; here, as in the left-hand-side of table 3.2, $q_{1000} = 2.342$. Also shown, for comparison, are the corresponding estimates found through equation (2.10), though these results are different to those in table 3.2 because here we have used “top-down” declustering. Equation (3.36) has also been used to obtain

	$q_{1000} = 2.342$			
	$\alpha = 1$	$\alpha = 0.8$	$\alpha = 0.5$	$\alpha = 0.2$
Original approach	2.370	2.352	2.345	2.101
New approach	2.343	2.343	2.341	2.298

Table 3.5: Estimates of 1000-observation return level using cluster peak excesses based on (a) original approach and (b) new approach, using the model for cluster peak excesses given in equation (3.35)

return level estimates for series which display ‘weak’ ($\alpha = 0.8$), ‘moderate’ ($\alpha = 0.5$) and ‘strong’ ($\alpha = 0.2$) temporal dependence, though incorrectly since this result is valid for independent processes only. From table 3.5, we can see that estimation of q_{1000} has been improved by using the model for cluster peak excesses given in expression (3.35). This was expected in the independent case, since here model (3.35) is valid. However, even when $\alpha = 0.2$, though estimates of q_{1000} are still lower than the true value, this under-estimation is less pronounced than under the original estimation procedure.

Though the results in table 3.5 and the distribution for cluster peak excesses derived in section A.2 go some way to helping us understand the effects of declustering on return level inference, the area lends itself to further research. Results analogous to (3.35) for *dependent* series might help to explain the under-estimation of return levels in series which display strong serial correlation (as are often found in practice), as might an extension of these results to cover the more commonly used “runs” declustering.

3.7 Discussion

This chapter has investigated the effects of declustering on parameter estimation, with particular attention given to return level inference. Chapter 2 highlighted the differences between ignoring serial correlation and filtering out such dependence as far as GPD parameter estimation and return level estimation were concerned; the principal aim of this chapter was to follow-up these results and determine which of the two approaches leads to the ‘best’ estimates. Datasets were generated with *known* marginal GPD parameters and return levels, and maximum likelihood was used to estimate these through the modelling of all threshold excesses (ignoring any dependence) and cluster peak excesses (filtering the series to obtain a set of independent excesses). The strength of serial correlation was determined by the logistic dependence parameter, which was used as a measure of dependence between consecutive pairs of the process. The results of the simulation study were at the very least informative, and could have far-reaching consequences in terms of

the methodology which statisticians recommend to practitioners in the analysis of environmental extremes.

The simulation study (full results given in appendix A.1) revealed systematic bias in estimates of the GPD scale and shape parameters obtained from the cluster peaks analyses; corresponding analyses using all threshold exceedances yielded far greater accuracy. The GPD scale parameter σ^* was systematically over-estimated, and the shape parameter ξ under-estimated, in the cluster peaks analyses. Often, the 95% confidence intervals for parameters estimated using cluster peak excesses – and sometimes the 99% intervals – failed to include the true parameter values; the models using all threshold excesses were more ‘on target’, the corresponding intervals *always* containing the true values. As well as providing more accurate estimates of the GPD parameters, using all threshold excesses also resulted in greater precision, with the inclusion of, in the case of ‘strong’ temporal dependence, up to five times as many extremes in the analysis. However, of far greater practical consequence are the effects of declustering on *return level estimation*. The simulation study revealed that, at levels of dependence often encountered in real-life data, the cluster peaks analyses were constantly under-estimating return levels relative to the analyses making use of all threshold excesses. This suggests that designing to the maximum likelihood estimates which use cluster peak excesses – an approach currently employed by most practitioners – would result in considerable underprotection. However, further investigation has also revealed that this under-estimation becomes even more pronounced as the strength of dependence is increased beyond $\alpha = 0.2$, and that, at weaker strengths of temporal dependence (in particular for α greater than about 0.4), the cluster peaks analyses actually *over-estimate* return levels.

Various follow-up studies have been undertaken in an attempt to determine the cause of these detrimental effects on parameter estimation in the cluster peaks analyses. The first of these considered using different declustering parameters in the runs declustering scheme used to identify sets of clusters; the choice of cluster termination interval was shown to have little effect on the under-estimation of return levels in highly correlated series. We investigated the issue of making arbitrary choices of declustering parameter still further by considering an “automatic” declustering scheme, proposed by Ferro and Segers (2003). This scheme makes use of knowledge of the temporal dependence of the process to ‘optimise’ the choice of runs declustering parameter. Again, return level estimates were still under-estimated in processes characterised by strong serial dependence. Other declustering schemes involving the use of different cluster inhabitants in the analysis were also investigated, some of which exacerbated the under-estimation of return levels, and

none of which improved the situation for cluster peaks analyses.

Finally, a distribution for cluster peak excesses for finite samples was derived, and was shown to be different to the distribution of all threshold excesses. This theoretical model, which is valid for cluster peak excesses obtained from independent threshold exceedances identified through “top–down” declustering, was supported by empirical estimates of excess using simulated data. Further, this model was used to obtain estimates of return levels – not only for independent data, but for series displaying different strengths of temporal dependence. Though the model is incorrect for temporally dependent series, estimates of the 1000–observation return level were shown to be less biased than those obtained under the standard approach *no matter what the strength of serial correlation*. As expected, using the correct model for cluster peak excesses for independent data ($\alpha = 1$) resulted in accurate estimates of the 1000–observation return level (as opposed to over–estimated return levels using the original procedure). However, return level estimation was improved at various levels of temporal dependence; even when $\alpha = 0.2$, though return level estimates *were* under–estimated using our model, the degree of under–estimation was less than that found under the original approach.

In this chapter and chapter 2, we have considered two approaches which deal with modelling extremes which are serially correlated: ignoring temporal dependence and fitting to the entire set of threshold excesses, and filtering out the temporal dependence using some declustering scheme (and fitting to the set of (hopefully) independent cluster peak excesses). We have seen that, by ignoring the serial correlation and fitting to all extremes, the accuracy and precision of our analysis is greatly improved; however, as chapter 2 discusses, standard errors attached to each maximum likelihood estimate are bound to be under–estimated, and should be inflated accordingly. In the following chapter, we investigate the potential merits of explicitly *modelling* the temporal dependence of our process through Markov chain models.

Chapter 4

Markov chain models for threshold exceedances

4.1 Introduction

Chapters 2 and 3 investigated the effects of both *removing* dependence (by declustering and fitting a GPD to the set of filtered cluster excesses) and *ignoring* dependence (simply fitting a GPD to the set of all threshold excesses) on parameter estimation and return level inference. The first approach mentioned here is by far the most popular, and is justified by the asymptotic theory which leads to the generalised Pareto distribution as being the limiting distribution of threshold excesses. The declustering is an *ad hoc* device to circumvent the difficulty that the joint distribution of successive threshold excesses is unspecified by the general theory. However, the disadvantages of this approach are numerous, and have been discussed in detail in the previous two chapters. Firstly, cluster identification is often arbitrary; secondly, information on extremes is discarded, and the opportunity for modelling within-cluster behaviour is lost. More importantly, as chapter 3 demonstrated, return levels can be under-estimated when analyses use the set of extracted cluster maxima only. Using all threshold excesses seems to provide more accurate estimates of parameters and return levels, but by assuming independence we are not exploiting the data fully. This chapter considers an approach which explicitly *models* the serial dependence present in many environmental time series, in the hope of capturing and exploiting as much information as possible from our set of threshold exceedances.

4.2 Markov chain models

The technique of declustering our series prior to analysis can be avoided if we are prepared to make stronger assumptions about the process under study. Other than independence,

(possibly) the simplest assumption that can be made is that the series $X_1, X_2, \dots$ forms a stationary first-order Markov chain (as used in the simulation study; see section 3.2.1). The stochastic properties of such a chain are completely determined by the joint distribution of consecutive pairs. This is because the joint density factorises as

$$f(x_1, \dots, x_n) = f(x_1) \times f(x_2|x_1) \times \dots \times f(x_n|x_{n-1}) \quad (4.1)$$

due to the Markov property. Given a model $f(x_i, x_{i+1}; \psi)$, where ψ is a generic parameter, it follows that the likelihood for ψ is

$$L(\psi) = f(x_1; \psi) \prod_{i=1}^{n-1} f(x_{i+1}|x_i; \psi) \quad (4.2)$$

$$= f(x_1; \psi) \frac{\prod_{i=1}^{n-1} f(x_i, x_{i+1}; \psi)}{\prod_{i=1}^{n-1} f(x_i; \psi)}. \quad (4.3)$$

To model threshold excesses, we can invoke bivariate extreme value considerations to model contributions to the numerator in (4.3); the denominator is simply replaced with the corresponding univariate GPD densities.

4.2.1 Bivariate threshold excess model

Recall that, with distribution function

$$G(x; \sigma, \xi) = 1 - \lambda_u \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]_+^{-1/\xi}, \quad (4.4)$$

the generalised Pareto distribution can be used to model the tail of an arbitrary distribution function F (see section 1.3). This means that there are parameters λ_u , σ and ξ such that, for large enough thresholds u , $F(x) \approx G(x)$, $x > u$. The aim now is to obtain a bivariate joint distribution $F(x, y)$ on regions of the form $x > u_x, y > u_y$, for large enough u_x and u_y .

Suppose $(x_1, y_1), \dots, (x_n, y_n)$ are independent realisations of a random variable (X, Y) with joint distribution function F . For suitable thresholds u_x and u_y the marginal distributions of F each have an approximation of the form (4.4), with respective parameter sets $(\lambda_{u_x}, \sigma_x, \xi_x)$ and $(\lambda_{u_y}, \sigma_y, \xi_y)$. The transformations

$$\tilde{X} = - \left(\log \left\{ 1 - \lambda_{u_x} \left[1 + \frac{\xi_x (X - u_x)}{\sigma_x} \right]^{-1/\xi_x} \right\} \right)^{-1} \quad (4.5)$$

and

$$\tilde{Y} = - \left(\log \left\{ 1 - \lambda_{u_y} \left[1 + \frac{\xi_y(Y - u_y)}{\sigma_y} \right]^{-1/\xi_y} \right\} \right)^{-1} \quad (4.6)$$

induce a variable $(\tilde{X}, \tilde{Y})$ whose distribution function $\tilde{F}$ has margins that are approximately standard Fréchet for $X > u_x$ and $Y > u_y$. Equation (3.4) gives the form of the (joint) survivor function for a bivariate extreme value distribution with unit exponential margins; it can also be shown (Theorem 8.1 in Coles (2001)) that the joint distribution function $G(x, y)$ for a bivariate extreme value distribution with standard Fréchet margins has the representation

$$G(x, y) = \exp \{ -V(x, y) \}, \quad (4.7)$$

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

$$V(x, y) = 2 \int_0^1 \max \left(\frac{q}{x}, \frac{1-q}{y} \right) dH(q), \quad (4.8)$$

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

$$\int_0^1 q dH(q) = \frac{1}{2}. \quad (4.9)$$

By equation (4.7), for large n ,

$$\tilde{F}(\tilde{x}, \tilde{y}) = \left\{ \tilde{F}^n(\tilde{x}, \tilde{y}) \right\}^{1/n} \quad (4.10)$$

$$\approx \left[\exp \left\{ -V \left(\frac{\tilde{x}}{n}, \frac{\tilde{y}}{n} \right) \right\} \right]^{1/n} \quad (4.11)$$

$$= \exp \{ -V(\tilde{x}, \tilde{y}) \}, \quad (4.12)$$

because of the homogeneity property of V . Finally, since $F(x, y) = \tilde{F}(\tilde{x}, \tilde{y})$, it follows that

$$F(x, y) \approx G(x, y) \quad (4.13)$$

$$\approx \exp \{ -V(\tilde{x}, \tilde{y}) \}, \quad (4.14)$$

for $x > u_x$ and $y > u_y$.

The logistic family

One standard class of parametric families for G is the logistic family, whose (joint) survival function is given in equation (3.21). The corresponding distribution function is given by

$$G(x, y; \alpha) = \exp \left\{ - \left(x^{-\frac{1}{\alpha}} + y^{-\frac{1}{\alpha}} \right)^\alpha \right\}, \quad (4.15)$$

for $x > 0, y > 0$ and $\alpha \in (0, 1]$. Thus, contributions to the numerator in (4.3) can be found by differentiation of (4.15) with respect to both x and y , if $x > u_x$ and $y > u_y$. However, inference for this model is complicated by the fact that a bivariate pair may exceed a specified threshold in just one of its components. Let

$$R_{0,0} = (-\infty, u_x) \times (-\infty, u_y), \quad (4.16)$$

$$R_{1,0} = [u_x, \infty) \times (-\infty, u_y), \quad (4.17)$$

$$R_{0,1} = (-\infty, u_x) \times [u_y, \infty) \quad \text{and} \quad (4.18)$$

$$R_{1,1} = [u_x, \infty) \times [u_y, \infty), \quad (4.19)$$

so that, for example, a point $(x, y) \in R_{0,1}$ if the y -component exceeds the threshold u_y , but the x -component is below u_x . For points in $R_{1,1}$, model (4.15) applies, the density of which gives the appropriate likelihood component. Elsewhere, it is necessary to censor the likelihood component. Thus, the likelihood contributions to the numerator in (4.3) become

$$f(x, y; \psi) = \begin{cases} \left. \frac{\partial^2 F}{\partial x \partial y} \right|_{(x,y)} & \text{if } (x, y) \in R_{1,1}, \\ \left. \frac{\partial F}{\partial x} \right|_{(x,u_y)} & \text{if } (x, y) \in R_{1,0}, \\ \left. \frac{\partial F}{\partial y} \right|_{(u_x,y)} & \text{if } (x, y) \in R_{0,1}, \\ F(u_x, u_y) & \text{if } (x, y) \in R_{0,0}, \end{cases} \quad (4.20)$$

where $x_i = x$ and $x_{i+1} = y$.

4.2.2 Modelling

Adopting the logistic model given in equation (4.15) for the distribution of consecutive pairs of observations, the derivatives in (4.20) become:

$$\begin{aligned} \left. \frac{\partial^2 F}{\partial x \partial y} \right|_{(x,y)} &= (xy)^{-(1/\alpha+1)} (x^{-1/\alpha} + y^{-1/\alpha})^{\alpha-2} \left[(x^{-1/\alpha} + y^{-1/\alpha})^\alpha - (1 - 1/\alpha) \right] \\ &\quad \times \exp \left\{ - (x^{-1/\alpha} + y^{-1/\alpha})^\alpha \right\}, \end{aligned} \quad (4.21)$$

$$\left. \frac{\partial F}{\partial x} \right|_{(x,u_y)} = x^{-(1/\alpha+1)} (x^{-1/\alpha} + y^{-1/\alpha})^{\alpha-1} \exp \left\{ - (x^{-1/\alpha} + y^{-1/\alpha})^\alpha \right\}, \quad (4.22)$$

$$\left. \frac{\partial F}{\partial y} \right|_{(u_x,y)} = y^{-(1/\alpha+1)} (x^{-1/\alpha} + y^{-1/\alpha})^{\alpha-1} \exp \left\{ - (x^{-1/\alpha} + y^{-1/\alpha})^\alpha \right\}, \quad (4.23)$$

$$F(u_x, u_y) = \exp \left\{ - (u_x^{-1/\alpha} + u_y^{-1/\alpha})^\alpha \right\}. \quad (4.24)$$

Replacing x and y with x_i and x_{i+1} (respectively) in (4.21) – (4.24) gives the required contributions to the numerator in (4.3) for the Markov chain model. Also, since u_x and

u_y are replaced with u_{x_i} and $u_{x_{i+1}}$, and for $i = 1, \dots, n-1$ we have that $u_{x_i} = u_{x_{i+1}} = u$, where u is the chosen threshold for our data, (4.24) becomes

$$F(u_x, u_y) = \exp \{-2^\alpha/u\}. \quad (4.25)$$

Maximisation of the log-likelihood leads to estimates and standard errors for the logistic dependence in the usual way, with a separate model being fitted to each season where seasonal variation is apparent (taking care to account for any discontinuities in the data; see section 4.3.1). The inference can be simplified by carrying out the marginal estimation, then transforming the data using (4.5), where σ_x , ξ_x and λ_{u_x} are replaced with σ_m , ξ_m and λ_{u_m} (respectively) to indicate the season to which the estimates belong. In this case, likelihood (4.3) is a function only of the dependence parameter contained in the model. Using the transformed series, the univariate densities in the denominator of (4.3) are now GPD if $X > u$; elsewhere, we use $F(u)$.

4.3 Implementation

4.3.1 Bradfield wind speed data

Having investigated the effects of both *ignoring* and *removing* temporal dependence on return level inference for the Bradfield wind speed data (chapter 2), we now use the (first-order) Markov chain model described in section 4.2 to explicitly *model* the dependence in the series. The method for selecting appropriate monthly thresholds was discussed in detail in section 2.5; maximum likelihood is used to obtain estimates for the dependence parameter α_m , as well as the marginal GPD scale and shape parameters σ_m and ξ_m , $m = 1, \dots, 12$. In each month, for all but nine pairs of observations, likelihood contributions to the Markov chain model are given by derivatives (4.21) – (4.24). However, for each month of data 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. Since we have ten years of data, each month of data will have nine such discontinuities, and for each of these nine pairs we assume independence and assign a value of 1 to α in (4.21) – (4.24).

From the MLEs for σ_m , ξ_m and α_m , we can obtain estimates of return levels for the 10, 50, 200 and 1000-year return periods (as before), and compare these to the estimates based on using all threshold excesses and cluster peak excesses. Table 4.1 shows estimates of the logistic dependence parameter α for each month, along with asymptotic standard

Month (m)	$\hat{\sigma}_m$	$\hat{\xi}_m$	$\hat{\alpha}_m$
1	8.149 (0.633)	-0.102 (0.055)	0.377 (0.014)
2	8.145 (0.673)	-0.104 (0.056)	0.396 (0.016)
3	9.009 (0.691)	-0.359 (0.050)	0.463 (0.017)
4	8.413 (0.655)	-0.271 (0.052)	0.383 (0.015)
5	6.870 (0.593)	-0.120 (0.062)	0.368 (0.015)
6	3.959 (0.358)	0.058 (0.069)	0.477 (0.017)
7	7.519 (0.571)	-0.414 (0.049)	0.441 (0.017)
8	6.972 (0.613)	-0.100 (0.063)	0.367 (0.016)
9	9.898 (0.400)	-0.137 (0.024)	0.324 (0.008)
10	7.758 (0.642)	-0.231 (0.060)	0.460 (0.016)
11	7.385 (0.583)	-0.061 (0.050)	0.450 (0.017)
12	8.411 (0.750)	-0.121 (0.068)	0.394 (0.016)

Table 4.1: Maximum likelihood estimation of the logistic dependence parameter α (and marginal GPD parameters σ and ξ) using the logistic model

errors (in parentheses). Also shown are estimates of the marginal GPD parameters σ and ξ , based on using all threshold excesses. We see that all months display medium to strong temporal dependence, with estimates of α ranging from 0.324 to 0.477; there appears to be no distinction between winter and summer months here (indeed, similarity of estimates across seasons might be indicative of a simpler model which fits a constant α for all seasons). Of course, the increased complexity (and computational expense) of modelling the 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 – or remove – the serial correlation observed in the series. To this end, we now focus on return level inference.

Return level estimation

As before, let $X_1, X_2, \dots, X_n$ form a stationary sequence, with marginal distribution function F . Standard arguments in Leadbetter, Lindgren and Rootzén (1983, Ch. 3) show that, for large n and u , it is typically the case that

$$\Pr \{\max(X_1, \dots, X_n) \leq u\} \approx \{F(u)\}^{n^\theta}, \quad (4.26)$$

where θ is the extremal index. Using the same notation as before, the value q_r , which equates (4.26) to $1 - 1/r$, is the r -year return level of the process. Again, since we have monthly varying parameters, q_r is found as the solution of

$$\sum_{m=1}^{12} \left[1 - \lambda_{u_m} \left[1 + \xi_m \left(\frac{q_r - u_m}{\sigma_m} \right) \right]_+^{-1/\xi_m} \right]^{h_m \theta_m} = 12 - \frac{1}{r}, \quad (4.27)$$

	Return period (years)			
	10	50	200	1000
Markov model	96.745 (2.864)	103.236 (5.930)	108.152 (8.786)	113.306 (12.219)
All exceedances	96.785 (3.534)	103.180 (6.045)	108.085 (10.702)	113.249 (15.678)
Cluster peaks	94.916 (3.970)	100.469 (15.971)	104.950 (23.793)	114.954 (44.865)

Table 4.2: Return level estimates with associated standard errors

where h_m is the number of hours in month m . We substitute σ_m , ξ_m and λ_{u_m} by their maximum likelihood estimates; the extremal index θ_m , which is implicitly defined through α_m , ensures we incorporate the temporal dependence of the series into the estimation of q_r . Smith (1992) gives *calculated* values of θ for specific values of α (in particular $\alpha = 0.5$, 0.333, 0.25 and 0.2 – see table 5.4), but the technique used is far from automatic and computationally very expensive. Thus, the scheme outlined below demonstrates how θ_m can be *estimated*.

1. For each month m , $m = 1, \dots, 12$, use $\hat{\alpha}_m$ to simulate a first-order Markov chain using the logistic family to model the distribution of successive pairs of realisations;
2. Use the method of Ferro and Segers to estimate the extremal index θ_m (equation 3.34), avoiding the need to decluster the series;
3. repeat steps 1 and 2 N times and use the sample mean of $\hat{\theta}_m$ to replace θ_m in equation (4.27).

We use $N = 1000$ repetitions of the scheme outlined above to estimate θ . Table 4.2 displays estimates of the 10, 50, 200 and 1000-year return levels, with standard errors in parentheses. Estimates based on removing dependence (cluster peaks), and ignoring dependence (all exceedances), are shown again here for comparison. As can be seen, estimates provided by the two methods which use all excesses seem to give fairly consistent results, with the Markov model giving greater precision. In most cases, the cluster maxima model under-estimates relative to the other two models, as well as giving far less precision to the parameter estimates (with estimates having much larger standard errors). In the simulation study of chapter 3 we found that using all excesses rather than cluster peak excesses seems to provide more accurate and efficient return level estimates; here, we see the Markov model appears to improve the situation still further, lending greater precision to our analysis. With this, and the fact that we can quantify the strength of dependence in the series through the estimation of the logistic dependence parameter α , we believe that the parametric Markov chain model provides considerable benefits over the cluster peaks approach.

4.3.2 Simulated data

Here, we assess our findings of section 4.3.1 by fitting the first-order Markov logistic model to simulated data. Data are simulated in accordance with section 3.2, using the bivariate logistic model for consecutive pairs in the series. As in the simulation study of chapter 3, for a range of ξ values, and for three levels of dependence ($\alpha = 0.8$, $\alpha = 0.5$ and $\alpha = 0.2$), the Markov chain model is fitted 1000 times to obtain sampling distributions for σ , ξ and α . Return levels are then estimated by equating (4.26) to $1 - 1/r$ and solving analytically, since the simulated chains are stationary and have no distinct seasonal components. Thus, the r -year return level q_r is estimated using

$$\hat{q}_r = u - \hat{\sigma} \hat{\xi}^{-1} \left(1 - \left[\hat{\lambda}_u^{-1} \left\{ 1 - (1 - 1/r)^{1/(n\hat{\theta})} \right\} \right]^{-\hat{\xi}} \right), \quad (4.28)$$

where θ is estimated using Ferro and Segers' method. The sampling distributions of the 1000-observation return level for the three strengths of dependence are shown in figure 4.1, along with the sampling distributions for the dependence parameters themselves. Here, only the results for $\xi = -0.4$ are shown, though similar results were obtained for the other values used (-0.1 , 0.0 , 0.3 and 0.8). From figure 4.1, we see that, as with the Bradfield wind speed data, the Markov chain model seems to give estimates of q_{1000} which are just as accurate as the all excesses model, with the cluster peaks approach under-estimating q_r when the series displays strong temporal dependence. Further, the sampling distributions are narrower for the Markov model, supporting the results from the Bradfield analysis where the standard errors attached to all return level estimates were smaller under this approach. The vertical lines, which represent the true parameter values, also show that the dependence parameter α has been adequately estimated.

4.4 Further considerations

4.4.1 Different choices of bivariate extreme value limit

The main reason for the popularity of the logistic model is its flexibility, with all levels of dependence – complete independence to perfect dependence – covered by the dependence parameter α . The derivation of the joint distribution function for the bivariate logistic model (given in equation (4.15)) from equation (4.8) is not obvious, but it can be shown that (4.15) can be obtained by letting H have the density function

$$h(q) = \frac{1}{2}(\alpha^{-1} - 1) \{q(1 - q)\}^{-1-1/\alpha} \{q^{-1/\alpha} + (1 - q)^{-1/\alpha}\}^{\alpha-2} \quad (4.29)$$

on $0 < q < 1$ (Coles, 2001). The mean constraint (4.9) is automatically satisfied for this model because of the symmetry about $w = 0.5$. However, one limitation of the logistic

Figure 4.1: Sampling distributions of $\hat{\alpha}$ and $\hat{q}_{1000}$ using all threshold excesses (heavy line), cluster peak excesses (thin line), and the Markov model (dashed line)

model due to the symmetry of h is the fact that the variables x and y in (4.15) 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.

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

$$h(q) = \frac{1}{2}(1-\alpha)(1-q)^{-1}q^{-2}(1-u)u^{1-\alpha}\{\alpha(1-u)+\beta u\}^{-1} \quad (4.30)$$

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

$$(1-\alpha)(1-q)(1-u)^\beta - (1-\beta)qu^\alpha = 0. \quad (4.31)$$

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

The Dirichlet model

An alternative (asymmetric) model proposed by Coles and Tawn (1991) is the Dirichlet model¹, where

$$h(q) = \frac{\alpha\beta\Gamma(\alpha+\beta+1)(\alpha q)^{\alpha-1}\{\beta(1-q)\}^{\beta-1}}{2\Gamma(\alpha)\Gamma(\beta)\{\alpha q + \beta(1-q)\}^{\alpha+\beta+1}} \quad (4.32)$$

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

4.4.2 Higher-order models

Extensions of the first-order Markov model to a general d -th order Markov chain are straightforward. The right-hand-side of (4.3) is simply replaced by

$$\frac{\prod_{i=1}^{n-d} f_d(x_i, \dots, x_{i+d-1}; \psi)}{\prod_{i=d}^{n-d} f_{d-1}(x_i, \dots, x_{i+d-2}; \psi)}, \quad (4.33)$$

¹Coles and Tawn use the name Dirichlet since the model is developed by transformation of the standard Dirichlet family of distributions

where f_i is the joint density of i consecutive observations. As previously, both the numerator and denominator may be approximated by multiplying together contributions similar in form to (4.20). In common with standard time series modelling, selection of the order, d , is difficult. Testing between orders $d - 1$ and d is equivalent to testing for conditional independence of variables within a unit simplex domain. Standard likelihood procedures can be used, though models typically will be non-nested. A combination of the usual time series procedures, such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) (Box *et al.*, 1994), and informal model goodness-of-fit assessments based on comparisons of predicted and observed functionals of interest, can be used (see section 4.5). The same procedures may also be used to select between different dependence models of a given order.

4.5 Suitability of the (first-order) Markov assumption at Bradfield

The first-order Markov chain model seems to perform well in terms of return level inference, increasing precision and allowing the quantification of dependence through the dependence parameter α . However, just as fitting the GPD to all threshold excesses requires that successive observations are independent, the fitting of the first-order Markov chain model is based on the assumption that extremes of the process are indeed first-order Markov. Though this is a much more plausible assumption than independence, it may be that a higher-order Markov chain model would better describe the behaviour of the process. The suitability of a d th-order Markov model relative to a $(d + 1)$ th-order model can be assessed via formal likelihood ratio tests, and via informal assessments based on comparisons of predicted and observed functionals of interest; this section looks at both. We begin by looking for conditional independence between observations at lag 2 within extreme events, before comparing model-based and empirical estimates of various quantities of interest for the first-order Markov model.

4.5.1 Conditional independence between observations at lag 2

We first 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. A simple *ad hoc* procedure is to look for serial correlation in the differences of the successive random walk steps during extreme events. Defining an extreme event as a cluster of extremes (clusters being identified using Ferro and Segers' method for identifying clusters), and using a simple peaks and troughs test to detect any non-

randomness between observation differences within clusters, we found there to be significant serial correlation between observations of lag 2 in most extreme events, suggesting that a higher-order Markov model might be more appropriate.

A more rigorous approach is to consider the trivariate distribution of consecutive triples of extremes. Creating a series $\tilde{X}_i$ by transforming the data (using equation 4.5) 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 & 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. Also, a plot of $W_{i,1}$ against $W_{i,3}$ can help determine the appropriateness of the first-order Markov model. 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), though this doesn't necessarily imply the first-order Markov assumption will completely fail. To illustrate the use of such a "simplex plot", we first consider some simulated data (data being simulated as in the study of chapter 3, i.e. first-order Markov with bivariate logistic transitionals). Figure 4.2 shows four such plots, each corresponding to a particular strength of dependence, ranging from very weak to very strong ($\alpha = 0.99$, $\alpha = 0.7$, $\alpha = 0.4$ and $\alpha = 0.1$). Though all the simulated chains are first-order Markov, we might still expect this dependence to induce serial correlation at lag 2 and beyond, especially as the strength of dependence increases. Indeed, this is illustrated in figure 4.2; in all four plots, there is a cluster of points on the interior, though this is more pronounced as α decreases (the first plot *does* shows many points tending towards the edges). This suggests that, though serial correlation at lag 2 has to some extent been induced even when $\alpha = 0.99$, the second-order Markov assumption becomes increasingly plausible as α decreases.

Figure 4.3 shows two simplex plots for the Bradfield data, one for the Markov chain model fit for January and one for July, though similar plots were obtained for all twelve months. In both we see a tendency for points to lie on the interior of the plot, suggesting that although the first-order Markov assumption might be adequate, a second-order Markov chain model might provide a better fit to the monthly data (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). Our conclusion is that, though the first-order Markov assumption might not fail, a second-order Markov chain model which fits to consecutive triples of extremes might provide

Figure 4.2: Simplex plots of $W_{i,i}$ versus $W_{i,3}$ for decreasing α (increasing (first-order) dependence)

Figure 4.3: Simplex plots of $W_{i,i}$ versus $W_{i,3}$ for the Bradfield data: January and July

considerable improvement upon the first-order fit. Further examination of the first-order assumption is given below.

4.5.2 The extremal index

Section 4.3.1 outlines a scheme for estimating the extremal index, θ_m , based on simulating first-order Markov chains with the logistic model for the joint distribution of consecutive pairs; this was a necessary step in the estimation of the r -year return level q_r , found by solving equation (4.27). However, we can use this simulation scheme to find model-based estimates of θ_m (for a range of thresholds), and assess the first-order Markov assumption by comparing these to threshold-dependent empirical estimates obtained as the reciprocal of the sample mean cluster size (this requires clusters to be subjectively identified; also, the theory states that θ^{-1} is the *limiting* mean cluster size, where clusters are identified as sets of consecutive threshold exceedances (see section 1.4.1)). Figure 4.4 shows estimates of θ_1 plotted against the threshold, u , used in its estimation, with runs declustering (see section 2.4.2) used to identify clusters of extremes; two declustering parameters are used for comparison – $\kappa = 30$ hours and $\kappa = 60$ hours. Since estimation of θ_m has been shown to be sensitive to the subjective choice of declustering parameter (Ancona-Navarrete & Tawn, 2000), Ferro and Segers' intervals estimator (see 3.5.2) is also used to estimate θ_m . The plot suggests an obvious discrepancy between estimates based on the first-order Markov chain model and estimates using the reciprocal of the mean cluster size. Even when the need to make a subjective choice of cluster separation interval is removed, and Ferro and Segers' intervals estimator is employed, we see the first-order Markov chain model still seems to give consistently higher estimates of θ_1 than its empirical counterparts. This suggests that the first-order Markov chain model is not allowing for as much extremal dependence as is present in the data, and so assuming extremes are second-order Markovian might well be more plausible (since stronger extremal dependence would induce longer-range dependence). Figure 4.4 shows little difference between the two methods which use 'runs' declustering and Ferro and Segers' intervals estimator; for a more comprehensive comparison of extremal index estimators, see chapter 5.

4.5.3 The cluster length distribution

A more general summary of cluster behaviour is the distribution, ω , of the number of exceedances per cluster. A comparison of the model-based and empirical estimate of ω (for January) is given in table 4.3; again, similar results were obtained for other months. Figure 4.5 shows density plots of ω , the dashed line showing the model-based distribution. From figure 4.5, both distributions seem reasonably similar; however, when we look at

Figure 4.4: Estimated extremal index for January (θ_1) plotted against the threshold, u , used in its estimation: $\oplus$ – first-order Markov chain model; $\bullet$ – empirical estimates using ‘runs’ declustering, $\kappa = 30$ hours; $\square$ – empirical estimates using ‘runs’ declustering, $\kappa = 60$ hours; ∇ – empirical estimates using the intervals estimator

the actual $\omega(j)$ given in the table, we see a poor agreement in places. For example, $\omega(7) = 0.063$ based on the first-order Markov chain model, compared to an empirical value of just 0.029. Similarly, we have a model-based estimate of $\omega(8) = 0.075$ compared to its empirical counterpart of 0.036. It is also apparent that, in contrast with some analytical examples (Hsing *et al.*, 1988), the estimated distribution ω is not geometric, having a larger $\omega(1)$ and a (much) heavier tail.

4.6 Second-order Markov chain model

Disagreement between model-based and empirical estimates of the extremal index (θ) and (arguably) the cluster length distribution (ω) suggests that the first-order Markov assumption might be inadequate for the Bradfield wind speed data. Indeed, the simplex plots suggest a second-order Markov assumption might be more valid; the correlograms in figure 4.6, which compare the autocorrelation function for a simulated first-order Markov chain with logistic dependence to that of the Bradfield data (for January), also suggest that the data have considerably longer-range dependence than the first-order Markov model allows for, with the correlogram for the simulated data tailing off much more quickly.

j	1	2	3	4	5	6	7	8	9	≥ 10
Model	0.098	0.080	0.075	0.052	0.069	0.040	0.063	0.075	0.057	0.391
Empirical	0.101	0.094	0.065	0.051	0.080	0.051	0.029	0.036	0.029	0.464

Table 4.3: Model-based and empirical estimates of the distribution, ω , of the number of exceedances per cluster

Figure 4.5: Density plots of ω , the cluster length distribution: dashed line – model-based distribution, solid line – empirical distribution

Figure 4.6: Autocorrelation functions for the Bradfield wind speed data (January) (Top) and the first-order Markov chain model (bottom)

As section 4.4.2 demonstrates, extensions of the first-order Markov model to a general d -th order Markov chain are quite straightforward – the d -dimensional analog to the likelihood given by equation (4.3) is shown in equation (4.33). The multivariate representation of the distribution function for the logistic model (whose bivariate distribution function is given in equation (4.15)) is given as

$$G(v_1, \dots, v_d; \alpha) = \exp \left\{ - \left(\sum_{i=1}^d v_i^{-1/\alpha} \right)^\alpha \right\}, \quad (4.34)$$

again for $\alpha \in (0, 1]$. Thus, if we consider a second-order Markov chain model², equation (4.33) becomes

$$L(\alpha) = \frac{\prod_{i=1}^{n-2} f(x_i, x_{i+1}, x_{i+2}; \alpha)}{\prod_{i=2}^{n-2} f(x_i, x_{i+1}; \alpha)}, \quad (4.35)$$

where the joint trivariate and bivariate densities (numerator and denominator respectively) are found by differentiation of (4.34). We should bear in mind, at this point, that the elements $v_1, \dots, v_d$ in equation (4.34) are exchangeable. However, inference is complicated by the fact that a bivariate pair (or trivariate triple) may exceed a specified threshold in just one of its components. Thus, contributions to the denominator are given by derivatives (4.21) – (4.24), where the likelihood component is censored for regions where a component lies below its threshold. Similarly, for contributions to the numerator, if we consider the regions

$$R_{0,0,0} = (-\infty, u_x) \times (-\infty, u_y) \times (-\infty, u_z), \quad (4.36)$$

$$R_{1,0,0} = [u_x, \infty) \times (-\infty, u_y) \times (-\infty, u_z), \quad (4.37)$$

$$R_{0,1,0} = (-\infty, u_x) \times [u_y, \infty) \times (-\infty, u_z), \quad (4.38)$$

$$R_{0,0,1} = (-\infty, u_x) \times (-\infty, u_y) \times [u_z, \infty), \quad (4.39)$$

$$R_{1,1,0} = [u_x, \infty) \times [u_y, \infty) \times (-\infty, u_z), \quad (4.40)$$

$$R_{0,1,1} = (-\infty, u_x) \times [u_y, \infty) \times [u_z, \infty), \quad (4.41)$$

$$R_{1,0,1} = [u_x, \infty) \times (-\infty, u_y) \times [u_z, \infty) \quad \text{and} \quad (4.42)$$

$$R_{1,1,1} = [u_x, \infty) \times [u_y, \infty) \times [u_z, \infty), \quad (4.43)$$

²Historically, the multivariate logisitc model has been used to measure spatial dependence between exchangeable sites, and not in a time series context for Markov chain modelling; the use of this model in this setting might be limited due to the temporal ordering of our data

then

$$f(x, y; \alpha) = \begin{cases} \left. \frac{\partial^3 F}{\partial x \partial y \partial z} \right|_{(x, y, z)} & \text{if } (x, y, z) \in R_{1,1,1}, \\ \left. \frac{\partial^2 F}{\partial x \partial y} \right|_{(x, y, u_z)} & \text{if } (x, y, z) \in R_{1,1,0}, \\ \left. \frac{\partial^2 F}{\partial y \partial z} \right|_{(u_x, y, z)} & \text{if } (x, y, z) \in R_{0,1,1}, \\ \left. \frac{\partial^2 F}{\partial x \partial z} \right|_{(x, u_y, z)} & \text{if } (x, y, z) \in R_{1,0,1}, \\ \left. \frac{\partial F}{\partial x} \right|_{(x, u_y, u_z)} & \text{if } (x, y, z) \in R_{1,0,0}, \\ \left. \frac{\partial F}{\partial y} \right|_{(u_x, y, u_z)} & \text{if } (x, y, z) \in R_{0,1,0}, \\ \left. \frac{\partial F}{\partial z} \right|_{(u_x, u_y, z)} & \text{if } (x, y, z) \in R_{0,0,1}, \\ F(u_x, u_y, u_z) & \text{if } (x, y, z) \in R_{0,0,0}, \end{cases} \quad (4.44)$$

where $x = x_i$, $y = x_{i+1}$ and $z = x_{i+2}$. As before, maximisation of the log-likelihood leads to estimates and standard errors in the usual way, with a separate model being fitted to each season. Again, inference can be simplified by carrying out the marginal estimation, then transforming the data to unit Fréchet, resulting in the likelihood being a function of the dependence parameter α only.

4.6.1 Implementation for the Bradfield data

Table 4.4 shows monthly estimates of α , with standard errors in parentheses. Unlike table 4.1, we do not show marginal estimates of σ_m and ξ_m here, as these are the same as before. As with the first-order Markov chain model, there seems to be no association between strength of dependence and season, and estimates of α range between 0.479 and 0.604. Comparing with estimates of α_m using the first-order model (table 4.1), we see estimates are generally higher for the second-order model. This might be expected; even in strongly dependent series (such as the Bradfield data), we can expect the joint dependence between three consecutive observations to be less than that between a pair of consecutive observations, though this difference would, of course, be more noticeable in data which are less serially correlated. Here, we see that all estimates of α_m indicate moderate dependence, with the weakest level of dependence being $\hat{\alpha}_{10} = 0.604$ (October). We also show (in figure 4.6) a correlogram for a simulated second-order Markov chain, using the estimate of α_1 from table 4.4 to model the dependence between successive triples. We see that this correlogram corresponds more closely to that of the Bradfield data than the correlogram for the first-order model. For each month, model-based estimates of the extremal index and the cluster length distribution for this model were also in closer agreement with their empirical counterparts than were the corresponding estimates for the first-order Markov chain model (see section 4.5). We also found there to be considerably

less serial correlation between observation differences at lag 3 than at lag 2, and a simple peaks and troughs test revealed conditional dependence between observation differences at lag 3 to be insignificant; this suggests there might be little to gain from extending the modelling procedure to a d -th order Markov chain, $d \geq 3$.

4.6.2 Model selection

We now consider different model choices for the joint distribution of consecutive triples (see section 4.4.1). Table 4.5 shows estimates of dependence parameters for three second-order Markov chain models; the first is the (symmetric) logistic model (given by equation 4.34), and the other two correspond to the bilogistic and Dirichlet (or Coles–Tawn) models outlined in section 4.4.1. As before, results are tabulated for January, though similar comparisons were made from fitting to other months. Also shown are the (maximised) log-likelihood values, and the mean number of threshold up-crossings and mean intra-cluster standard deviations (m_u and m_s respectively). Since both the bilogistic and Dirichlet models reduce to the symmetric logistic model when $\alpha = \beta$, standard likelihood ratio tests can be used to detect any model improvement when we allow for asymmetry in the dependence structure. Comparing the logistic and bilogistic models, referring $D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\}$ to χ^2_1 tables we see no significant improvement by including the asymmetry parameter β_1 . Similarly, $D = 2\{\ell_2(\mathcal{M}_2) - \ell_0(\mathcal{M}_0)\}$ is non-significant at any reasonable level. These tests are supported by the similarity between both dependence parameters within the bilogistic and Dirichlet models; recall that, in both models, when $\alpha = \beta$ we are reduced to the (symmetric) logistic model. Also, the similarity of the mean number of threshold up-crossings (m_u) and within-cluster standard deviations (m_s) across models suggests that dependence model selection is a relatively unimportant aspect of the modelling procedure, and catering for asymmetry in the dependence structure is unnecessary.

4.6.3 Return level estimation

To complete the analysis of the second-order Markov chain model, we turn our attention to return level estimation. The procedure is the same as that outlined in section 4.3.1 for the first-order Markov chain model, using estimates of α_m from table 4.4 (using the second-order Markov chain model) instead of those shown in table 4.1. The extremal index for each month m , θ_m , is also estimated in the same fashion, simulating second-order Markov chains using the values in table 4.4 and estimating θ_m using the intervals estimator, to form a sampling distribution for $\hat{\theta}_m$. The mean of this sampling distribution is then used to replace θ_m in equation (4.27). Estimates of q_r for the 10, 50, 200 and 1000-year return

m	1	2	3	4	5	6
$\hat{\alpha}_m$	0.542	0.546	0.602	0.505	0.509	0.597
(s.e.)	(0.021)	(0.022)	(0.023)	(0.021)	(0.021)	(0.021)
m	7	8	9	10	11	12
$\hat{\alpha}_m$	0.570	0.504	0.479	0.604	0.601	0.547
(s.e.)	(0.022)	(0.021)	(0.013)	(0.022)	(0.023)	(0.022)

Table 4.4: Maximum likelihood estimation of the logistic parameter α using the logistic model; second-order Markov chain model

Model	Dependence parameters	Log-likelihood	m_u	m_s
Logistic ($\mathcal{M}_0$)	$\hat{\alpha}_1 = 0.542(0.021)$	$\ell_0(\mathcal{M}_0) = 92036.6$	1.45	0.75
Bilogistic ($\mathcal{M}_1$)	$\hat{\alpha}_1 = 0.621(0.022)$ $\hat{\beta}_1 = 0.614(0.028)$	$\ell_1(\mathcal{M}_1) = 92037.1$	1.53	0.82
Dirichlet ($\mathcal{M}_2$)	$\hat{\alpha}_1 = 0.423(0.030)$ $\hat{\beta}_1 = 0.430(0.026)$	$\ell_2(\mathcal{M}_2) = 92037.6$	1.49	0.79

Table 4.5: Estimates of dependence parameters for three (second-order) Markov chain models, with associated maximised log-likelihoods. m_u and m_s are mean number of threshold up-crossings and mean intra-cluster standard deviations (respectively), clusters defined by Ferro and Segers’ “automatic” method

	Return period (years)			
	10	50	200	1000
First-order model	96.745 (2.864)	103.236 (5.930)	108.152 (8.786)	113.306 (12.219)
Second-order model	96.786 (2.859)	103.283 (5.845)	108.154 (8.709)	113.412 (11.992)

Table 4.6: Return level estimates with associated standard errors

periods are shown in table 4.6, with estimates based on the first-order Markov chain model shown again for comparison. We see that both the first-order and second-order Markov chain models give very similar estimates of return levels, the second-order model having the advantage of greater precision. The precision is only slightly improved, however; consequently, the greater complexity and computational expense (save the advantage of slightly smaller standard errors) has been unnecessary if we use return level estimation as the chief motivating factor for model selection. Though the evidence thus far has pointed towards the second-order Markov chain model as being more appropriate than the first-order model, estimated return levels have hardly changed.

4.7 Discussion

This chapter has shown that if we are prepared to make stronger assumptions about the process we are studying, we can make considerable improvements to return level inference. The results of the simulation study in chapter 3 highlighted the advantages of using all threshold excesses in an extreme value analysis over the more common procedure of using a filtered set of ‘independent’ threshold excesses. These included increased precision due to the inclusion of more extreme data, more accurate estimates of model parameters, and most importantly more accurate estimates of return levels; using cluster peak excesses in processes which displayed strong serial correlation resulted in substantially under-estimated return levels. However, the main drawback of simply using all threshold excesses is, of course, the implicit assumption of independence between successive threshold exceedances. The aim of this chapter was to investigate the possibility of improving return level inference further by assuming short-term temporal dependence, and accounting for this dependence through a first-order Markov chain model. Using both the Bradfield wind speed data and simulated data, the Markov chain model gave return level estimates similar to the approach which ignores dependence and fits to all threshold excesses; however, the estimates under the first-order Markov assumption were more precise than those found under the assumption of complete independence. Return level estimates for the Bradfield wind speed data had smaller asymptotic standard errors when dependence had been properly accounted for in the likelihood, and the sampling distributions of q_{1000} (using simulated data) showed less variability under the Markov chain model than when temporal dependence was simply ignored. Since in any extremal analysis the value of information is especially high, such improvements in efficiency through more structured statistical modelling (which can lead to substantial improvements in any practical application) can justify the increased complexity of such models.

Though the first-order Markov assumption has greater credibility than the assumption of complete independence, higher-order models which allow for longer-range dependence can also be considered. Informal diagnostic checks attempted to assess the suitability of the first-order Markov assumption at Bradfield by constructing simplex plots and comparing model-based and empirical estimates of various quantities of interest. These checks suggested that, though a first-order Markov assumption may be plausible, a second-order Markov chain might capture the temporal dependence inherent in the process more accurately. Indeed, model-based and empirical estimates of the extremal index and the cluster length distribution, for example, were in closer agreement under the second-order model. However, increased model complexity can only be justified if this leads to substantial real improvements in any practical setting; return level estimates under the second-order model hardly changed from the estimates under the first-order model, with only (*very*) slight improvements in efficiency. Thus, in subsequent chapters where Markov chain models are considered, model-fitting is limited to the simpler first-order model. Different models for the joint distribution of consecutive values also showed that the simpler logistic model might be preferred over more complicated asymmetric models. Of the two asymmetric models studied (the bilogistic and Dirichlet models), neither resulted in significantly different (maximised) log-likelihoods. Further, estimates of both ‘asymmetry’ parameters were similar within both the bilogistic and Dirichlet models, suggesting symmetry in the dependence structure anyway.

We have shown that, by using a simple first-order Markov chain model using the bivariate logistic model as the distribution for consecutive pairs of the process, we can improve the efficiency of our return level inference. However, to account for temporal dependence (quantified by the logistic dependence parameter) in the estimate of return levels, a model-based estimate of the extremal index θ_m in each month m is required so we can solve equation (4.27) for q_r . We used the intervals estimator of Ferro and Segers (2003), since this does not require the (arbitrary) specification of a declustering parameter (as in the estimation of θ_m using runs or blocks declustering). The next chapter examines the sensitivity of return level inference to the choice of extremal index estimator, in a bid to possibly improve return level inference still further.

Chapter 5

Inference for the extremal index

5.1 Introduction

The extremal index, θ ($0 \leq \theta \leq 1$), is an important parameter measuring the degree of clustering of extremes in a stationary process. Apart from being of interest in its own right, it is the key parameter for extending extreme value theory for independent, identically distributed (IID) random variables to stationary processes. The concept of the extremal index was developed in a series of papers including Newell (1964), Loynes (1965), O'Brien (1974) and Leadbetter *et al.* (1983), with a review by Leadbetter and Rootzén (1988). From a statistical point of view, since there are already many publications on estimating extreme value distributions from IID data, it can be seen that estimating the extremal index is a key problem.

Section 1.4.1 gives a formal definition of the extremal index; an alternative characterisation, provided by Hsing *et al.* (1988), is that θ^{-1} is the limiting mean cluster size in the point process of exceedance times over a high threshold. This suggests that a suitable way to estimate the extremal index can be found through methods which formally identify ‘clusters’ of extremes, though such methods often require an arbitrary choice of declustering parameter. Section 2.4.2 outlines one such method – ‘runs’ declustering – where the declustering parameter κ (or ‘cluster termination interval’) is chosen such that a cluster of extremes above a chosen threshold u is terminated if at least κ consecutive observations fall below that threshold. Other cluster identification techniques have been developed, such as ‘blocks’ declustering; here, the data are divided into k blocks of length τ , with each block being treated as a cluster of extremes if it contains at least one exceedance of the threshold u . However, as with ‘runs’ declustering, this method requires an arbitrary choice of declustering parameter, the declustering parameter this time being τ , the block length; Ancona–Navarrete and Tawn (2000) show estimators based on the

formation of clusters to be highly sensitive to such arbitrary choices. To this end, Ferro and Segers (2003) have devised a method of estimating θ without any prior declustering. Their intervals estimator (outlined in section 3.5.2) is based on the inter-arrival times of exceedances, and so does not require the formation of clusters of extremes, avoiding the need to make an arbitrary choice of declustering parameter. Estimators of the extremal index have also been shown to be sensitive to the choice of threshold, u ; Smith and Weissman (1994) discuss how u can be chosen optimally, depending on the cluster identification procedure employed.

Other methods for estimating θ have been suggested (see Ancona-Navarrete and Tawn, 2000), and Smith (1992) presents a method for *calculating* the extremal index for a Markov chain (see section 5.3.2). After a more involved discussion of these methods, this chapter will attempt to compare the performance of the different estimators using both simulated data and the wind speed data introduced in chapter 2. The methods used to estimate the extremal index assume stationarity; though non-stationary sequences are of practical interest, an exact analog of the extremal index may not exist (Hüsler, 1986). In practice (and as has been demonstrated in previous chapters), a pragmatic approach is to break up the data into seasonal blocks (or months in the case of the wind speed data) within which they may be assumed stationary.

5.2 Estimation

Some of the following estimators have been described or outlined in previous chapters; here, more formal definitions are given, with superscripts to differentiate between them. For each, $(X_1, X_2, \dots, X_n)$ denotes the first n observations from a stationary sequence which satisfies the $D(u_n)$ condition (see section 1.4.1), where n is large and u_n is a high threshold.

5.2.1 Cluster size methods

These estimators use declustering techniques to identify clusters of extremes, and then estimate the extremal index as the reciprocal of the sample mean cluster size. In each, $N(u)$ is the number of threshold exceedances.

A baseline estimator

As a first attempt to identify clusters, we consider a cluster of exceedances to be terminated as soon as an observation falls below the threshold u . Thus,

$$\hat{\theta}^{(1)} = \frac{\sum_{j=1}^n I(X_{j+1} \leq u < X_j)}{N(u)}, \quad (5.1)$$

where I is the indicator function which takes the value 1 if observation j is greater than the threshold (which itself must be greater than or equal to the next observation $j + 1$), and 0 otherwise. Put more simply, $\hat{\theta}^{(1)}$ is the sample proportion of threshold downcrossings.

The runs estimator

Here, we take any sequence of at least κ consecutive observations below the threshold as separating two clusters. Define W_j to be 1 if the j th observation is above the threshold (i.e. $X_j > u$) and 0 otherwise. Obviously,

$$N(u) = \sum_{j=1}^n W_j; \quad (5.2)$$

Let

$$Z = \sum_{j=1}^{n-\kappa} W_j (1 - W_{j+1}) \cdots (1 - W_{j+\kappa}). \quad (5.3)$$

Then

$$\hat{\theta}^{(2)} = \frac{Z}{N(u)}. \quad (5.4)$$

The definition of Z ensures that an exceedance in position j is counted if and only if the following κ observations are all below the threshold u (Smith & Weissman, 1994). Thus, we count the number of clusters by counting the number of ‘last exceedances’. Note that the baseline estimator and the runs estimator are equivalent for $\kappa = 1$.

The blocks estimator

The blocks method of declustering divides the data into approximately k blocks of length τ , where $n \approx k\tau$. Each block is treated as one cluster; thus if Z^* denotes the number of blocks in which there is at least one exceedance of the threshold u , we consider $N(u)/Z^*$ to estimate the mean size of a cluster, and hence estimate θ by

$$\hat{\theta}^{(3)} = \frac{Z^*}{N(u)}. \quad (5.5)$$

5.2.2 Maxima methods

Consider a sequence divided into k blocks of length τ (as above). If we denote the maximum of the i th block by $M_{\tau,i}$, for $i = 1, \dots, k$, then it follows that for large enough τ , the $M_{\tau,i}$ are approximately independent observations from a GEV distribution. For any stationary sequence, an exchangeable sequence with the same marginal distribution as the dependent sequence can be generated by randomising the index of the observations. This exchangeable series acts as a good approximation to an associated IID series.

For this associated series, denote the block maxima by $\tilde{M}_{\tau,i}$ for $i = 1, \dots, k$. We model the $\tilde{M}_{\tau,i}$ as independent and identically distributed $\text{GEV}(\mu, \sigma, \xi)$ random variables. If the extremal index of the series is θ , then the dependent series block maxima, $M_{\tau,i}$, are modelled by a $\text{GEV}(\mu_\theta, \sigma_\theta, \xi_\theta)$ distribution, where

$$\mu_\theta = \mu - \sigma(1 - \theta^\xi)/\xi, \quad (5.6)$$

$$\sigma_\theta = \sigma\theta^\xi \quad \text{and} \quad (5.7)$$

$$\xi_\theta = \xi, \quad (5.8)$$

and (μ, σ, ξ) are the parameters of the independent series maxima (see Coles, 2001). This idea leads to the following estimators.

Gomes' estimator

Gomes (1993) proposed estimating θ from the estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and $(\hat{\mu}_\theta, \hat{\sigma}_\theta, \hat{\xi}_\theta)$ obtained from the separate fits to the two sets of block maxima. A pooled estimate of ξ is calculated as

$$\tilde{\xi} = \frac{\hat{\sigma} - \hat{\sigma}_\theta}{\hat{\mu} - \hat{\mu}_\theta}. \quad (5.9)$$

Then an estimate of θ is given by

$$\hat{\theta}^{(4)} = \left(\frac{\hat{\sigma}}{\hat{\sigma}_\theta} \right)^{-1/\tilde{\xi}}. \quad (5.10)$$

A maximum likelihood estimator

A newer approach based on block maxima, developed by Ancona–Navarrete and Tawn (2000), has been suggested. Here, instead of separately estimating the two sets of GEV parameters and then making subsequent adjustments to these to estimate θ , simultaneous estimation of the parameter vector $(\mu, \sigma, \xi, \theta)$ is proposed which determines all the

parameters of the distributions of M_τ and $\tilde{M}_\tau$. To undertake this simultaneous estimation we treat

$$(M_{\tau,1}, \dots, M_{\tau,k}, \tilde{M}_{\tau,1}, \dots, \tilde{M}_{\tau,k})$$

as independent, but not identically distributed, GEV random variables (with marginal distributions given above). Thus, the joint log-likelihood is given by

$$\begin{aligned} \ell(\mu, \sigma, \xi, \theta) = & -k(\log \sigma + \xi \log \theta) - \sum_{i=1}^k \left[\frac{\sigma + \xi(M_{\tau,i} - \mu)}{\sigma \theta^\xi} \right]^{-1/\xi} \\ & - (1/\xi + 1) \sum_{i=1}^k \log \left[\frac{\sigma + \xi(M_{\tau,i} - \mu)}{\sigma \theta^\xi} \right] \\ & - k \log \sigma - \sum_{i=1}^k \left[1 + \xi \left(\frac{\tilde{M}_{\tau,i} - \mu}{\sigma} \right) \right]^{-1/\xi} \\ & - (1/\xi + 1) \sum_{i=1}^k \log \left[1 + \xi \left(\frac{\tilde{M}_{\tau,i} - \mu}{\sigma} \right) \right]. \end{aligned} \quad (5.11)$$

The joint vector is not independent but Tawn shows that the impact of this approximation is negligible in practice and insignificant asymptotically. By this construction, θ is a parameter of the joint distribution, so a maximum likelihood estimator, $\hat{\theta}^{(5)}$, can be found by maximisation of the log-likelihood function given above.

5.2.3 An intervals estimator

The cluster size methods for estimating the extremal index require the formation of clusters prior to estimation. For example, the method of runs declustering requires the choice of separation interval κ to identify clusters, and this choice is made arbitrarily. Similarly with blocks declustering, the choice of block length τ is also made subjectively. Hsing (1991) points out that the choice of block length or run length usually has a significant influence on the estimate of θ . Ferro and Segers (2003) propose an estimator which doesn't require the formation of clusters. They call their estimator the intervals estimator, as it is based on the inter-arrival times between successive threshold exceedances. The intervals estimator is outlined in section 3.5.2, and is given as

$$\hat{\theta}^{(6)} = \frac{2 \left\{ \sum_{i=1}^{N-1} (T_i - 1) \right\}^2}{(N-1) \sum_{i=1}^{N-1} (T_i - 1)(T_i - 2)}, \quad (5.12)$$

where N is the number of exceedance times observed and T_i are the inter-arrival times between exceedances.

5.3 Performance of estimators

We compare the performance of the six extremal index estimators $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(6)}$ for different processes where $\theta = 1$ and $\theta < 1$. The estimators are compared using the mean bias and root mean square error (RMSE) criteria. In each case, we have samples of size 10,000, and these criteria are compared using 200 independent replications of the estimation procedures. To reduce sampling variation, for each process all the estimators are evaluated using the same simulated data sets.

All estimators based on the formation of clusters (i.e. $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$ and $\hat{\theta}^{(3)}$), as well as the intervals estimator ($\hat{\theta}^{(6)}$), are threshold-dependent; if u is too high there will be just a few exceedances of the threshold and so estimates will be highly variable, but if u is too low the approximation to the limiting characterisation may be poor, leading to biased estimates. Obviously, the sample size is important here, as this directly influences what is too high a threshold – see Hsing (1993) for more detail. Where the estimators depend on a threshold, we obtain estimates corresponding to 95% and 99% marginal quantiles. A similar trade-off exists in the selection of κ and τ in runs and blocks estimation; when these parameters are taken to be much too small/large, θ is often over/under-estimated (respectively). Smith and Weissman (1994) use an optimality criterion based on a bias-variance trade-off to find ‘optimal’ declustering parameters for various processes; these values are given in table 5.1. Thus, when comparing the estimators using simulated data, $\hat{\theta}^{(2)}$ and $\hat{\theta}^{(3)}$ have a substantial advantage over the other estimators as they are ‘tailored’ to do well in a way which can not be achieved in practice.

Section 5.3.1 investigates the performance of the extremal index estimators for processes which have $\theta = 1$ (investigating the performance of cluster size estimators for independent processes, for example, might help to assess the relative shortcomings of different declustering techniques); section 5.3.2 looks at processes with $\theta < 1$. Both sections also examine the sensitivity of the runs and blocks estimators to the choice of declustering parameters κ and τ . For the maxima methods (section 5.2.2), we take the number of blocks and length of blocks to be equal, in accord with Gomes (1993).

Process	Runs, κ	Blocks, τ
Independent	1	2
AR(1), $A = 0.5$	4	20
Markov bivariate extreme value, $\theta = 0.328$	10	20

Table 5.1: Optimal values for cluster size estimators for some processes

5.3.1 Processes with $\theta = 1$

Independent process

For any independent process, $\theta = 1$. Though it may seem rather strange to attempt to quantify the extremal dependence of an independent process, investigating the performance of cluster size estimators for such a process might indicate where certain declustering procedures fail; here, they attempt to identify clusters when there are none! To consider such a process, we simulate a series of completely independent random variables from a Gumbel distribution. We first investigate the sensitivity of the runs and blocks estimators to the choice of declustering parameters κ and τ (respectively). Figure 5.1 shows the mean value and RMSE for both estimators, with a threshold set at the 99% quantile (the “99% threshold”), using various cluster identification parameters. It shows the runs estimator is much more sensitive to κ than the blocks estimator is to τ , with significantly increased bias and RMSE for $\hat{\theta}^{(2)}$ as κ increases and relative stability in estimates of θ using the blocks estimator. For an independent process, cluster identification using the runs approach is difficult. The reason for this is that, for a threshold u with associated exceedance probability λ_u , the probability of independently re-exceeding the threshold in the $\kappa - 1$ observations following a downcrossing of that threshold is $1 - (1 - \lambda_u)^{\kappa-1}$, so is not negligible for thresholds in the 95% – 99% range which are typically used. Figure 5.1 clearly shows the optimal values for κ and τ are both 1, but since we must have $\tau \geq 2$ to be able to identify blocks, we have an optimal value of $\tau = 2$ for the blocks estimator. These findings are in agreement with the theoretical findings of Smith and Weismann (1994), shown in the first row of table 5.1.

The performance of the estimators for such an independent process is presented in table 5.2, which shows the sampling distribution means for each estimator and (in parentheses) the RMSE. Note that the baseline and runs estimators are equivalent here, since the optimal declustering parameter for the runs estimator is $\kappa = 1$. All estimators appear to perform reasonably well, with Gomes’ estimator ($\hat{\theta}^{(4)}$) performing particularly well, with a sample mean of 1.000 (to three decimal places) and a very small RMSE. Of the cluster size estimators, the blocks estimator seems best, though the difference between runs and blocks is negligible. We also see that threshold selection has had little impact upon estimation, with only slight variation between thresholds for the runs and intervals estimators. The blocks estimator shows most stability here, being insensitive to threshold change.

Figure 5.1: Mean and RMSE for estimates of $\theta = 1$ for an independent process, plotted against cluster identification parameter for a 99% threshold. Runs ($\hat{\theta}^{(2)}$) and blocks ($\hat{\theta}^{(3)}$) are shown by solid and broken lines respectively

Estimator	Threshold employed	
	95% quantile	99% quantile
$\hat{\theta}^{(1)} = \hat{\theta}^{(2)}$	0.949 (0.055)	0.991 (0.000)
$\hat{\theta}^{(3)}$	0.996 (0.000)	0.996 (0.000)
$\hat{\theta}^{(6)}$	0.986 (0.032)	0.962 (0.063)
$\hat{\theta}^{(4)}$	1.000 (0.000)	
$\hat{\theta}^{(5)}$	0.991 (0.032)	

Table 5.2: Estimation of the extremal index, $\theta = 1$, for an independent sample: sampling mean, and, in brackets, the RMSE about θ

Gaussian AR(1) process

Many dependent sequences also have an extremal index of 1; a series for which $\theta = 1$ means that dependence is negligible at very high levels, but not necessarily so at extreme levels that are relevant for any particular application. In such a series, the limiting clusters occur as single values. Therefore, we need to find out if the conclusions for independent processes hold when there is sub-asymptotic dependence. To investigate, we consider a stationary first-order Markov chain with Gaussian dependence and Gumbel marginal distribution, i.e. a first-order autoregressive, or AR(1), process; Leadbetter *et al.* (1983) show such a process has $\theta = 1$ since both the $D(u_n)$ and $D'(u_n)$ conditions are satisfied, where the $D'(u_n)$ condition is a further restriction on *local* dependence of extremes (see section 1.4.1 and Leadbetter *et al.*, 1983). Specifically, we have $Z_j = AZ_{j-1} + \varepsilon_j$, where the ε_j are IID normal random variables with zero mean and variance given by $1 - A^2$, and Z_1 is drawn from a $N(0,1)$ distribution. A transformation from standard normal to Gumbel margins is attained by letting $X_j = -\log[-\log\{\Phi(Z_j)\}]$, where Φ is the distribution function of a standard normal random variable; this transformation is used to enable fair comparison with the independent process and the case $\theta < 1$ in section 5.3.2 (below), where the margins are also of Gumbel form. The lag t correlation of the process is A^t , for $-1 < A < 1$. We take $A = 0.5$, so the process has positive dependence at all lags.

As before, we begin by looking at the sensitivity of the runs and blocks estimators ($\hat{\theta}^{(2)}$ and $\hat{\theta}^{(3)}$ respectively) to the choice of declustering parameter. Figure 5.2 shows the mean values and RMSE for both estimators, with a 99% threshold, for various values of declustering parameters κ and τ . The main conclusion from these plots is that the runs estimator is much more sensitive to the choice of κ than the blocks estimator is to τ ; though both show bias for all values of cluster identification parameter, the blocks estimator seems relatively stable after about $\tau = 10$, and certainly after $\tau = 20$ (as suggested by table 5.1). In contrast, the runs estimator has a sampling distribution mean which appears to decrease continuously as κ increases, with an associated increase in the RMSE (as shown by the bottom plot). Thus, we might be tempted to suggest an optimal value of $\kappa = 1$ for the runs estimator (i.e. the baseline estimator), since this value gives a sampling distribution mean closest to the true value of $\theta = 1$ and the smallest RMSE. However, this estimator would classify a cluster of extremes with two (or more) up-crossings as two (or more) independent clusters; since we have sub-asymptotic dependence governed by the autoregressive parameter $A = 0.5$, these clusters might not necessarily be independent. This reflects Smith and Weissman's optimal choice of $\kappa = 4$ which is small, but is chosen to help ensure that distinct clusters of extremes are correctly identified.

Figure 5.2: Mean and RMSE for estimates of $\theta = 1$ for an AR(1) process, plotted against cluster identification parameter for a 99% threshold. Runs ($\hat{\theta}^{(2)}$) and blocks ($\hat{\theta}^{(3)}$) are shown by solid and broken lines respectively

Estimator	Threshold employed	
	95% quantile	99% quantile
$\hat{\theta}^{(1)}$	0.754 (0.247)	0.869 (0.134)
$\hat{\theta}^{(2)}$	0.627 (0.373)	0.822 (0.182)
$\hat{\theta}^{(3)}$	0.757 (0.245)	0.857 (0.148)
$\hat{\theta}^{(6)}$	0.745 (0.259)	0.876 (0.152)
$\hat{\theta}^{(4)}$	0.746 (0.255)	
$\hat{\theta}^{(5)}$	0.702 (0.270)	

Table 5.3: Estimation of the extremal index, $\theta = 1$, for an AR(1) process with $A = 0.5$: sampling mean, and, in brackets, the RMSE about θ

Table 5.3 summarises the performance of the six estimators, again showing sampling distribution means and the root mean square error (in parentheses). We see that $\hat{\theta}^{(6)}$ (at the 99% threshold) seems to be less biased than the other estimators, though $\hat{\theta}^{(1)}$ (the baseline estimator) is a close second and has the smallest RMSE. The runs estimator appears to perform poorly in comparison, suggesting that selection of the optimal κ value is not clear-cut, with $\kappa = 1$ being optimal in this example. In contrast to the independent process, here threshold-based estimators are sensitive to the choice of threshold, with all four of these estimators providing better estimates at the higher threshold level. The maxima estimators ($\hat{\theta}^{(5)}$ and $\hat{\theta}^{(6)}$) compare well with the other estimators at the lower threshold level, but are out-performed by the threshold-based estimators at the 99% level. These results are in accordance with the findings of Ancona-Navarrete and Tawn (2000), who show that in fact the estimators are estimating $\theta(u)$ rather than θ itself, where $\theta(u)$ is a threshold-based extremal index. This threshold-based index is a penultimate expression for θ , which allows us to study dependence within the extreme values that are of practical interest; see Ancona-Navarrete and Tawn (2000) for more detail. Ancona-Navarrete and Tawn give the values of $\theta(u)$ for an AR(1) process to be 0.711 and 0.855 for thresholds set at the 95% and 99% quantiles respectively. Thus, we find that all four of the threshold-based estimators are estimating $\theta(u)$ rather well. This is in contrast to the study which examined an independent process, where all methods used gave reasonably accurate estimates of θ itself (though this could be because, for independent processes, $\theta(0.95)$ and $\theta(0.99)$ are close to θ).

5.3.2 Processes with $\theta < 1$

Bivariate extreme value Markov process

Here, we look at a process for which $\theta < 1$; namely, a first-order Markov chain for which the joint distribution of consecutive pairs of values is provided by the logistic model, whose distribution function is given by equation (4.15). For completely independent processes, and indeed for autoregressive processes, we know $\theta = 1$. However, for a bivariate extreme value Markov chain, the extremal index will be dependent on the parameter α , which itself quantifies the strength of dependence between consecutive extremes of the process. Smith (1992) exploits some scaling properties of bivariate extremes which leads to a random walk representation for the tail behaviour of a stationary (first-order) Markov process, and so to the computation of the extremal index in terms of the fluctuation properties of that random walk. Thus the limiting values of θ for various values of logistic dependence (given by α – see chapter 4) can be found. Specifically, Smith shows the extremal index is given

	$\alpha = 0.5$	$\alpha = 0.333$	$\alpha = 0.25$	$\alpha = 0.2$
Limiting θ	0.328	0.158	0.0925	0.0616

Table 5.4: Calculated values of θ for various values of logistic dependence α

by

$$\theta = \int_{-\infty}^0 e^x Q(x) dx, \quad (5.13)$$

where Q is a solution to the Wiener–Hopf equation

$$Q(x) = \int_0^\infty Q(y) H(x - dy) \quad (5.14)$$

(where $Y_1, Y_2, \dots$ form a random walk and H is the distribution function of the differences $Y_2 - Y_1, Y_3 - Y_2, \dots$), subject to the normalising condition

$$\lim_{x \rightarrow \infty} Q(x) = 1. \quad (5.15)$$

Table 5.4 gives the limiting values of θ based on equation (5.13); for more detail, see Smith (1992).

Again, we first investigate the sensitivity of the runs and blocks estimators to the choice of declustering parameter used to identify clusters of threshold exceedances. Figure 5.3 shows the mean value and RMSE for both estimators, with a 99% threshold, using various values of κ and τ . Since consecutive pairs of the process are drawn from a bivariate logistic model with $\alpha = 0.5$, the true value of θ is 0.328 (shown by the dotted line in the top plot of figure 5.3). We see that both the runs and blocks estimators are sensitive to the choice of cluster identification parameter, with the sampling means continuously decreasing as κ/τ increases. However, the blocks estimator is larger (on average) for all $\kappa = \tau$ used in this study. For the runs estimator, the RMSE appears to be minimised at about $\kappa = 10$, where this value of declustering parameter also gives a sampling mean closest to the true value of $\theta = 0.328$ (see top plot). For the blocks estimator, we see that the RMSE is minimised, and relatively stable, for all $\tau \geq 20$. These conclusions support the findings of Smith and Weissman (1994), who suggest optimal declustering parameters of $\kappa = 10$ and $\tau = 20$ for the runs estimator and blocks estimator (respectively) of a bivariate extreme value Markov process (see table 5.1). For the values of κ and τ used in this study, the runs estimator performs better than the blocks estimator; however, for some values of $\kappa = \tau > 40$, figure 5.3 suggests the blocks estimator might be best.

We now compare the six extremal index estimators for bivariate extreme value Markov

Figure 5.3: Mean and RMSE for estimates of $\theta = 0.328$ for a bivariate extreme value Markov chain, plotted against cluster identification parameter for a 99% threshold. Runs ($\hat{\theta}^{(2)}$) and blocks ($\hat{\theta}^{(3)}$) are shown by solid and broken lines respectively; the dotted line in the top plot represents the true value of θ

Estimator	$\alpha = 0.5$ ($\theta = 0.328$)		$\alpha = 0.2$ ($\theta = 0.0616$)	
	95% quantile	99% quantile	95% quantile	99% quantile
$\hat{\theta}^{(1)}$	0.400 (0.077)	0.417 (0.106)	0.150 (0.094)	0.198 (0.181)
$\hat{\theta}^{(2)}$	0.276 (0.057)	0.325 (0.051)	0.074 (0.024)	0.107 (0.097)
$\hat{\theta}^{(3)}$	0.375 (0.053)	0.376 (0.072)	0.130 (0.072)	0.161 (0.130)
$\hat{\theta}^{(6)}$	0.338 (0.035)	0.360 (0.088)	0.070 (0.023)	0.113 (0.128)
$\hat{\theta}^{(4)}$	0.373 (0.051)		0.121 (0.063)	
$\hat{\theta}^{(5)}$	0.334 (0.042)		0.134 (0.026)	

Table 5.5: Estimation of the extremal index for an extreme value Markov chain (with logistic dependence) for $\alpha = 0.5$ and $\alpha = 0.2$ ($\theta = 0.328$ and $\theta = 0.0616$ respectively): sampling mean, and, in brackets, the RMSE about θ

processes, with two strengths of logistic dependence, $\alpha = 0.5$ and $\alpha = 0.2$; from table 5.4, we see the ‘target’ values for these processes are 0.328 and 0.0616 respectively. Table 5.5 summarises our findings, again showing results for both the 95% and 99% thresholds for threshold-dependent estimators. For $\alpha = 0.5$, The runs estimator ($\hat{\theta}^{(2)}$) seems to perform best, with a sampling mean of 0.325 at the 99% threshold level – closer to the true value of 0.328 than any other estimator. The intervals estimator ($\hat{\theta}^{(6)}$) also appears to fair well, with a sampling mean close to the true value and the smallest RMSE (0.035 at the 95% threshold level). Of the maxima estimators, Ancona–Navarrete and Tawn’s maximum likelihood estimator, $\hat{\theta}^{(5)}$, performs best, with a smaller mean bias and RMSE than $\hat{\theta}^{(4)}$. Relative to the other estimators, the naïve baseline estimator performs poorly. Since, often, some clusters will have two (or more) upcrossings of the threshold, this estimator will mis-classify a single cluster as two or more independent clusters. This extends to the case when $\alpha = 0.2$, with $\hat{\theta}^{(1)}$ having the largest mean bias and RMSE, no matter what the threshold level. When $\alpha = 0.2$, the intervals estimator appears to be more ‘on target’ than any other, with a sampling distribution mean of 0.070 (recall the true value when $\alpha = 0.2$ is 0.0616). This estimator also has the smallest RMSE. Of the cluster size methods, the runs estimator is again better than both the baseline estimator and the blocks estimator, having smaller RMSE and being less biased. Most threshold-based estimators appear to be insensitive to threshold change at the lower level of dependence, but when $\alpha = 0.2$ sampling means of these estimators change substantially when the threshold is raised. Again, it should be noted that $\hat{\theta}^{(2)}$ and $\hat{\theta}^{(3)}$ have a significant advantage over the other estimators, as they have been tailored to do well by choosing κ and τ optimally.

5.4 Application to wind speed data

Here, we compare extremal index estimation techniques using the wind speed data collected at High Bradfield in the Peak District (see chapter 2 for a discussion of these data). Estimating θ separately for each month (owing to seasonal variation), we use the notation $\hat{\theta}_m^{(j)}$, $m = 1, \dots, 12$, $j = 1, \dots, 6$, to denote the extremal index estimator $\hat{\theta}^{(j)}$ in month m . Section 4.3.1 showed that estimating the extremal index was a necessary step in estimating the r -year return level, q_r , when modelling the temporal dependence of a process using a Markov chain model; $\hat{q}_r$ is found by solving equation (4.27), where σ_m , ξ_m and λ_{u_m} are replaced by their maximum likelihood estimates, and θ_m is estimated via simulation, using the maximum likelihood estimate for an appropriate dependence parameter. Within this simulation scheme, we used Ferro and Segers’ intervals estimator ($\hat{\theta}^{(6)}$) to estimate θ since it avoids the need to make an arbitrary choice of declustering parameter (as in runs or blocks estimation). However, as this chapter has shown, if we

can assume the data are, for example, first-order extreme value Markovian, we can examine the performance of the runs and blocks estimators and find optimal values of κ or τ . Such an approach could indicate that the runs or blocks estimators outperform the intervals estimator (using these optimal declustering parameters), and so our return level estimation procedure might gain accuracy and efficiency.

5.4.1 Extremal index estimation

We first investigate the performance of the runs and blocks estimators using the 95% and 99% thresholds. Figure 5.4 shows estimates of θ_1 (January) and θ_6 (June) using the runs and blocks estimators, for a range of cluster identification parameters κ and τ (similar results were obtained for other months). Three features are apparent from figure 5.4: the blocks estimator gives higher estimates of both θ_1 and θ_6 for all values of declustering parameter; estimates increase with threshold; and the extremal dependence is quite strong.

Obviously, without knowing the *exact* value of θ_1 or θ_6 , it is difficult to determine which values of declustering parameter lead to the best estimators. However, when first-order Markov chain models were fitted to the monthly data at Bradfield (see table 4.1), the logistic dependence parameter, α , was estimated as 0.377 in January and 0.477 in June. Table 5.4 gives true values of θ for various levels of logistic dependence; $\alpha = 0.333$ gives $\theta = 0.158$ and $\alpha = 0.5$ corresponds to $\theta = 0.328$. Thus, $\hat{\alpha}_1 = 0.377$ would correspond to an extremal index close to 0.158, and $\hat{\alpha}_6 = 0.477$ would match up closely to an extremal index of just less than 0.328. We have interpolated between $\alpha = 0.333$ and $\alpha = 0.5$ to obtain approximate values of θ between these levels of dependence – these approximate values are shown as the dotted (horizontal) lines in figure 5.4. Obviously, this assumes the Bradfield data are first-order Markov; section 4.5 concludes that a second-order Markov chain model provides a better approximation, but this does not necessarily imply the first-order Markov chain model is completely inadequate. Further, differences between the first- and second-order models as far as return level estimation is concerned were found to be negligible. If we are prepared to accept a first-order Markov assumption, then we can see that, at the higher thresholds, the optimal declustering parameter for the runs estimator is about $\kappa = 10$ and $\kappa = 5$ for January and June respectively. At the higher thresholds, the blocks estimator seems to be quite stable for $\tau \geq 20$ in both months. These findings (loosely) correspond to the optimal κ and τ for a Markov bivariate extreme value process given in table 5.1 (after Smith & Weissman, 1994). However, if we prefer the second-order Markov assumption, then this longer-range dependence will induce stronger extremal dependence and so result in a lower extremal index. In this case, the dotted lines in figure 5.4 are misleading.

Table 5.6 summarises the results for January and June for the runs and blocks estimators (using ‘optimal’ declustering values obtained from figure 5.4, assuming the data are first-order Markov), as well as the other four extremal index estimators. Comparing seasons, we see that estimates are higher in June than in January; this is to be expected, since the fitted Markov chain models suggested stronger extremal dependence in January than in June (recall that $\hat{\alpha}_1 = 0.377$ and $\hat{\alpha}_6 = 0.477$). All threshold-based estimators are threshold-sensitive, with estimates increasing at the higher threshold level, though the intervals estimator ($\hat{\theta}^{(6)}$) is probably the least sensitive to changes in threshold level. We might assume that the baseline estimator is inadequate here; as before, this estimator will incorrectly classify a single cluster of extremes as two independent clusters if there is a single down-crossing within that cluster. Generally, the maxima estimators give lower values than the cluster size methods. A reason for this could be non-stationarity within each season, so that the maxima over blocks in some periods of the month may be lower than in others, leading to the GEV being a poor representation of the block maxima distribution. The evidence in table 5.6 (from the lower thresholds) points to θ_1 being in the range (0.07, 0.14) and θ_6 in the range (0.13, 0.31).

5.4.2 Consequences for return level estimation

We now examine the sensitivity of return level estimation to extremal index estimation, by using the method outlined in section 4.3.1 to estimate the r -year return level q_r . Using the Markov chain models methodology, we use the fitted values of α_m to simulate a Markov chain, using the logistic family to model the joint distribution of consecutive pairs of realisations. Since section 4.6.3 showed there to be little advantage of fitting a second-order Markov chain model as far as return level estimation is concerned, we use the results from fitting first-order Markov chains to monthly exceedances (given in table 4.1). However, unlike before, when the intervals estimator was used to then estimate the extremal index θ_m , we now consider the other estimators too (the results for January and June as given in table 5.6). As before, the process of simulating a Markov chain and estimating the extremal index is repeated to create a sampling distribution for the $\hat{\theta}_m^{(j)}$, and then the means of these sampling distributions are used to replace θ_m in equation (4.27).

Table 5.7 shows estimated return levels for the 10, 50, 200 and 1000-year return periods, based on first-order Markov chain models fitted to sets of monthly threshold excesses (see table 4.1 for estimates of α_m). Equation (4.27) has been solved by replacing σ_m , ξ_m and λ_{u_m} with their maximum likelihood estimates and θ_m with $\hat{\theta}_m^{(j)}$, $j = 1, \dots, 6$, to investigate the effects of all six estimators of θ_m on return level inference. Since threshold-dependent

Figure 5.4: Estimation of θ_1 (January, top) and θ_6 (June, bottom) for Bradfield wind speed data. Estimates are obtained using the runs (solid line) and blocks (broken line) estimators. The two lines for each estimator show estimates for the 95% and 99% thresholds, with the higher line coming from the higher threshold in each case

Estimator	January ($m = 1$)		June ($m = 6$)	
	95% quantile	99% quantile	95% quantile	99% quantile
$\hat{\theta}_m^{(1)}$	0.209	0.330	0.264	0.481
$\hat{\theta}_m^{(2)}$	0.101	0.205	0.147	0.299
$\hat{\theta}_m^{(3)}$	0.136	0.233	0.312	0.336
$\hat{\theta}_m^{(6)}$	0.070	0.112	0.125	0.178
$\hat{\theta}_m^{(4)}$	0.117		0.140	
$\hat{\theta}_m^{(5)}$	0.121		0.134	

Table 5.6: Estimates of the extremal index for the Bradfield wind speed data

θ_m replaced with	$\hat{q}_{10}$		$\hat{q}_{50}$		$\hat{q}_{200}$		$\hat{q}_{1000}$	
	95%	99%	95%	99%	99%	99%	95%	99%
$\hat{\theta}_m^{(1)}$	98.729	99.658	104.117	104.937	109.719	110.441	114.173	114.442
$\hat{\theta}_m^{(2)}$	96.973	98.415	102.546	103.831	108.324	109.446	112.873	113.247
$\hat{\theta}_m^{(3)}$	97.939	98.676	103.411	104.057	109.104	109.650	113.609	113.643
$\hat{\theta}_m^{(6)}$	96.745	98.043	103.236	103.490	108.152	109.131	113.306	113.563
$\hat{\theta}_m^{(4)}$	97.672		103.176		108.899		113.428	
$\hat{\theta}_m^{(5)}$	98.271		103.722		109.352		113.795	

Table 5.7: Return level estimates based on first-order Markov assumption; equation (4.27) solved using various extremal index estimators

extremal index estimators have been shown to be sensitive to the choice of threshold (table 5.6), we show results for estimators based on both the 95% and 99% thresholds. Comparing estimates using threshold-dependent estimators of θ_m , we see that the higher estimates of θ_m at the 99% threshold level have also resulted in higher estimates of q_r for all values of r studied. We also see very little difference between estimates of q_r which replace θ_m with $\hat{\theta}_m^{(6)}$ (as used in section 4.3.1) and estimates which use the runs estimator for θ_m ($\hat{\theta}_m^{(2)}$). Estimates using the baseline estimator ($\hat{\theta}_m^{(1)}$) are always slightly higher than those which use other cluster-size methods to estimate θ_m ; as already discussed, using such a method to estimate θ_m in a series which displays strong serial correlation is inappropriate, since too many ‘independent’ clusters will be identified. The two maxima methods for estimating the extremal index ($\hat{\theta}_m^{(4)}$ and $\hat{\theta}_m^{(5)}$) seem to give relatively similar estimates of q_r . Generally, the different methods for estimating the extremal index appear to have little consequence for return level inference. Comparing these estimates to those in table 4.2, we have further evidence that using cluster peak excesses in a simple extreme value analysis (i.e. removing temporal dependence) gives under-estimated return levels (when there is moderately strong extremal dependence). Using all threshold exceedances (i.e. ignoring dependence) gives estimates of q_r which agree closely to those which use the Markov chains methodology, but with less precision.

5.5 Discussion

As well as extending work from the previous chapter by investigating the sensitivity of return level estimation to the choice of extremal index estimator, the current chapter has also examined extremal index estimation more generally by comprehensively comparing the performance of six different estimators. Three different ‘types’ of estimator have been compared – ‘cluster size methods’, ‘maxima methods’ and Ferro and Segers’ intervals esti-

mator. A number of papers have already compared the performance of various estimators of θ when $\theta < 1$; however, these papers have ignored the $\theta = 1$ case, and fail to investigate the impact of different estimators in any practical application. We have investigated the performance of six extremal index estimators for two processes with $\theta = 1$ – namely a completely independence process with Gumbel margins, and a Gaussian AR(1) process with sub-asymptotic dependence. We have also investigated the performance of these estimators for two processes with $\theta < 1$, both of which were extreme value Markov processes, one with dependence parameter $\alpha = 0.5$ and the other with $\alpha = 0.2$. For all processes, the sensitivity of the ‘cluster size’ estimators to the choice of declustering parameter was investigated. Sensitivity to threshold level was also considered for threshold-dependent estimators. Attention was then focused on a real-life data set – the Bradfield wind speed data – and the sensitivity of return level estimation to the choice of extremal index estimator (using a first-order Markov chain model as discussed in chapter 4) was considered.

For the completely independent process, the blocks estimator appears to be least affected by the choice of declustering parameter, with relatively stable estimates of θ and consistently lower RMSE than the runs estimator. Neither the runs nor blocks estimators were particularly affected by threshold level. Overall, Gomes’ estimator seemed to perform best, with a sampling mean exactly equal to the true parameter value (correct to 3 decimal places) and the smallest RMSE. Similarly, for the Gaussian AR(1) process, the runs estimator performed relatively poorly, and seemed to be most influenced by the choice of declustering parameter. As discussed, all extremal index estimators for the AR(1) process appear to estimate the threshold-based extremal index $\theta(u)$ rather than the limiting value θ . The baseline estimator performs well here, with a sampling distribution mean closest to the true (limiting) value of $\theta = 1$ and the smallest RMSE. In contrast to the completely independent process, threshold-based estimators appear sensitive to threshold level, with more accurate estimates obtained at the higher threshold.

For an extreme value Markov process (where $\theta < 1$), we found that both runs and blocks estimators were sensitive to the choice of declustering parameter, and, generally, the runs estimator had a smaller RMSE for all κ/τ used in the study. We concluded that, for the runs estimator, an optimal choice of cluster termination interval is (about) $\kappa = 10$; this value gave an estimate of θ closest to the true value, and so also returned the smallest RMSE. For the blocks estimator, the RMSE appeared to stabilise after about $\tau = 20$, and so we suggested a block length of $\tau \geq 20$ as an optimal choice of declustering parameter under the blocks estimator. Of all the estimators, the runs estimator performed best (especially for $\alpha = 0.2$), whilst the somewhat naïve baseline estimator performed poorly.

Care must be taken when interpreting the accuracy of the threshold-based estimators; in the studies for which $\theta = 1$, the estimators $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$, $\hat{\theta}^{(3)}$ and $\hat{\theta}^{(6)}$ all appear to give more accurate estimates at the higher threshold. However, rather than the raised threshold returning estimates closer to the limiting value of θ , it is more likely that as the threshold is raised there are more clusters of length 1 being identified. Thus, the reciprocal of the mean cluster size is bound to move towards 1 as the threshold is increased.

When the different estimators were applied to the Bradfield wind speed data, and these estimates used (in turn) to replace θ_m in equation (4.27), the effect on return level inference was found to be negligible. With the possible exception of the baseline estimator (which seemed to result in higher estimates of q_r), all other estimators gave reasonably similar estimates of the 10, 50, 200 and 1000-year return levels; since we are assuming a first-order Markov dependence structure, and for such processes the baseline estimator was shown to perform poorly anyway, the fact that results based on using the baseline estimator are inconsistent with other results is unsurprising. Indeed, for such extreme value Markov processes (as we are assuming by fitting a first-order Markov chain model), section 5.3.2 showed the runs estimator to perform best; thus, if we were to choose an estimator from which we could most confidently trust return level estimates for the Bradfield wind speed data, we would probably use the runs estimator. Optimal declustering parameters κ_m for each month m can be found via plots such as those in figure 5.4.

Ancona-Navarrete and Tawn (2000) suggest ways in which the Markov chain assumption can be built into the estimation procedure for θ . Though they found that knowing such a feature about the process could improve extremal index estimation, they also found these improvements to be limited. Nonetheless, this avenue could provide a useful path for further research.

Chapter 6

Bayesian inference for extremes

6.1 Introduction

Throughout this thesis, the method of maximum likelihood has provided a general and flexible technique for parameter estimation. Given a (generic) parameter vector $\boldsymbol{\psi}$ within a family $\boldsymbol{\Psi}$, the likelihood function is the probability (density) of the observed data as a function of $\boldsymbol{\psi}$. Values of $\boldsymbol{\psi}$ that have high likelihood correspond to models which give high probability to the observed data. The principle of maximum likelihood estimation is to adopt the model with greatest likelihood; of all the models under consideration, this is the one that assigns the highest probability to the observed data. Other inferential procedures, such as “method of moments”, provide viable alternatives to maximum likelihood estimation; moments-based techniques choose $\boldsymbol{\psi}$ optimally by equating model-based and empirical moments, and solving for $\boldsymbol{\psi}$ to obtain parameter estimates. These, and other procedures (such as probability weighted moments, L -moments and ranked set estimation), are discussed in detail in, amongst other places, Kotz and Nadarajah (2000).

6.1.1 General theory

Bayesian techniques offer an alternative way to draw inferences from the likelihood function, which many practitioners often prefer. As in the non-Bayesian setting, we assume data $\boldsymbol{x} = (x_1, \dots, x_n)$ to be realisations of a random variable whose density falls within a parametric family $\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{\psi}) : \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$. However, parameters of a distribution are now treated as random variables, for which we specify *prior distributions* – distributions of the parameters *prior* to the inclusion of data. The specification of these prior distributions enables us to supplement the information provided by the data – which, in extreme value analyses, is often very limited – with other sources of information. At the same time, it can be contended that, since different analysts might specify different priors, conclusions

become subjective.

Leaving aside the arguments for and against the Bayesian methodology, suppose we model our observed data $\mathbf{x}$ using the probability density function $f(\mathbf{x}; \boldsymbol{\psi})$. The likelihood function for $\boldsymbol{\psi}$ is therefore $L(\boldsymbol{\psi}|\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\psi})$. Also, suppose our prior beliefs about likely values of $\boldsymbol{\psi}$ are expressed by the probability density function $\pi(\boldsymbol{\psi})$. We can combine both pieces of information using Bayes Theorem, which states that

$$\pi(\boldsymbol{\psi}|\mathbf{x}) = \frac{\pi(\boldsymbol{\psi})L(\boldsymbol{\psi}|\mathbf{x})}{f(\mathbf{x})}, \quad (6.1)$$

where

$$f(\mathbf{x}) = \begin{cases} \int_{\Psi} \pi(\boldsymbol{\psi})L(\boldsymbol{\psi}|\mathbf{x})d\boldsymbol{\psi} & \text{if } \boldsymbol{\psi} \text{ is continuous,} \\ \sum_{\Psi} \pi(\boldsymbol{\psi})L(\boldsymbol{\psi}|\mathbf{x}) & \text{if } \boldsymbol{\psi} \text{ is discrete.} \end{cases} \quad (6.2)$$

Since $f(\mathbf{x})$ is not a function of $\boldsymbol{\psi}$, Bayes Theorem can be written as

$$\begin{aligned} \pi(\boldsymbol{\psi}|\mathbf{x}) &\propto \pi(\boldsymbol{\psi}) \times L(\boldsymbol{\psi}|\mathbf{x}) \\ \text{i.e. posterior} &\propto \text{prior} \times \text{likelihood.} \end{aligned} \quad (6.3)$$

In equation (6.1), $\pi(\boldsymbol{\psi}|\mathbf{x})$ is the *posterior* distribution of the parameter vector $\boldsymbol{\psi}$, $\boldsymbol{\psi} \in \Psi$, i.e. the distribution of $\boldsymbol{\psi}$ *after* the inclusion of the data. This prior distribution is often of great interest, since the prior–posterior changes represent the changes in our beliefs after the data has been included in the analysis. However, computation of the denominator in (6.1) can be problematic, and usually analytically intractable. There is nothing particularly special about the fact that equation (6.1) represents a Bayesian posterior; given any complex non–standard probability distribution, we need ways to understand it, to calculate its moments, to compute its conditional and marginal distributions and their moments, all of which could require troublesome integration as in the denominator of equation (6.1). We need a way of understanding posterior densities which does not rely on being able to analytically integrate the kernel of the posterior; stochastic simulation is one possible solution.

6.1.2 Markov chain Monte Carlo

The recent explosion in Markov chain Monte Carlo (MCMC) techniques owes largely to their application in Bayesian inference. The idea here is to produce simulated values from the posterior distribution – not exactly, as this is usually unachievable, but through an appropriate MCMC technique.

The Gibbs sampler

The Gibbs sampler is a way of simulating from multivariate distributions based only on the ability to simulate from conditional distributions. Suppose the density of interest (usually the posterior density) is $\pi(\boldsymbol{\psi})$, where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)'$, and that the full conditionals

$$\pi(\psi_i | \psi_1, \dots, \psi_{i-1}, \psi_{i+1}, \dots, \psi_d) = \pi(\psi_i | \boldsymbol{\psi}_{-i}) = \pi_i(\psi_i), \quad i = 1, \dots, d$$

are available for simulating from ($\boldsymbol{\psi}_{-i}$ denotes the parameter vector $\boldsymbol{\psi}$ excluding ψ_i). The Gibbs sampler uses the following algorithm:

1. Initialise the iteration counter to $k = 1$. Initialise the state of the chain to $\boldsymbol{\psi}^{(0)} = (\psi_1^{(0)}, \dots, \psi_d^{(0)})'$;
2. Obtain a new value $\boldsymbol{\psi}^{(k)}$ from $\boldsymbol{\psi}^{(k-1)}$ by successive generation of values

$$\begin{aligned} \psi_1^{(k)} &\sim \pi(\psi_1 | \psi_2^{(k-1)}, \dots, \psi_d^{(k-1)}) \\ \psi_2^{(k)} &\sim \pi(\psi_2 | \psi_1^{(k)}, \psi_3^{(k-1)}, \dots, \psi_d^{(k-1)}) \\ &\vdots \\ \psi_d^{(k)} &\sim \pi(\psi_d | \psi_1^{(k)}, \dots, \psi_{d-1}^{(k)}); \end{aligned}$$

3. Change counter k to $k + 1$, and return to step 2.

Each simulated value depends only on the previous simulated value, and not any other previous values or the iteration counter k . The Gibbs sampler can be used in isolation if we can readily simulate from the full conditional distributions; however, this is not always the case. Fortunately, the Gibbs sampler can be combined with Metropolis–Hastings schemes when the full conditionals are difficult to simulate from.

Metropolis–Hastings sampling

Suppose again that $\pi(\boldsymbol{\psi})$ is the density of interest. Further, suppose that we have some arbitrary transition kernel $p(\boldsymbol{\psi}_{i+1}, \boldsymbol{\psi}_i)$ (which is easy to simulate from) for iterative simulation of successive values. Then consider the following algorithm:

1. Initialise the iteration counter to $k = 1$, and initialise the chain to $\boldsymbol{\psi}^{(0)}$;
2. Generate a proposed value $\boldsymbol{\psi}'$ using the kernel $p(\boldsymbol{\psi}^{(k-1)}, \boldsymbol{\psi}')$;
3. Evaluate the *acceptance probability* $A(\boldsymbol{\psi}^{(k)}, \boldsymbol{\psi}')$ of the proposed move, where

$$A(\boldsymbol{\psi}, \boldsymbol{\psi}') = \min \left\{ 1, \frac{\pi(\boldsymbol{\psi}') L(\boldsymbol{\psi}' | \boldsymbol{x}) p(\boldsymbol{\psi}', \boldsymbol{\psi})}{\pi(\boldsymbol{\psi}) L(\boldsymbol{\psi} | \boldsymbol{x}) p(\boldsymbol{\psi}, \boldsymbol{\psi}')} \right\};$$

4. Put $\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}'$ with probability $A(\boldsymbol{\psi}^{(k-1)}, \boldsymbol{\psi}')$, and put $\boldsymbol{\psi}^{(k)} = \boldsymbol{\psi}^{(k-1)}$ otherwise;
5. Change the counter from k to $k + 1$ and return to step 2.

So at each stage, a new value is generated from the proposal distribution. This is either accepted, in which case the chain moves, or rejected, in which case the chain stays where it is. Whether or not the move is accepted or rejected depends on the acceptance probability which itself depends on the relationship between the density of interest and the proposal distribution. Common choices for the proposal distribution include symmetric chains, where $p(\boldsymbol{\psi}, \boldsymbol{\psi}') = p(\boldsymbol{\psi}', \boldsymbol{\psi})$, and random walk chains, where the proposal $\boldsymbol{\psi}'$ at iteration k is $\boldsymbol{\psi}' = \boldsymbol{\psi} + \varepsilon_k$, where the ε_k are IID random variables.

Hybrid methods

Here, we combine Gibbs sampling and Metropolis–Hastings schemes to form hybrid Markov chains whose stationary distribution is the distribution of interest. For example, given a multivariate distribution whose full conditionals are awkward to simulate from directly, we can define a Metropolis–Hastings scheme for each full conditional, and apply them to each component in turn for each iteration. This is similar to Gibbs sampling, but each component update is a Metropolis–Hastings update, instead of a direct simulation from the full conditional. Another scheme, known as “Metropolis within Gibbs”, goes through each full conditional in turn, simulating directly from the full conditionals wherever possible, and carrying out a Metropolis–Hastings update elsewhere.

6.1.3 Bayesian inference for extremes

There are various (and some may say compelling) reasons for preferring a Bayesian analysis of extremes over the more traditional likelihood approach. As already discussed, since extreme data are (by their very nature) quite scarce, the ability to incorporate other sources of information through a prior distribution has obvious appeal. Bayes’ Theorem also leads to an inference that comprises a complete distribution, meaning that the variance of the posterior distribution, for example, can be used to summarise the precision of the inference, without having to rely upon asymptotic theory. Also, implicit in the Bayesian framework is the concept of the *predictive distribution*. This distribution describes how likely are different outcomes of a future experiment. The predictive probability density function is given by

$$f(y|\mathbf{x}) = \int_{\Psi} f(y|\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{x})d\boldsymbol{\psi} \quad (6.4)$$

when $\boldsymbol{\psi}$ is continuous. From equation (6.4), we can see that the predictive distribution is formed by weighting the possible values for $\boldsymbol{\psi}$ in the future experiment $f(y|\boldsymbol{\psi})$ by how

likely we believe they are to occur after seeing the data. For example, a suitable model for the threshold excess Y of a process is $Y \sim \text{GPD}(\sigma, \xi)$. Estimation of $\boldsymbol{\psi} = (\sigma, \xi)$ could be made on the basis of previous observations $\mathbf{x} = (x_1, \dots, x_n)$ using, for example, the Markov chain model approach of chapter 4. Thus, in the Bayesian framework, we would have

$$\Pr \{Y \leq y | x_1, \dots, x_n\} = \int_{\Psi} \Pr \{Y \leq y | \boldsymbol{\psi}\} \pi(\boldsymbol{\psi} | \mathbf{x}) d\boldsymbol{\psi}. \quad (6.5)$$

Equation (6.5) gives the distribution of a future threshold excess, allowing for both parameter uncertainty and randomness in future observations. Solving

$$\Pr \{Y \leq q_{r,\text{pred}} | x_1, \dots, x_n\} = 1 - \frac{1}{r} \quad (6.6)$$

for $q_{r,\text{pred}}$ therefore gives an estimate of the r -year return level that incorporates uncertainty due to model estimation. Though (6.5) may seem analytically intractable, it can be approximated if the posterior distribution has been estimated using, for example, MCMC. After removal of the “burn-in” period, the MCMC procedure gives a sample $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_B$ that can be regarded as realisations from the stationary distribution $\pi(\boldsymbol{\psi} | \mathbf{x})$. Thus

$$\Pr \{Y \leq q_{r,\text{pred}} | x_1, \dots, x_n\} \approx \frac{1}{B} \sum_{i=1}^B \Pr \{Y \leq q_{r,\text{pred}} | \boldsymbol{\psi}_i\}, \quad (6.7)$$

which we can solve for $q_{r,\text{pred}}$ using a numerical solver. Another reason lending appeal to Bayesian inference for extremes is that it is not dependent on the regularity assumptions required by the theory of maximum likelihood. For example, when $\xi < -0.5$, maximum likelihood estimation breaks down – in this situation, a Bayesian approach provides a feasible alternative¹.

6.2 Implementation for threshold excesses

We demonstrate the use of the Bayesian methodology via two illustrative examples, both focusing on the technique of modelling sets of threshold excesses with the generalised Pareto distribution: the first uses simulated data, in the form of a first-order Markov chain, the joint distribution of consecutive pairs of observations given by the (symmetric) logistic model; the second examines the wind speed data at High Bradfield. In both examples, to maximise precision, all threshold excesses are used in the analysis. Indeed,

¹Recall (from section 1.2.2) that, when $\xi < -0.5$, maximum likelihood fails due to the unboundedness of the likelihood function. A Bayesian analysis is not, in principle, affected by the unboundedness so long as the likelihood is integrable with respect to the prior (though the computational burden is substantially increased); for more information (and a worked example), see Coles and Powell (1996)

chapters 3 and 4 support the use of all threshold excesses no matter how strong the serial correlation, with analyses based on sets of cluster peak excesses resulting in under-estimated return levels in strongly dependent sequences. Without the availability of any expert prior knowledge, prior distributions with large variances are adopted for both σ and ξ .

6.2.1 Analysis of simulated data

Here, we perform a Bayesian analysis on a set of threshold excesses, using the techniques of chapter 3 to simulate our data: a first-order Markov chain is simulated, with the joint distribution of consecutive pairs of the process given by a bivariate extreme value distribution with logistic dependence function. In particular, the series is constructed such that the marginals have a generalised Pareto distribution, with parameters $\sigma = 1$ (so $\sigma^* = 0.30171$ – see section 3.3) and $\xi = -0.4$; the logistic dependence parameter used is $\alpha = 0.8$. For computational convenience (and to retain the positivity of σ^*), it is often more practical to reparameterise to $\text{GPD}(\eta, \xi)$, where $\eta = \log \sigma^*$ – simple exponentiation of the simulated values of η will then give posterior values for σ^* . In the absence of genuine prior information, independent priors with very large variances are adopted – we use

$$\pi(\eta, \xi) = \pi(\eta) \times \pi(\xi) \quad (6.8)$$

with

$$\pi(\eta) \sim N(0, 10000) \quad \text{and} \quad (6.9)$$

$$\pi(\xi) \sim N(0, 100), \quad (6.10)$$

following the example of Coles and Powell (1996) (of course, improper, completely flat priors *could* have been used). We use Metropolis–Hastings sampling (see section 6.1.2), to simulate values from the posterior, i.e.

1. Set an initial value $\boldsymbol{\psi} = (\eta, \xi)$;
2. Specify an arbitrary probability rule $p(\boldsymbol{\psi}_{i+1} | \boldsymbol{\psi}_i)$ for iterative simulation of successive values;
3. Use this probability rule to generate a candidate value $\boldsymbol{\psi}'$ for $\boldsymbol{\psi}_{i+1}$, accepting this with probability A ; otherwise $\boldsymbol{\psi}_{i+1} = \boldsymbol{\psi}_i$.

In this example, a random walk chain is proposed; thus, densities for p are specified corresponding to random walks in the two component directions:

$$\eta' = \eta_i + \varepsilon_\eta, \quad (6.11)$$

$$\xi' = \xi_i + \varepsilon_\xi, \quad (6.12)$$

with the ε being normally distributed with zero mean and variances v_η and v_ξ respectively. Choices of v_η and v_ξ need to be ‘tuned’ to maximise the efficiency of the algorithm. Ideally, we’d like the autocorrelations to decrease rapidly with increasing lag; increasing the innovations v_η and v_ξ reduces serial correlation in the chain, but also has the effect of reducing the overall acceptance probability and so the chain tends to get “stuck”. In this example, $v_\eta = 0.015$ and $v_\xi = 0.03$ were found to work reasonably well, giving an overall acceptance probability of about 30% for each parameter (0.320 for η and 0.339 for ξ). The lag 1 autocorrelations for σ^* and ξ were 0.975 and 0.941 (respectively), and were insignificant by about lag 40.

Figure 6.1 shows the output from the MCMC, with corresponding output for the 1000–observation return level. Recall that the *actual* parameter values are $\sigma^* = 0.30171$ and $\xi = -0.4$, giving a 1000–observation return level (using equation (2.10)) of $q_{1000} = 2.34226$. Thus for each iteration k , the posterior draws $\eta^{(k)}$ and $\xi^{(k)}$ are used to estimate the return level $q_{1000}^{(k)}$, to obtain a posterior distribution for the return level also. From figure 6.1, we see that the MCMC scheme appears to perform well, with rapid convergence to the stationary distributions. The chains also appear to ‘mix well’, owing to the choices of v_η and v_ξ . Plots of the posterior densities for σ^* , ξ and q_{1000} show the accuracy of the Bayesian analysis, with the vertical (dashed) lines, which represent the true parameter values, being close to the centre of each posterior. At a risk of being over-cautious, the first 5000 realisations were discarded as “burn-in” to ensure the remaining observations were from the target densities; thus, inferences are based on a sample of 45000 realisations.

Table 6.1 summarises the results of figure 6.1, showing the posterior means and standard deviations, as well as the 95% confidence intervals (or ‘credibility intervals’ as they are often termed in the Bayesian setting). Also shown, for comparison, are the corresponding results from a maximum likelihood analysis on the same simulated dataset; the standard errors (and hence the 95% confidence intervals) for the 1000–observation return level have been obtained via the delta method (see section 2.7.2). Comparing the Bayesian and likelihood–based analyses, we see that the Bayesian approach provides greater accuracy, with the posterior means of σ^* , ξ and q_{1000} lying closer to the true values than their maximum likelihood counterparts; the posterior standard deviations and likelihood–based standard errors are in close agreement.

6.2.2 Analysis of Bradfield wind speed data

We now analyse the Bradfield wind speed data using the Bayesian methodology. The MCMC scheme is set up as for the simulated data, i.e. we use a Metropolis–Hastings

Figure 6.1: Trace plots and posterior densities of GPD parameters (and 1000–observation return level) in a Bayesian analysis of simulated data

		$\sigma^* = \mathbf{0.30171}$	$\xi = \mathbf{-0.4}$	$q_{1000} = \mathbf{2.34226}$
Posterior distribution	mean (st. dev.)	0.305 (0.016)	−0.404 (0.030)	2.342 (0.013)
	95% C.I.	(0.274, 0.337)	(−0.459, −0.341)	(2.319, 2.369)
Maximum Likelihood	mean (st. dev.)	0.293 (0.015)	−0.410 (0.031)	2.317 (0.015)
	95% C.I.	(0.263, 0.323)	(−0.472, −0.348)	(2.287, 2.347)

Table 6.1: Summary statistics for the posterior scale, shape and 1000–observation return level, and corresponding maximum likelihood estimates

scheme with random walk updates for both the log(scale) and shape parameters η_m and ξ_m , $m = 1, \dots, 12$. In the absence of any prior information, we specify diffuse, independent priors for both the log(scale) and shape parameters, both having a normal distribution with zero mean and large variance. We then use each draw from the posteriors for η_m and ξ_m to estimate the 10, 50, 200 and 1000-year return levels, to obtain posterior distributions for these return levels. However, since the data display quite distinct seasonal variation, and we fit a different GPD to each month of threshold excesses separately (see chapter 2), for each iteration k the return level $q_r^{(k)}$ is obtained as the solution to

$$\sum_{i=1}^{12} \lambda_{u_m} \left[1 + \xi_m^{(k)} \left(\frac{q_r^{(k)} - u_m}{\exp(\eta_m^{(k)})} \right) \right]_+^{-1/\xi_m^{(k)}} = \frac{1}{r}, \quad (6.13)$$

$m = 1, \dots, 12$. As before, this equation is solved for $q_r^{(k)}$ using a program containing the NAG routine C05AJF (Numerical Algorithms Group, 1999).

The agreeable performance of the Metropolis–Hastings sampler for excesses in January, as shown by the trace plots in figure 6.2, is typical for data in other months, with chains mixing well and converging relatively quickly. Table 6.2 shows the posterior means (and standard deviations in parentheses) for the GPD parameters σ_m and ξ_m , which can be compared to the maximum likelihood estimates (and associated standard errors) given in table 2.2. Indeed, comparing tables 6.2 and 2.2, we see that the maximum likelihood estimates and the posterior means for both σ_m and ξ_m are in close agreement for all m . However, all posterior means for σ are slightly lower than (and posterior means for ξ slightly higher than) their maximum likelihood counterparts. Asymptotic standard errors and posterior standard deviations are extremely similar for all parameter estimates, though these *should* be inflated in the likelihood analysis (see section 2.6.3).

Figure 6.3 shows posterior distributions for the 10, 50, 200 and 1000-year return levels; table 6.3 summarises these distributions. Also displayed are the median and inter-quartile range for each posterior distribution. Due to the asymmetry of the return level posterior distributions, these summaries might be preferred; indeed, we see that the posterior mean is always higher than the posterior median, suggesting that the mean has been unduly weighted by some relatively large values. Comparing these posterior summaries to the maximum likelihood estimates for the 10, 50, 200 and 1000-year return levels (table 2.5), we see that the posterior medians are in close agreement to the maximum likelihood estimates. The posterior standard deviations are slightly larger than the asymptotic standard errors attached to the maximum likelihood estimates, probably due to the severe asymmetry of the return level posterior distributions.

Figure 6.2: Trace plots and posterior densities of GPD parameters in a Bayesian analysis of the Bradfield wind speed data – January

Month (m)	σ_m	ξ_m
1	8.078 (0.640)	−0.084 (0.057)
2	8.123 (0.665)	−0.090 (0.057)
3	8.898 (0.668)	−0.339 (0.051)
4	8.390 (0.665)	−0.259 (0.054)
5	6.843 (0.586)	−0.105 (0.063)
6	3.932 (0.350)	0.076 (0.068)
7	7.451 (0.566)	−0.397 (0.050)
8	6.955 (0.610)	−0.086 (0.063)
9	9.894 (0.405)	−0.133 (0.025)
10	7.718 (0.634)	−0.217 (0.060)
11	7.349 (0.594)	−0.046 (0.052)
12	8.332 (0.746)	−0.103 (0.069)

Table 6.2: Posterior means (and standard deviations) of the GPD scale and shape parameters

Figure 6.3: Trace plots and posterior densities for the 10, 50, 200 and 1000-year return levels in a Bayesian analysis of the Bradfield wind speed data

		Return period (years)			
		10	50	200	1000
Posterior	mean	97.963	104.598	109.120	115.576
	(st. dev.)	(4.227)	(7.797)	(12.874)	(21.376)
Posterior	median	96.694	103.108	107.926	112.649
	IQR	(5.17)	(8.74)	(13.36)	(19.83)

Table 6.3: Posterior summaries for the 10, 50, 200 and 1000-year return levels

6.3 Bayesian inference for measures of extremal dependence

We have discussed that one advantage of a Bayesian analysis over the traditional maximum likelihood approach is the ability to incorporate parameter uncertainty and randomness in future observations through predictive distributions, and in the context of extremes *predictive return levels* (see section 6.1.3). To fully exploit the extension to predictive densities, some form of temporal relationship can be imposed – for this, we could use the Markov chain models described in chapter 4. Other stochastic processes, which properly account for temporal dependence through the likelihood, are also available, many of which use the extremal index as a parameter for the model. This section extends the implementation of the Bayesian techniques illustrated in the previous section by looking at Bayesian inference for threshold excesses using Markov chain models, and Bayesian inference for the extremal index.

6.3.1 Inference for Markov chain models

In chapter 4 we discussed how temporal dependence of a process can be explicitly modelled using Markov chains. Specifically, we drew upon results from multivariate extreme value theory to model the joint distribution of consecutive values of a process, longer-range dependence being modelled by higher-order Markov chains. We found that, although the wind speed data at Bradfield displays strong serial dependence, and that a second-order Markov chain might be more appropriate than a first-order model to capture this dependence, in terms of return level inference nothing much was gained from the increased complexity and computational expense of the higher-order model. Further consideration was given to different choices of dependence structure (logistic, bilogistic and Dirichlet), though again, the increased complexity of the asymmetric models gave little, if any, benefits over the simpler logistic model.

We now consider a Bayesian analysis of the Bradfield data, using the Markov chain methodology described in chapter 4 to model the dependence aspects of the process. A first-order Markov chain model is employed, so we model the distribution of consecutive pairs of the series. The (bivariate) logistic model, whose distribution function is given by equation (4.15), is used to model such pairs, contributions to the likelihood arising from (4.20) depending on whether or not specific observations lie above or below the threshold. The modelling procedure follows that of section 4.2.2, i.e. a two-step method which estimates the GPD parameters σ_m and ξ_m first, then transforms the data to a unit Fréchet distribution (using these parameters) and estimates the dependence parameter

α_m separately. The full procedure in the Bayesian framework then is:

- In the absence of any expert prior information, specify independent, non-informative prior distributions for the parameters:

$$\pi(\eta_m, \xi_m, \alpha_m) = \pi(\eta_m) \times \pi(\xi_m) \times \pi(\alpha_m)$$

with

$$\begin{aligned}\pi(\eta_m) &\sim N(0, 10000), \\ \pi(\xi_m) &\sim N(0, 100) \quad \text{and} \\ \pi(\alpha_m) &\sim U(0, 1);\end{aligned}$$

- Set an initial value for $\boldsymbol{\psi} = (\eta_m, \xi_m, \alpha_m)$;
- Use a random walk procedure to generate a candidate value for each of η_m and ξ_m , i.e.

$$\begin{aligned}\eta'_m &= \eta_m + \varepsilon_{\eta_m}, \quad \text{and} \\ \xi'_m &= \xi_m + \varepsilon_{\xi_m};\end{aligned}$$

the ε being normally distributed with zero mean and variances v_{η_m} and v_{ξ_m} respectively;

- Accept η'_m (and ξ'_m) as the next value in the chain with probability A , otherwise reject the candidate value (and set the next value equal to the current value);
- At each iteration, use the simulated values for η_m and ξ_m to transform the data to unit Fréchet;
- Use a random walk procedure to generate a candidate value for α_m , i.e.

$$\alpha'_m = \alpha_m + \varepsilon_{\alpha_m},$$

ε_{α_m} being drawn from a normal distribution with zero mean and variance v_{α_m} ;

- Accept α'_m as the next value in the chain with probability A , this time using the Markov chain likelihood (contributions to which are given by equations (4.21) – (4.24)) in the calculation of A .

As before, v_{η_m} , v_{ξ_m} and v_{α_m} need to be ‘tuned’ to maximise the efficiency of the algorithm (and avoid chains which are too “hot” or “cold”). We found that using $v_{\eta_m} = 0.015$, $v_{\xi_m} = 0.03$ and $v_{\alpha_m} = 0.001$ seemed to work reasonably well for all m . The trace plots

Figure 6.4: Trace plots and posterior densities of logistic dependence parameters α_m for $m = 1$ (January) and $m = 7$ (July)

m		1	2	3	4	5	6
Posterior	mean	0.346	0.368	0.429	0.360	0.354	0.449
	(st. dev.)	(0.016)	(0.018)	(0.019)	(0.018)	(0.018)	(0.019)
m		7	8	9	10	11	12
Posterior	mean	0.424	0.352	0.294	0.425	0.416	0.370
	(st. dev.)	(0.020)	(0.018)	(0.009)	(0.018)	(0.019)	(0.017)

Table 6.4: Posterior means (and standard deviations) of the logistic dependence parameter α_m

for α_1 and α_7 (shown in figure 6.4), whose convergence is reflected in plots for all other months, lend encouragement to the success of the MCMC algorithm employed. Below the trace plots we show plots of the corresponding posterior densities after deletion of the settling-in period (taken to be 5000 iterations). Table 6.4 summarises these results, as well as the posterior densities for all other months, showing the posterior means and standard deviations (in parentheses). The posterior means compare reasonably well with the maximum likelihood estimates for α_m (table 4.1); however, the maximum likelihood estimates seem to be consistently higher than the posterior means, with (slightly) smaller asymptotic standard errors (*c.f.* the posterior standard deviations). As in the likelihood analysis, α_m doesn't appear to depend much on the season, with posterior means for α_m reflecting moderate to strong dependence no matter what time of year.

Return level estimation

As usual, the posterior distributions for our parameters can now be used to obtain posterior distributions for various return levels. However, to allow for the fact that we are explicitly modelling the temporal dependence inherent in the data, and to incorporate our simulated values for α_m , we now need to solve

$$\sum_{m=1}^{12} \left[1 - \lambda_{u_m} \left[1 + \xi_m^{(k)} \left(\frac{q_r^{(k)} - u_m}{\exp(\eta_m^{(k)})} \right) \right]_+^{-1/\xi_m^{(k)}} \right]^{h_m \theta_m^{(k)}} = 12 - \frac{1}{r} \quad (6.14)$$

for $q_r^{(k)}$ at each iteration k , to obtain a posterior distribution for q_r . Thus, estimation is more complicated due to the fact that at each iteration, the extremal index θ_m must be estimated. However, using the simulation technique of section 4.3.1, θ_m can be estimated at each iteration by repeatedly simulating an extreme value Markov chain using the current simulated posterior value of α_m (and marginal GPD parameters η_m and ξ_m); then we can use the intervals estimator, for example, to repeatedly estimate θ_m . The mean of the sampling distribution for θ_m can then be used as our posterior draw for $\theta_m^{(k)}$. In this way, θ_m is not classed as a parameter of the model, but is estimated using the posterior draws for η_m , ξ_m and α_m . As chapter 5 showed, as far as return level inference is concerned, the choice of estimator for θ_m is largely unimportant, so we use the intervals estimator since there is no need for any arbitrary specification of declustering parameter prior to estimation.

Figure 6.5 and table 6.5 show the posterior densities and summaries (respectively) for the 10, 50, 200 and 1000-year return levels. As before, both the (posterior) mean and standard deviation *and* median and interquartile range are given; the severe asymmetry

Figure 6.5: Posterior densities for the 10, 50, 200 and 1000-year return levels in a Bayesian analysis of the Bradfield wind speed data (using Markov chain models for each month)

		Return period (years)			
		10	50	200	1000
Posterior	mean	96.484	103.864	109.648	116.727
	(st. dev.)	(2.383)	(4.512)	(7.736)	(12.158)
Posterior	median	96.213	103.143	108.152	113.553
	(IQR)	(2.95)	(5.3)	(8.7)	(12.5)

Table 6.5: Posterior summaries for the 10, 50, 200 and 1000-year return levels (using Markov chain models for each month)

of the posterior densities for these return levels suggests the median and interquartile range might be preferable summaries to use. Comparing these results to those in table 6.3, the increased precision of the Markov chain approach is obvious – all return levels have a smaller posterior standard deviation and interquartile range than those obtained under the standard threshold exceedance approach where extremal dependence was ignored. The summaries of posterior location match up closely with those found under the standard approach (table 6.3), and are in close agreement with the maximum likelihood estimates obtained using all threshold exceedances when temporal dependence was both ignored and explicitly modelled using Markov chains (table 4.2). Indeed, there is increased efficiency over the maximum likelihood approach, the posterior standard deviations for all return levels being smaller than the asymptotic standard errors attached to the maximum likelihood estimates.

Predictive return levels

As discussed in section 6.1.3, an appealing feature of a Bayesian analysis is the ability to incorporate parameter uncertainty and randomness in future observations through predictive distributions. Since, in the field of extremes, the return level is often the parameter of most practical importance, we can use the Bayesian framework to estimate *predictive return levels*. Equation (6.7) provides a practical way of approximating the posterior distribution of the predictive return level, $q_{r,\text{pred}}$. Specifically, if we use the Markov chain model methodology, then the r -year predictive return level, $q_{r,\text{pred}}$, is obtained by solving

$$\frac{1}{B} \sum_{k=1}^B \sum_{m=1}^{12} \left[1 - \lambda_{u_m} \left[1 + \xi_m^{(k)} \left(\frac{q_{r,\text{pred}} - u_m}{\exp(\eta_m^{(k)})} \right) \right]_{+}^{-1/\xi_m^{(k)}} \right]^{h_m \theta_m^{(k)}} = 12 - \frac{1}{r} \quad (6.15)$$

for $q_{r,\text{pred}}$.

Figure 6.6 shows a plot of return levels in which uncertainty due to both process variability and sampling variability has been accounted for (on a $-\log\{-\log(1 - 1/r)\}$ scale). The corresponding estimative return level, based on maximum likelihood estimates, is also included. Clearly, designing to a level determined by maximum likelihood estimates may lead to substantial underprotection due to uncertainty in parameter estimation, which is accounted for in the predictive analysis. Table 6.6 reports some of these predictive return levels and shows, for comparison, return level estimates based on some other methods used so far in this thesis. We show the posterior mean and median for the 10, 50, 200 and 1000-year return levels in a Bayesian analysis which uses Markov chain models to account for the dependence in the series. We also show maximum likelihood estimates

Figure 6.6: Estimative (thin line) and predictive (bold line) return level curves for Bradfield

	Return period (years)			
	10	50	200	1000
Predictive distribution	100.711	111.963	124.283	144.939
Posterior mean (modelling dependence)	96.484	103.864	109.648	116.727
Posterior median (modelling dependence)	96.213	103.143	108.152	113.553
MLE (all excesses, modelling dependence)	96.745	103.236	108.152	113.306
MLE (all excesses, ignoring dependence)	96.785	103.180	108.085	113.249
MLE (cluster peak excesses)	94.916	100.469	104.950	114.954
Profile likelihood upper 95% point	100.35	107.91	113.10	118.79

Table 6.6: Predictive return levels for Bradfield, compared with estimates based on other methods used

obtained in earlier chapters, which use all excesses and cluster peak excesses in their estimation. In practice, even when using maximum likelihood estimates, some attempt is made to account for parameter uncertainty by designing to the upper end points of confidence intervals; to this end, we also report the upper end point of the 95% profile likelihood (chapter 2). As shown by table 6.6, designing to the maximum likelihood estimate which uses cluster peak excesses – a common approach used in practice – would result in considerable underprotection. Even designs which use the upper end point of the profile likelihood completely under-estimate relative to the predictive analysis.

6.3.2 Bayesian methods for the extremal index

When using Markov chain models for threshold exceedances, we have discussed how the extremal index can be used to ensure that the temporal dependence we are modelling is properly accounted for in our estimation of return levels. The estimated logistic dependence parameter α is used to simulate an extreme value Markov process, and then the extremal index is estimated from this. In the previous section, and in chapter 4, it was necessary to estimate the extremal index in order to be able to estimate return levels. Chapter 5 investigated the consequences of using different techniques for estimating the extremal index on return level estimation; we concluded that the choice of estimator was an unimportant one as far as return levels were concerned. However, the extremal index is also of great interest in its own right; it is the key parameter when extending discussions of the limiting behaviour of extreme values from independent and identically distributed sequences to stationary sequences (see section 1.4.1). Indeed, the extremal index has been the focus of much recent theoretical and practical research – see Ferro and Segers (2003), Ancona-Navarrete & Tawn (2000), Smith & Weismann (1994) and Smith (1992) for a review of some of this work – none of which has been in the Bayesian framework.

This section attempts to look at the extremal index in a Bayesian setting. Although the previous section used the extremal index to obtain posterior distributions for return levels and predictive return levels, the extremal index was not a part of the MCMC – it was simply estimated using the posterior distribution for the logistic dependence parameter α . We now attempt to specify an appropriate prior distribution for the extremal index θ and estimate this with other parameters simultaneously. We use simulated data to assess the accuracy of our inference, drawing on the results of Smith (1992) for the true values of θ for extreme value Markov chains with different (logistic) dependence parameters α .

Inference using a maxima method

Section 5.2.2 outlined two procedures for estimating the extremal index without resorting to declustering. A maximum likelihood estimator (developed by Ancona-Navarrete and Tawn) was described, the log-likelihood function for which is given by equation (5.11). The original idea, proposed by Gomes (1993), suggests fitting the GEV (with parameters μ_θ , σ_θ and ξ_θ) to the set of block maxima; we then fit another GEV (with parameters μ , σ and ξ) to a set of block maxima obtained from the original series after randomising the index of the observations. Thus the second fit should be to a series of block maxima which have the same marginal distribution as the dependent series, but which are truly independent. Subsequent adjustments are then made to these two sets of estimates to obtain an estimate of θ . Ancona-Navarrete and Tawn (2000) suggest a method which simultaneously estimates the parameter vector $\psi = (\mu, \sigma, \xi, \theta)$ by treating both sets of block maxima (together) as independent GEV random variables, leading to the (joint) log-likelihood function given in equation (5.11). Instead of using (5.11) to obtain maximum likelihood estimates for μ , σ , ξ and θ , we can use the log-likelihood function in a Bayesian setting. Again, specifying independent, non-informative priors for the parameters, we can use a Metropolis-Hastings scheme (with random walk updates) to sample from the posteriors. Normal priors with zero mean and variances 10000, 10000, and 100 are used for μ , $\eta = \log(\sigma)$ and ξ (respectively); since the extremal index $\theta \in (0, 1]$, we specify a uniform $U(0, 1)$ prior for θ .

Figure 6.7 shows the trace plot and associated posterior density for the extremal index for a simulated dataset; table 6.7 summarises these plots, as well as the posterior distributions for the other estimable parameters. The data were simulated in accord with simulated data in other chapters, i.e. an extreme value Markov chain, the joint distribution of consecutive pairs being given by the bivariate (symmetric) logistic model with dependence parameter α . In this case, data were simulated such that $\alpha = 0.5$; the results of Smith (1992), summarised in table 5.4, show such a series to have an extremal index of 0.328. We see from figure 6.7 and table 6.7 that the Bayesian inference performs quite well, with a posterior mean of 0.341. Indeed, the true value of 0.328 lies within two (posterior) standard deviations of this mean value. The gain in efficiency over the standard maximum likelihood analysis can be seen if we compare the posterior standard deviation to the sampling distribution standard deviation of the maximum likelihood estimator (given in table 5.5); these are 0.025 and 0.042 (respectively). The other parameters, though not of (direct) interest here, also have posterior means close to their true values.

Figure 6.7: MCMC realisations of the extremal index for simulated data

		μ	σ	ξ	θ
Posterior	mean	5.450	3.025	0.279	0.341
	(st. dev.)	(0.031)	(0.094)	(0.025)	(0.025)

Table 6.7: Summary statistics for the posterior location, scale, shape and extremal index

Inference using the intervals estimator

Ferro and Segers (2003) describe two models for the inter-arrival times used in the calculation of their intervals estimator for θ . They call the first a “limiting model”, where they show the distribution of the inter-arrival times, T_i , to be the mixture distribution

$$(1 - \theta)\epsilon_0 + \theta\zeta_\theta, \quad (6.16)$$

where ϵ_0 is the degenerate probability distribution at 0 and ζ_θ is the exponential distribution with rate θ . Writing $t_i = NT_i/n$, where N is the number of exceedances and n is the total number of points in the process, the log-likelihood from this limiting model is

$$\ell(\theta) = 2(N - 1)\log\theta - \theta \sum_{i=1}^{N-1} t_i. \quad (6.17)$$

The resulting maximum likelihood estimator for θ is $\min\{1, 2/\bar{t}\}$, where $\bar{t}$ is the mean of the t_i . Ferro and Segers discuss that this estimator tends in distribution to one as n increases: a failure arising from the model assigning all of the inter-arrival times to the exponential component of the mixture distribution.

The second model is a “penultimate approximation” to the limiting model, where, to a first-order approximation, the distribution of inter-arrival times is given by

$$\Pr(T > n) = \theta(1 - \lambda_u)^{n\theta}, \quad (6.18)$$

where $\lambda_u \in (0, 1)$ may be thought of as the exceedance rate of the threshold u . The log-likelihood from this model is given by

$$\begin{aligned} \ell(\theta) = & a_1 \log [1 - \theta(1 - \lambda_u)^\theta] \\ & + \{ \log\theta + \log [1 - (1 - \lambda_u)^\theta] \} \sum_{i \geq 2} a_i + \theta \log [1 - \lambda_u] \sum_{i \geq 2} (i - 1)a_i, \end{aligned} \quad (6.19)$$

where a_i is the number of arrival times equal to i . Maximum likelihood estimates can be found by numerical optimisation.

As before, we use a simulated first-order extreme value Markov chain, with logistic dependence parameter $\alpha = 0.5$, to examine the performance of the MCMC for the extremal index. We consider both models for the inter-arrival times of threshold excesses – the “limiting” model (equation 6.16) and the “penultimate” model (equation 6.18) – in turn, thus obtaining two separate MCMC outputs. The results are shown in figures 6.8 – 6.9 and 6.10 for the “limiting” and “penultimate” models respectively. For the “limiting” model, the length of the simulated dataset is varied, from $n = 100$ to $n = 1000$, to

demonstrate the problem with this model; as expected, the posterior estimates of θ tend to one as the length of the process increases (recall that this process has an extremal index of $\theta = 0.328$). Figures 6.8 and 6.9 illustrate this problem, with the trace plots and posterior densities clearly converging to one as n increases. Table 6.8 also supports this, with posterior means (after “burn-in”) of 0.909, 0.934, 0.959 and 0.987 for $n = 100, 250, 500$ and 1000 respectively. Further, the posterior standard deviations suggest convergence to 1, decreasing in size as n increases, from 0.075 when $n = 100$ to 0.012 when $n = 1000$. The “penultimate” model solves this problem, and (as can be seen from figure 6.10 and table 6.9) gives posterior draws which are close to the true value of $\theta = 0.328$.

Comparing the Bayesian analyses for θ using the maxima method and the models for the inter-arrival times of threshold exceedances, we can conclude that analyses based on the “penultimate” model for inter-arrival times provide the best inferences, with a posterior mean closer to the true value of $\theta = 0.328$ than that given by both the “limiting” model and the maxima method. This approach also seems to give the most precise inference, with the smallest posterior standard deviation for θ . Inferences using the maxima method are reasonable, with a posterior mean for θ of 0.341 (table 6.7). However, the “limiting” model for the inter-arrival times of threshold exceedances has been shown to be flawed, with a posterior distribution tending to 1 as the size of the simulated dataset increases.

6.3.3 A Bayesian declustering scheme for dependent extremes

While we have discussed that, as far as GPD parameter estimation and return level inference are concerned, using all threshold excesses in an analysis of extremes (either by ignoring dependence or explicitly modelling dependence) is preferable to filtering out a set of independent threshold excesses using a declustering technique, estimating other cluster characteristics may require clusters to be identified. For example, suppose we are interested in making inferences about a cluster functional, H . This could be the ‘excess height statistic’ (Leadbetter, 1995) or one of the functionals discussed in Smith *et al.* (1997) relating to the peak or duration of extreme events, such as the number of threshold upcrossings per cluster or the average cluster length. In such situations, we need to employ a technique which can systematically identify clusters of threshold exceedances in a sensible way. Cluster functionals can also be used as an informal tool to assess (and compare) the fit of non-nested statistical models, where more formal tests based on the likelihood ratio might not be appropriate; see section 4.5.

Most declustering schemes proposed in the literature require an auxiliary parameter, the

Figure 6.8: MCMC realizations of θ using the “limit” model

	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
Mean	0.909	0.934	0.959	0.987
Standard deviation	0.075	0.060	0.038	0.012
Median	0.927	0.950	0.956	0.991
Inter-Quartile Range	0.103	0.074	0.044	0.015

Table 6.8: Summary statistics for the posterior estimates of the extremal index using the “limit” model

Figure 6.9: Posterior densities for the extremal index (using the “limiting” model), — $n = 100$; — $n = 250$; — $n = 500$; - - - $n = 1000$

Figure 6.10: MCMC realizations of θ using the “penultimate” model

	θ
Posterior mean	0.32841
Posterior st. dev.	0.05914
Posterior median	0.32610
Posterior IQR	0.0779

Table 6.9: Summary statistics for the posterior distribution of θ using the “penultimate” model

choice of which is largely arbitrary; for example, runs declustering requires a cluster termination interval (κ), and blocks declustering requires the specification of a block length (τ), prior to declustering. Section 3.5.2 introduced the method of “automatic” declustering of extreme values, based on Ferro and Segers’ intervals estimator for θ , which can be justified by an asymptotic result for the arrival times between threshold exceedances. It was explained that “automatic” declustering was equivalent to runs declustering, but without the need to arbitrarily choose the cluster identification parameter κ . Briefly, the procedure is:

- Estimate the extremal index using the intervals estimator (equation 3.34);
- Calculate $C = \lfloor \hat{\theta}N \rfloor + 1$, where N is the total number of threshold exceedances;
- Find the largest C inter-arrival times;
- Use these inter-arrival times to partition the series into clusters of threshold exceedances.

This is equivalent to runs declustering with cluster identification parameter κ set equal to the C -th largest inter-arrival time.

Now in a Bayesian setting, consider a vector of posterior estimates for θ , i.e. $\theta^{(1)}, \dots, \theta^{(R)}$, where R is the number of iterations. Then we can find $C^{(k)} = \lfloor \theta^{(k)}N \rfloor + 1$ for $k = 1, \dots, R$. From this, we can obtain a vector of posterior draws for the declustering parameter κ , i.e. $\kappa^{(1)}, \dots, \kappa^{(R)}$, by setting $\kappa^{(k)}$ equal to the $C^{(k)}$ -th largest inter-arrival time at iteration i . For example, if, at iteration 1, we use the likelihood for the “penultimate” model to obtain a posterior draw $\theta^{(1)}$ for θ , and from this find $C^{(1)} = 6$, then we find the six largest inter-arrival times of threshold exceedances of the process. We then decluster using runs declustering with $\kappa^{(1)}$ being equal to the smallest of these six inter-arrival times. We can then repeat this procedure for each posterior draw for θ to obtain a sequence of posterior declustering parameters $\kappa^{(k)}$ associated with each $\theta^{(k)}$, and look at characteristics of the posterior distribution for κ . Thus, we can advise on the choice of declustering parameter κ depending on the estimated extremal index θ , the Bayesian analysis allowing us to express our uncertainty about κ (which cannot be achieved in the frequentist setting) through the construction of posterior confidence intervals for κ . We now demonstrate this Bayesian declustering scheme using both simulated data and the real-life wind speed data at Bradfield, and use the scheme to estimate certain cluster functionals of interest.

Simulated data

As before, we use a simulated first-order Markov chain using the bivariate logistic model (equation 4.15) for the joint distribution of consecutive pairs in the series. We use $\alpha = 0.8$, $\alpha = 0.5$ and $\alpha = 0.2$ to simulate series with different strengths of temporal dependence (‘weak’, ‘moderate’ and ‘strong’ dependence respectively) to get an idea of how the declustering parameter κ might change with dependence. The resulting MCMC trace plots and posterior densities for κ (after “burn-in”) are shown in figure 6.11, with numerical summaries given in table 6.10. The associated trace plots for θ are not shown; the trace plot for θ for the case $\alpha = 0.5$ is shown in figure 6.10, and similar output was obtained for θ for $\alpha = 0.2$ and $\alpha = 0.8$. The MCMC was run for 10,000 iterations, with the first 1000 realisations discarded as “burn-in”.

Comparing the posterior means for θ with the true values of θ for extreme value Markov chains (table 5.4), we see that, for $\alpha = 0.5$, the posterior mean (0.328) is *exactly* (to 3 decimal places) the true value for such a process (the results for $\alpha = 0.5$ are the same as those given in table 6.9). The posterior means of θ for the other two processes ($\alpha = 0.8$ and $\alpha = 0.2$) seem reasonable, though the true value of θ when $\alpha = 0.2$ is 0.0616, which is slightly lower than our posterior mean of 0.136. The true value of θ when $\alpha = 0.8$ is unknown (Smith (1992) calculates θ for a range of values of α between 0.2 and 0.5 only), though extrapolating beyond $\alpha = 0.5$ suggests a value of θ close to the posterior mean of 0.797 seems sensible. The similarity between posterior means and medians suggests the posterior distributions are reasonably symmetric, so both the mean and median provide representative location summaries for the posteriors. Of greater interest here, however, are the posterior densities for the declustering parameter κ . From table 6.10 we see that, for processes with first-order logistic dependence $\alpha = 0.8$, 0.5 and 0.2, we have corresponding posterior means for κ of 4.032, 14.035 and 33.162 respectively. This suggests that, for processes with these three strengths of dependence, cluster termination intervals of 4, 14 and 33 (respectively) will be sufficient to identify sets of independent clusters of extremes (since the cluster termination interval κ must take an integer value). These conclusions seem sensible, since we would expect to use larger cluster termination intervals for processes with stronger temporal dependence, to ensure clusters are far enough apart to be deemed independent; a series with much weaker dependence should not require such a large interval to guarantee independence between successive clusters of extremes. In any practical application, then, where it is *necessary* to identify clusters of extremes, this Bayesian declustering scheme could be employed not only to find a suitable declustering parameter κ , but also to gain insight into the variability of this declustering parameter, and subsequently any cluster functional H .

Figure 6.11: MCMC realisations of the declustering parameter κ for three strengths of temporal dependence

		$\alpha = 0.8$ ($\theta \approx 0.775$)		$\alpha = 0.5$ ($\theta = 0.328$)		$\alpha = 0.2$ ($\theta = 0.0616$)	
		θ	κ	θ	κ	θ	κ
Posterior	mean	0.797	4.032	0.328	14.035	0.136	33.162
	(st. dev)	(0.062)	(2.097)	(0.060)	(2.381)	(0.030)	(2.571)
Posterior	median	0.8001	4.000	0.326	14.000	0.134	33.000
	(IQR)	(0.088)	(2.000)	(0.078)	(4.000)	(0.039)	(4.000)

Table 6.10: Posterior summaries for θ and κ for various strengths of dependence

Application to Bradfield wind speed data

We now apply the Bayesian declustering scheme to the Bradfield wind speed data, and use the posterior distribution of κ to obtain posterior distributions for two cluster functionals commonly studied: (1) the mean cluster size, and (2) the mean number of threshold upcrossings per cluster.

The mean cluster size (H_M) is a cluster functional often studied in hydrology, where it is interpreted as the average duration of a storm. Obviously, this functional can be calculated in any standard likelihood-based analysis; runs declustering with a pre-determined cluster termination interval κ can be used to identify clusters of threshold exceedances, and the mean cluster size obtained empirically. However, κ is chosen arbitrarily and so does not take into account the extremal dependence of the process. Using our Bayesian declustering scheme, the cluster termination interval κ is found via estimation of the extremal index θ , for which we have a posterior distribution. Thus, for each posterior draw $\theta^{(k)}$, $k = 1, \dots, R$ (where R is the number of iterations), we have a corresponding declustering parameter $\kappa^{(k)}$, and so can obtain an associated estimate of the cluster functional $H_M^{(k)}$. So instead of a point estimate for the mean cluster size H_M , we have knowledge about the variability of H_M through the posterior distribution, and so have a more complete inferential procedure. This has even greater appeal if we can quantify any prior knowledge about the extremal dependence of the process through the prior distribution for θ . Figure 6.12 shows the MCMC output for the extremal index θ (again, using a non-informative $U(0, 1)$ prior for θ), the declustering parameter κ and mean cluster size H_M for the Bradfield wind speed data (results shown for January); table 6.11 summarises these results.

We see that the posterior mean for κ_1 is 29.259, suggesting that for this process, declustering using the runs method would require a cluster termination interval of about 29 or 30 hours to ensure successive clusters are independent. With a posterior mean for the extremal index of 0.196, this seems sensible, and is consistent with the results for simulated data shown in table 6.10; here, posterior means for the declustering parameter κ were 14.035 and 33.162 corresponding to posterior means for the extremal index of 0.328 and 0.136 respectively. Both posterior distributions for θ_1 and κ_1 seem symmetric, with posterior means and medians being similar in value in both cases. The posterior distribution for the functional H_M (the mean cluster size) indicates that the mean cluster size is between 8 and 20. In a meteorological or hydrological setting, this would suggest that we can expect the mean length of storms in January to be between 8 and 20 hours. This is also apparent from the 95% confidence interval taken from the posterior distribution,

Figure 6.12: MCMC realisations of the extremal index (θ), declustering parameter (κ), and two cluster functionals – the mean cluster length (H_M) and the mean number of threshold upcrossings per cluster (H_U) for the Bradfield wind speed data (January)

		θ_1	κ_1	H_M	H_U
Posterior	mean	0.196	29.259	12.800	1.670
	(st. dev)	(0.026)	(9.790)	(1.697)	(0.354)
Posterior	median	0.195	32.000	12.800	1.670
	(IQR)	(0.035)	(10.000)	(1.940)	(0.404)
	95% CI	(0.148, 0.251)	(9.000, 48.000)	(9.879, 16.300)	(1.061, 2.431)

Table 6.11: Posterior summaries of the extremal index, declustering parameter, and two cluster functionals – mean cluster length (H_M) and mean number of threshold upcrossings per cluster (H_U)

ranging from just under 10 hours to just over 16 hours. From this, we can also see that the distribution of H_M is positively skewed, with the median value lying closer to the lower quartile than the upper quartile.

We now investigate another cluster functional – the mean number of threshold upcrossings per cluster – which we call H_U . Again, we repeatedly decluster the data using each draw from the posterior distribution for κ_1 , calculating the average number of threshold upcrossings per cluster at each step, to obtain a posterior distribution for H_U . The results are shown (and summarised) alongside the results for θ_1 , κ_1 and H_M in figure 6.12 (and table 6.11). The posterior mean for H_U is 1.670, suggesting that most clusters have, on average, only one or two threshold upcrossings. This can be explained by the strong temporal dependence present in the process. In processes which exhibit weaker short-term temporal dependence, successive within-cluster observations are more likely to alternate above and below the threshold, and so such processes are more likely to have a higher number of within-cluster threshold upcrossings. However, processes in which successive observations are more strongly related (as in the Bradfield wind speed data) are less likely to have as many within-cluster threshold upcrossings; once an observation has fallen sub-threshold, this is more likely to terminate a cluster since its successors are also likely to fall sub-threshold. However, in some instances less than κ_1 observations will fall below the threshold before another observation exceeds the threshold, and so some clusters will have one or more threshold upcrossings – though this is likely to occur with greater frequency when a process is less serially correlated. In meteorological terms, the posterior distribution for H_U suggests that, during a storm, periods of relative calm (i.e. times when a within-cluster observation falls below the threshold) will be infrequent, only occurring, on average, between once and three times (of course, this is for the Bradfield data in January only).

The more complete Bayesian inference allows us to understand the variability of storm duration and mean number of threshold upcrossings, and obtain plausible intervals for such cluster functionals.

6.4 Discussion

The principal aim of this chapter was to investigate the use of Bayesian techniques for extremes. The potential benefits of working within the Bayesian framework over a standard likelihood-based analysis are compelling, particularly when working with extreme values, which are (by their very nature) scarce; any extra information which can augment our data through a prior distribution, for example, could greatly enhance our inference and improve the efficiency of our analysis. The primary objective of an extreme value analysis is often prediction, i.e. given the history of a process (and other external information), the main interest is in estimating the extremal characteristics over some future epoch of the process. Again, the Bayesian methodology caters for such estimation through the predictive distribution, which allows for uncertainty in the estimation of the parameter components. A Bayesian analysis also delivers a more complete inference than the point estimates obtained under the maximum likelihood approach via the posterior distribution; parameter variability can be quantified using simple posterior summaries, instead of relying upon asymptotic theory. Indeed, as has already been discussed, in some cases maximum likelihood estimates are non-regular, or even unavailable – for example, when $\xi < -0.5$ or $\xi < -1$ (respectively) – to this problem, the Bayesian framework provides an attractive solution.

The two illustrative examples, which showed the MCMC “in action” for simulated data and the Bradfield wind speed data, highlight the main advantages of the Bayesian analysis over the standard likelihood-based inference: increased accuracy and precision. Comparing the posterior distributions of the GPD parameters (and the 1000-observation return level) with the sampling distributions obtained from the simulation study in chapter 3, the posterior means were closer to the true parameter values than were their sampling distribution counterparts, and the posterior standard deviations were consistently smaller than the standard deviations obtained from the sampling distributions. In contrast, in the analysis of the Bradfield wind speed data, the posterior standard deviations of the GPD parameters were very similar to the asymptotic standard errors in the standard likelihood-based analysis, and sometimes even larger (comparing tables 6.2 and 2.2), suggesting a loss of precision in the Bayesian analysis. However, the maximum likelihood analysis assumes independence of successive observations so the standard errors should be inflated to account for this. Table 2.4 reports these adjusted standard errors and, after this adjustment, we see that the Bayesian posterior standard deviations are consistently (and substantially) smaller than the standard errors attached to the maximum likelihood estimates. Thus, the gain in efficiency in both examples is obvious.

The Markov chain model proposed in chapter 4 was shown to increase the precision of return level estimation compared to the approaches which ignore or remove temporal dependence; we have further improved this inference by fitting the Markov chain model in the Bayesian set-up. The resulting analysis was shown to improve return level inference by considerably reducing the variability of our estimates (*c.f.* the approach which ignores temporal dependence and uses all threshold excesses). A predictive analysis of the Bradfield wind speed data was also carried out, allowing us to account for all sources of possible uncertainty.

To further the work carried out in chapter 5, we have also focused on Bayesian methods for the extremal index. We have exploited the likelihoods for Ancona–Navarrete and Tawn’s extremal index estimator, where the extremal index is made a model parameter and is jointly estimated alongside the GEV parameters, as well as the likelihoods for Ferro and Segers’ two models for the inter-arrival times of threshold exceedances. We found that using Ferro and Segers’ “penultimate” model for the exceedance inter-arrival times proved very successful, with a posterior mean for θ very close to the true value. We then proposed a Bayesian declustering scheme based on the posterior distribution for the extremal index, which was shown to open up the possibility of obtaining posterior distributions for various cluster functionals. This was demonstrated using two such functionals commonly of interest in the hydrological literature: the mean cluster size and the mean number of threshold upcrossings per cluster. Such analyses allow greater insight into the properties of storms within particular months.

The Bayesian analyses throughout this chapter have used largely non-informative prior distributions with large variances. Even so, we have seen how the Bayesian methodology can help to improve inference. Further improvements could be sought if any prior information regarding the extremes could be obtained; however, how such information is best elicited for extremes is often problematic. Future work could involve investigating the inclusion of such ‘expert’ prior information; for more information, see Coles and Tawn’s analysis of extreme rainfall data (1996). A possible oversight in the Bayesian analyses of this chapter is the fact that the GPD scale parameter σ is dependent on the choice of threshold u ; hence, an uninformative prior for σ at one threshold becomes informative at other thresholds. One way of circumventing this problem, whilst at the same time avoiding a block maxima approach (which is wasteful of data), is to work with a point process model wherein the parameters are independent of the threshold level. However, a simpler approach is considered in the following chapter, where the GPD is reparameterised to give parameters which are threshold-independent.

Chapter 7

A hierarchical model for extreme wind speeds

7.1 Introduction

Complex relationships which permeate many real-life datasets often remain undetected in standard statistical analyses. However, the recent explosion in simulation-based inference has enabled statistical models of far greater complexity to be built and estimated, paving the way for a statistical analysis which identifies, and even quantifies, such relationships. In chapter 2, a simple extreme value analysis of the Bradfield wind speed data was demonstrated, and the use of all threshold excesses and cluster peak excesses to fit the GPD to each season (month) separately was compared. Chapter 4 then described how we can reflect any temporal dependence present in the process through the likelihood, by using Markov chain models for short-term serial correlation. However, seasonal variation was still dealt with by fitting separate models to each month. We now develop a hierarchical model for the hourly gust maximum wind speed data which attempts to identify, and quantify, the seasonal effects present in the data instead of fitting a different model to each month. Thanks to the U.K. Meteorological Office, we now also have corresponding wind speed data for eleven other sites (see figure 7.1) across England, allowing the potential for a model which identifies site effects (compared to a site-by-site analysis which makes no attempt to quantify any such effects) as well as seasonal effects. Temporal dependence can be incorporated into the analysis (as before) via Markov chain models, invoking bivariate extreme value theory for the joint distribution of consecutive pairs of the process. Given the scarcity (and hence the value) of extremes, even limited improvements in efficiency as a result of such structured statistical modelling could lead to significant improvements in any practical application. The aim of this chapter is to explore the feasibility and robustness of such hierarchical models in extreme value analyses,

Figure 7.1: Location of wind speed stations

in the context of the U.K. wind speed data. The results of chapter 4, in which a separate Markov chain model was fitted to each month, are used as a baseline to compare with the results of a fully Bayesian fixed effect analysis in which site and seasonal effects, as well as temporal dependence, are identified.

7.2 The wind speed data

We now have available data from eleven other wind speed stations across England; the location of all these stations, including High Bradfield, is shown on the map in figure 7.1. Table 7.1 gives the mean gusts for all twelve sites in January, April, July and October. The “DCNN” code is the wind speed station identifier, the first and second digits representing “district” and “county” (respectively), the last two digits giving the station number within that county. Thus, Leeming and Linton on Ouse are both in the same district and county, Leeming being station 45 and Linton on Ouse station 55. Excluding Bradfield and Sheffield, whose datasets span a period of ten years, the hourly gust maximum wind speeds have been recorded over a period of eighteen years, from 1st January 1974 (which is also the start of the Bradfield and Sheffield datasets) to 31st December 1991 (inclusive), constituting approximately 157,000 observations at each site. These sites represent a variety of geographical locations – both urban and rural, high and low altitudes, and both east and west facing sites. This range of geographical conditions will

Figure 7.2: Time series plots (first three years), histograms and box-plots of gusts at Bradfield and Nottingham

Site	DCNN	j	January	April	July	October
Sheffield	4062	1	20.605	16.532	14.434	15.730
Bradfield	4063	2	28.802	20.614	18.314	23.210
Leeming	2245	3	18.965	15.962	14.321	16.033
Linton on Ouse	2255	4	17.238	15.247	13.250	15.062
Scampton	2416	5	17.706	15.001	13.427	15.483
Cranwell	2432	6	19.124	16.022	14.701	16.071
Bingley	4036	7	20.995	15.923	15.188	17.180
Nottingham	4206	8	17.902	15.595	14.066	14.998
Finningley	4238	9	17.645	14.911	13.961	14.916
Keele	4617	10	17.201	14.657	12.839	14.517
Aigburth	7221	11	18.160	15.197	14.238	15.335
Ringway	7377	12	17.453	15.098	13.355	14.858

Table 7.1: Mean gust speeds (knots) observed at the twelve wind speed stations

obviously influence the strength of winds observed at any particular site. For example, wind movements are modified by cityscapes because buildings increase the friction on air flowing over and around them, slowing the speed of winds by as much as 80% in some city centres (de Blij and Müller, 1996). Also, locations at higher altitudes, being more exposed, often observe stronger wind gusts than those closer to sea-level.

Of the twelve sites being studied, sites 4063 (High Bradfield) and 4206 (Nottingham) most noticeably demonstrate these differences. Figure 7.2 shows time series plots and histograms of the gusts observed at these sites, as well as plots of the time series against the lagged series (the plots for Nottingham and Bradfield show seasonal variation and temporal dependence similar to those observed at the other ten sites). High Bradfield lies 380 metres above sea-level. The anemometer is in an isolated location in the Peak District National Park, and is not surrounded by any buildings or other structures which could reduce wind speeds observed at the station. In contrast, Nottingham is comparatively low-lying, being just 50 metres above sea-level. Nottingham is also a city, surrounded by many tall buildings and smaller towns, which induces a micro-climate in the city centre with lower wind speeds being observed than in surrounding areas. Such physical characteristics force considerably lower wind speeds to be observed at Nottingham than at Bradfield, as seen in figure 7.2. For a detailed description of the behaviour of the wind climate in the U.K., see Shellard (1976).

Section 2.3 discussed the seasonal nature of the wind climate in the U.K., with the winter months (usually October – March) observing the strongest wind gusts. Indeed, it is very uncommon for structural damage caused by the wind to occur outside this period. We also discussed the importance of including summer extremes in the analysis; although such gusts are not extreme relative to a gust in, say, January, the same mechanism (an alternating sequence of anticyclones and depressions) is responsible for the generation of storms throughout the year, and so extremes from summer months can help us to understand what happens in winter months (when genuinely large events *can* occur). The time series plots in figure 7.2 illustrate the seasonal nature of the wind speed data at Bradfield and Nottingham, with similar cycles being observed at all other sites. Table 7.1, showing mean gust speeds at the twelve sites in January, April, July and October, also demonstrates this seasonal variation. All sites have their highest mean gust speeds in January and their lowest in July (or August), with values for April and October being similar to each other; this suggests a cyclic nature in gust speeds, decreasing throughout summer months towards July and August, and then increasing throughout the winter towards January.

7.3 A hierarchical model

Chapter 4 investigated the use of Markov chains to model short-term temporal dependence in the Bradfield wind speed data. Until now, seasonal effects have been removed by fitting a separate model to each season of data. As section 7.2 discussed, we now also have the added complexity of possible site effects. One solution to deal with such effects (rather than ignore or remove them) involves the construction of a model which allows a pooling of information across the sources of these effects (in this case sites and seasons). The feasibility and robustness of such models in extreme value analyses has been, to date, largely unexplored.

Typically, site and seasonal variation are dealt with by fitting a separate model to each season within each site. This ‘site-by-site, separate seasons’ approach, using the GPD as a model for threshold excesses, would return parameter estimates $(\hat{\sigma}, \hat{\xi})$ for each month within each site. In this chapter, the results of such an analysis (as carried out in chapters 3 and 4 for Bradfield) are used as a baseline to compare with the results of a fully Bayesian fixed effects analysis in which site and seasonal effects, as well as temporal dependence, are quantified. For all seasons within each of the twelve sites, mean residual life plots (section 2.5) have been used to aid the threshold selection procedure.

7.3.1 Model specification

We denote by $(\tilde{\sigma}_{m,j}, \xi_{m,j})$ the parameters of the GPD model that is assumed to be valid for threshold excesses in season (month) m and site j . We use $\tilde{\sigma}_{m,j} = \sigma_{m,j} - \xi_{m,j}u_{m,j}$ in place of the usual scale parameter $\sigma_{m,j}$ to avoid the problem of threshold dependency; the GPD shape parameter $\xi_{m,j}$ is constant above a suitably high threshold $u_{m,j}$, but the scale parameter changes with $u_{m,j}$ unless $\xi_{m,j} = 0$. The reparameterised scale parameter $\tilde{\sigma}_{m,j}$ remains constant at all thresholds $u'_{m,j} > u_{m,j}$. In a Bayesian setting, the usual parameterisation might be restrictive, since an uninformative prior for $\sigma_{m,j}$ becomes informative at other thresholds. Using the Markov chain model set-up, we also model the joint distribution of consecutive pairs of the process using the logistic model (equation (4.15)). However, given the similarity between estimates of α across months (table 4.1), we allow α to vary across sites only; thus, we denote by α_j the logistic dependence parameter α in site j . Allowing for seasonal and site effects for $\tilde{\sigma}$ and ξ , and site effects only for α , we specify the following fixed effects model (in accord with Coles, 2001b):

$$\log(\tilde{\sigma}_{m,j}) = \gamma_{\tilde{\sigma}}^{(m)} + \epsilon_{\tilde{\sigma}}^{(j)}, \quad (7.1)$$

$$\xi_{m,j} = \gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)} \quad \text{and} \quad (7.2)$$

$$\alpha_j = \epsilon_{\alpha}^{(j)}, \quad (7.3)$$

where γ and ϵ represent seasonal and site effects respectively. We work with $\log(\tilde{\sigma}_{m,j})$ for computational convenience, and to retain the positivity of the scale parameter $\tilde{\sigma}_{m,j}$. All effects for $\log(\tilde{\sigma}_{m,j})$ and $\xi_{m,j}$ are assumed to be normally and independently distributed:

$$\gamma_{\tilde{\sigma}}^{(m)} \sim N_0(0, \tau_{\tilde{\sigma}}) \quad \text{and} \quad (7.4)$$

$$\gamma_{\xi}^{(m)} \sim N_0(0, \tau_{\xi}), \quad m = 1, \dots, 12, \quad (7.5)$$

for seasonal effects, and

$$\epsilon_{\tilde{\sigma}}^{(j)} \sim N_0(a_{\tilde{\sigma}}, \zeta_{\tilde{\sigma}}) \quad \text{and} \quad (7.6)$$

$$\epsilon_{\xi}^{(j)} \sim N_0(a_{\xi}, \zeta_{\xi}), \quad j = 1, \dots, 12, \quad (7.7)$$

for site effects, where $N_0(A, B)$ is the normal distribution with mean A and *precision* B (used for notational convenience). We choose the mean of the normal distribution of the seasonal effects to be fixed at zero to avoid over-parameterisation and problems of identifiability; however, we could equally have fixed the mean for the distribution of site effects to achieve this. Since the logistic dependence parameter α must lie between 0 and 1, we draw the site effect for α from a uniform distribution, and so

$$\epsilon_{\alpha}^{(j)} \sim U(0, 1). \quad (7.8)$$

The final layer of the model is to specify prior distributions for the random effect distribution parameters. Here, we have chosen largely non-informative priors, adopting conjugacy wherever possible to simplify computations. Thus,

$$a_{\tilde{\sigma}} \sim N_0(b_{\tilde{\sigma}}, c_{\tilde{\sigma}}), \quad a_{\xi} \sim N_0(b_{\xi}, c_{\xi}), \quad (7.9)$$

$$\tau_{\tilde{\sigma}} \sim Ga(d_{\tilde{\sigma}}, e_{\tilde{\sigma}}), \quad \tau_{\xi} \sim Ga(d_{\xi}, e_{\xi}), \quad \text{and} \quad (7.10)$$

$$\zeta_{\tilde{\sigma}} \sim Ga(f_{\tilde{\sigma}}, g_{\tilde{\sigma}}), \quad \zeta_{\xi} \sim Ga(f_{\xi}, g_{\xi}), \quad (7.11)$$

with a suitable specification of hyper-parameters.

7.3.2 Inference

The most feasible way of estimating the model outlined in 7.3.1 is through Markov chain Monte Carlo. ‘One level down’ from the data, conjugate priors have been chosen for the parameters of the fixed effects distributions, and so here, the full conditional distributions are available for Gibbs sampling. Consider, for example, the site effect for $\log(\tilde{\sigma}_{m,j})$, with distribution given by (7.6). The conjugate priors for the mean, $a_{\tilde{\sigma}}$ and precision, $\zeta_{\tilde{\sigma}}$, are given in (7.9) and (7.11) respectively. Now the density of each site effect $\epsilon_{\tilde{\sigma}}^{(j)}$, $j = 1, \dots, 12$, is given by

$$f(\epsilon_{\tilde{\sigma}}^{(j)}) = \sqrt{\frac{\zeta_{\tilde{\sigma}}}{2\pi}} \exp \left\{ -\frac{\zeta_{\tilde{\sigma}}}{2} (\epsilon_{\tilde{\sigma}}^{(j)} - a_{\tilde{\sigma}})^2 \right\}. \quad (7.12)$$

Thus, the full likelihood is

$$L(a_{\bar{\sigma}}, \zeta_{\bar{\sigma}}) = \left(\sqrt{\frac{\zeta_{\bar{\sigma}}}{2\pi}} \right)^{12} \exp \left\{ -\frac{\zeta_{\bar{\sigma}}}{2} \sum_{j=1}^{12} (\epsilon_{\bar{\sigma}}^{(j)} - a_{\bar{\sigma}})^2 \right\}. \quad (7.13)$$

The prior takes the form

$$\pi(a_{\bar{\sigma}}, \zeta_{\bar{\sigma}}) = \pi(a_{\bar{\sigma}})\pi(\zeta_{\bar{\sigma}}), \quad (7.14)$$

where

$$\pi(a_{\bar{\sigma}}) \propto \exp \left\{ -\frac{c_{\bar{\sigma}}}{2} (a_{\bar{\sigma}} - b_{\bar{\sigma}})^2 \right\} \quad \text{and} \quad (7.15)$$

$$\pi(\zeta_{\bar{\sigma}}) \propto \zeta_{\bar{\sigma}}^{f_{\bar{\sigma}}-1} \exp \{ -g_{\bar{\sigma}} \zeta_{\bar{\sigma}} \}, \quad (7.16)$$

and therefore,

$$\pi(a_{\bar{\sigma}}, \zeta_{\bar{\sigma}}) \propto \zeta_{\bar{\sigma}}^{f_{\bar{\sigma}}-1} \exp \left\{ -\frac{c_{\bar{\sigma}}}{2} (a_{\bar{\sigma}} - b_{\bar{\sigma}})^2 - g_{\bar{\sigma}} \zeta_{\bar{\sigma}} \right\}. \quad (7.17)$$

Now we have the likelihood and the prior, we can write down the posterior distribution:

$$\begin{aligned} \pi(a_{\bar{\sigma}}, \zeta_{\bar{\sigma}} | \epsilon_{\bar{\sigma}}^{(j)}) &\propto \left(\frac{\zeta_{\bar{\sigma}}}{2\pi} \right)^6 \zeta_{\bar{\sigma}}^{f_{\bar{\sigma}}-1} \\ &\times \exp \left\{ -\frac{1}{2} \left[2g_{\bar{\sigma}} \zeta_{\bar{\sigma}} + c_{\bar{\sigma}} (a_{\bar{\sigma}} - b_{\bar{\sigma}})^2 + \zeta_{\bar{\sigma}} \sum_{j=1}^{12} (\epsilon_{\bar{\sigma}}^{(j)} - a_{\bar{\sigma}})^2 \right] \right\}. \end{aligned} \quad (7.18)$$

The full conditional for $a_{\bar{\sigma}}$ then is

$$\pi(a_{\bar{\sigma}} | \cdot) \propto \exp \left\{ -\frac{1}{2} \left[c_{\bar{\sigma}} (a_{\bar{\sigma}} - b_{\bar{\sigma}})^2 + \zeta_{\bar{\sigma}} \sum_{j=1}^{12} (\epsilon_{\bar{\sigma}}^{(j)} - a_{\bar{\sigma}})^2 \right] \right\} \quad (7.19)$$

$$\propto \exp \left\{ -\frac{1}{2} (c_{\bar{\sigma}} + 12\zeta_{\bar{\sigma}}) \left(a_{\bar{\sigma}} - \frac{b_{\bar{\sigma}} c_{\bar{\sigma}} + \zeta_{\bar{\sigma}} \sum_{j=1}^{12} \epsilon_{\bar{\sigma}}^{(j)}}{c_{\bar{\sigma}} + 12\zeta_{\bar{\sigma}}} \right)^2 \right\}, \quad (7.20)$$

that is

$$a_{\bar{\sigma}} | \cdot \sim N_0 \left(\frac{b_{\bar{\sigma}} c_{\bar{\sigma}} + \zeta_{\bar{\sigma}} \sum_{j=1}^{12} \epsilon_{\bar{\sigma}}^{(j)}}{c_{\bar{\sigma}} + 12\zeta_{\bar{\sigma}}}, c_{\bar{\sigma}} + 12\zeta_{\bar{\sigma}} \right). \quad (7.21)$$

Similarly, the full conditional for $\zeta_{\bar{\sigma}}$ is given by

$$\pi(\zeta_{\bar{\sigma}} | \cdot) \propto \left(\frac{\zeta_{\bar{\sigma}}}{2\pi} \right)^6 \zeta_{\bar{\sigma}}^{f_{\bar{\sigma}}-1} \exp \left\{ -\frac{1}{2} \left[2g_{\bar{\sigma}} \zeta_{\bar{\sigma}} + \zeta_{\bar{\sigma}} \sum_{j=1}^{12} (\epsilon_{\bar{\sigma}}^{(j)} - a_{\bar{\sigma}})^2 \right] \right\} \quad (7.22)$$

$$\propto \zeta_{\bar{\sigma}}^{(f_{\bar{\sigma}}+6)-1} \exp \left\{ -\zeta_{\bar{\sigma}} \left[g_{\bar{\sigma}} + \frac{1}{2} \sum_{j=1}^{12} (\epsilon_{\bar{\sigma}}^{(j)} - a_{\bar{\sigma}})^2 \right] \right\}, \quad (7.23)$$

that is

$$\zeta_{\tilde{\sigma}}|\cdot \sim Ga\left(f_{\tilde{\sigma}} + 6, g_{\tilde{\sigma}} + \frac{1}{2} \sum_{j=1}^{12} (\epsilon_{\tilde{\sigma}}^{(j)} - a_{\tilde{\sigma}})^2\right). \quad (7.24)$$

The full conditionals for the other four parameters in (7.9) – (7.11) can be obtained in a similar way. For these parameters, the MCMC is straightforward, since we can use Gibbs sampling (see 6.1.2) to sample directly from the full conditionals. However, ‘one step up’ (at the level of the data), the complexity of the GPD likelihood means that conjugacy is unattainable for each of the fixed effect parameters. For each month m within each site j , the likelihood is given by

$$\begin{aligned} L\left(\gamma_{\tilde{\sigma}}^{(m)}, \gamma_{\xi}^{(m)}, \epsilon_{\tilde{\sigma}}^{(j)}, \epsilon_{\xi}^{(j)}\right) &= -n_{m,j} \log\left(\exp\left\{\gamma_{\tilde{\sigma}}^{(m)} + \epsilon_{\tilde{\sigma}}^{(j)}\right\} + \left(\gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)}\right) u_{m,j}\right) \\ &\quad - \left[1 + \left(\gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)}\right)^{-1}\right] \\ &\quad \times \left\{ \sum_{i=1}^{n_{m,j}} \log\left(\exp\left\{\gamma_{\tilde{\sigma}}^{(m)} + \epsilon_{\tilde{\sigma}}^{(j)}\right\} + \left(\gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)}\right) (y_i + u_{m,j})\right) \right. \\ &\quad \left. - \sum_{i=1}^{n_{m,j}} \log\left(\exp\left\{\gamma_{\tilde{\sigma}}^{(m)} + \epsilon_{\tilde{\sigma}}^{(j)}\right\} + \left(\gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)}\right) u_{m,j}\right) \right\}, \quad (7.25) \end{aligned}$$

where $n_{m,j}$ is the number of exceedances in month m and site j . Here, we use a Metropolis step (see 6.1.2) to update the site and seasonal effects, employing a random walk update procedure. As throughout chapter 6, there are some design considerations to be made in the choice of the random walk maximum step length for each parameter having a Metropolis update, with bad choices leading to poor convergence properties of the chain.

7.3.3 Results

The following results have been obtained on the basis of the MCMC scheme described in 7.3.2. The hierarchical model is of sufficient complexity that assessing convergence of the sampler to its stationary distribution is a non-trivial task; to this end, three large simulations have been run with different starting points. The MCMC was allowed to run for half a million iterations¹, and three starting points were chosen for each fixed effects parameter. Figure 7.3 shows convergence of the site effects $\epsilon_{\tilde{\sigma}}^{(j)}$ – the site effects for $\log(\tilde{\sigma})$ – for all sites $j = 1, \dots, 12$. The first 1000 realisations from the thinned series are shown, with convergence after around 200 observations in each plot. Thus, of the thinned

¹though only every 100th iteration is reported; taking iterations at even intervals reduces dependence and it is often preferable to obtain summary statistics from such a “thinned” chain. Thinning also saves computer storage space

series of length 5000, the first 200 observations are discarded as “burn-in”, leaving 4800 realisations on which to base our inference. Prior parameters were chosen to give a highly non-informative specification:

$$b_{\tilde{\sigma}} = b_{\xi} = 0, c_{\tilde{\sigma}} = c_{\xi} = 10^{-6}, d_{\tilde{\sigma}} = d_{\xi} = e_{\tilde{\sigma}} = e_{\xi} = f_{\tilde{\sigma}} = f_{\xi} = g_{\tilde{\sigma}} = g_{\xi} = 10^{-3}. \quad (7.26)$$

The MCMC returned 60 trace plots in total – the 12 site effects for both GPD marginal parameters and the logistic dependence parameter (those for $\log(\tilde{\sigma})$ as shown in figure 7.3), and the 12 seasonal effects for the two GPD parameters. Trace plots for the other site and seasonal effects parameters not shown were similar to those in figure 7.3. Figure 7.4 shows the MCMC output for Bradfield in January; the top row are the seasonal effects (for $\log(\tilde{\sigma})$ and ξ), the second row the site effects (for $\log(\tilde{\sigma})$, ξ and α), the third row the overall parameters for January in Bradfield (seasonal effect plus site effect), and the bottom row shows posterior densities for these parameters after the removal of “burn-in”. Table 7.2 summarises the results shown in figure 7.4, and compares these with those based on a site-by-site maximum likelihood analysis (see chapter 4). To contrast, the corresponding results for Nottingham in July are also shown. The main advantage of the Bayesian hierarchical model over the standard likelihood analysis is the reduction in sampling variation. Table 7.2 shows that for all parameters, the posterior standard deviations are substantially smaller than the asymptotic standard errors of the corresponding maximum likelihood estimates; thus, the pooling of information is seen to add considerably to the precision of the analysis.

Figure 7.5, which shows plots of maximum likelihood estimates against their hierarchical model counterparts for each season within each site (and for sites only for α and the 1000-year return level), also supports this. Here, we see *shrinkage* (Coles, 2001b) for each of the GPD parameters relative to the maximum likelihood estimates, especially for the shape parameter ξ , with the posterior means from the hierarchical model occupying a smaller range than the MLEs. This means that the extent of information at each site and in each season admits smoothing in the scale and shape of the process. The plot showing MLEs against posterior means for the 1000-year return level reveals another possible benefit of the hierarchical model: the results of maximum likelihood estimation are very unstable for such extreme quantiles, and can provide unrealistic estimates, whereas the hierarchical model reveals a degree of stability through the pooling of information across sites, enabling sensible estimates to be obtained at all locations.

Figure 7.3: Trace plots of the site effects for $\log(\tilde{\sigma})$ for each site in the study

Figure 7.4: MCMC output for Bradfield in January

	Bradfield, January		Nottingham, July	
	Mean (st. dev.)	MLE (asympt. s.e.)	Mean (st. dev.)	MLE (asympt. s.e.)
$\gamma_{\tilde{\sigma}}^{(m)}$	1.891 (0.042)		1.294 (0.042)	
$\gamma_{\xi}^{(m)}$	0.021 (0.018)		0.002 (0.018)	
$\epsilon_{\tilde{\sigma}}^{(j)}$	0.367 (0.044)		-0.121 (0.041)	
$\epsilon_{\xi}^{(j)}$	-0.105 (0.020)		-0.059 (0.017)	
$\epsilon_{\alpha}^{(j)}$	0.385 (0.009)		0.300 (0.011)	
$\tilde{\sigma}_{m,j}$	7.267 (0.211)	8.149 (0.633)	3.234 (0.061)	2.914 (0.163)
$\xi_{m,j}$	-0.084 (0.015)	-0.102 (0.055)	-0.057 (0.013)	0.018 (0.044)
α_j	0.385 (0.009)	0.368 (0.012)	0.400 (0.011)	0.412 (0.020)

Table 7.2: Bayesian fixed effects analysis of extreme wind speeds – Bradfield (January) and Nottingham (July)

Figure 7.5: Posterior means against maximum likelihood estimates of GPD parameters, logistic dependence parameter and 1000-year return level

Site effects

Figure 7.6 shows posterior means for the site effects for each parameter, together with 95% credible intervals. It is obvious from these plots that there is evidence of a site effect for the scale parameter ξ and the logistic dependence parameter α , with all effects being significantly different from zero. There are, however, effects for $\log(\tilde{\sigma})$ which are *not* significantly different from zero; namely, the site effects for sites 9, 10, 11 and 12. It is interesting to note that three of these sites – Keele, Aigburth and Ringway – are the most westerly locations in the study. Linton on Ouse and Bradfield (which observe, on average, the lowest and highest wind speeds respectively) have the most significant site effects for the scale parameter. Similarly, Bradfield also has the most significant site effect for the shape parameter ξ .

Seasonal effects

Figure 7.7 shows the estimated seasonal (monthly) effects for each of the GPD parameters, again with associated 95% credible intervals; examination of these seasonal effects reveals interesting patterns. For the scale parameter σ there is clear evidence of a seasonal effect, with no credible intervals encompassing zero. Further, we can see that the seasons which have the greatest effect on gusts are the winter months, with a steady decline in the significance of the effects as we move through spring and summer. After season 7 (July) there is an increase in the magnitude of these seasonal effects, most sharply between August and September, with January having the most significant effect on gust speeds. In contrast, we see that for ξ , most seasons have no significant effect on the parameter, with the exception of October (and possibly February). Thus, it may be possible to simplify the analysis and model ξ through site effects only.

In summary, the hierarchical model has allowed the identification of structural relationships in the process over the region studied, as well as producing a reduction in sampling variation for inferences at any particular site due to the transfer of information through the model. We should note, however, the distinction in interpretation between Bayesian and classical inferences; nonetheless, the comparison in table 7.2 (and for other sites) is, at the very least, informative.

Figure 7.6: Posterior means and 95% credible intervals for site effects for $\log(\tilde{\sigma})$, ξ and α

Figure 7.7: Posterior means and 95% credible intervals for seasonal effects for $\log(\tilde{\sigma})$ and ξ

7.3.4 Diagnostics

The above results suggest that the hierarchical model is informative, and the closeness of the posterior means to their maximum likelihood counterparts is reassuring. However, more formal goodness-of-fit checks at the local level can be made via probability plots and quantile-quantile plots. By using the fitted parameter values to transform each monthly set of excesses at each site to a common margin (say unit exponential), the fit to all seasons can be assessed simultaneously.

Figure 7.8 shows probability plots and quantile-quantile plots for Bradfield and Nottingham, with similar plots being obtained for all other sites in the study. The overall adequacy of the hierarchical model for each site is strongly supported by these plots, with very little deviation from the diagonal line (which represents equality between fitted and empirical quantities). More recently, methods for checking model goodness-of-fit specific to complex Bayesian models have been developed; for a discussion, see Gelfand and Trevisani (2002).

Figure 7.8: Probability plots and quantile-quantile plots for Bradfield and Nottingham

7.3.5 Return level estimation

As always, in a practical setting we are more interested in estimating extreme quantiles than the actual GPD parameters themselves. We adopt the approach of section 6.3.1; i.e., at each site j , $j = 1, \dots, 12$, the overall r -year return level q_r is obtained as the solution of the equation

$$\sum_{m=1}^{12} \left[1 - \lambda_{u_{m,j}} \left[1 + \xi_{m,j}^{(k)} \left(\frac{q_r^{(k)} - u_{m,j}}{\tilde{\sigma}_{m,j}^{(k)}} \right) \right]_+^{-1/\xi_{m,j}^{(k)}} \right]^{h_m \theta_j^{(k)}} = 12 - \frac{1}{r}, \quad (7.27)$$

solved for each posterior draw k to obtain a posterior distribution for q_r . As before, at each iteration in the MCMC, the current posterior draw for the logistic dependence parameter α is used to simulate a Markov chain with the required strength of dependence, and a method used to estimate the extremal index θ from this simulated chain. As we have already seen from chapter 5, the exact method used to estimate θ has very little effect on return level estimation; nonetheless, to avoid the need to make arbitrary choices of declustering parameter, we use Ferro and Segers' intervals estimator (see chapter 5). We can then solve equation (7.27) numerically using the Fortran NAG subroutine C05AJF (Numerical Algorithms Group, 1999).

Tables 7.3 and 7.4 show the posterior means and standard deviations for the 10, 50, 200 and 1000-year return levels (for Bradfield and Nottingham). These posterior means are compared to the estimates of q_r obtained from a standard likelihood approach (modelling dependence using a Markov chain model – see table 4.2), where asymptotic standard errors were obtained using the delta method. Comparing the results from the hierarchical model to those found under maximum likelihood, there is obvious disagreement. At Bradfield, the 10 and 50 year estimates are similar, but the 200 and 1000-year return levels are considerably lower in the maximum likelihood approach. At Nottingham, return levels appear to be under-estimated under the hierarchical model relative to the maximum likelihood estimates. Other sites with relatively strong gusts (e.g. Bingley) gave results in the direction of those found for Bradfield – i.e. lower maximum likelihood estimates for the longer period return levels than those obtained under the hierarchical model, whereas sites with lower observed wind speeds (e.g. Linton on Ouse) gave results in the direction of those found for Nottingham. However, common to all sites was the increased precision of the hierarchical model; as tables 7.3 and 7.4 show, asymptotic standard errors for return levels at Bradfield and Nottingham are substantially larger than the posterior standard deviations for return levels in the hierarchical model – testimony to the pooling of information across sites and seasons in the latter – and this pattern is repeated for all sites in the study. Thus, the hierarchical model can be seen to add considerably to the

precision of the analysis. To assess the performance of the hierarchical model with regard to return level inference, in section 7.4 we turn our attention to simulated data, where the actual return level q_r is *known*.

	Return Period (years)			
	10	50	200	1000
Hierarchical model	96.887 (0.982)	103.463 (1.333)	112.158 (2.023)	128.862 (3.750)
Maximum likelihood	96.745 (2.864)	103.236 (5.930)	108.152 (8.786)	113.306 (12.219)

Table 7.3: Return levels for Bradfield

	Return Period (years)			
	10	50	200	1000
Hierarchical model	68.294 (0.958)	75.390 (1.474)	81.411 (2.024)	88.292 (2.806)
Maximum likelihood	71.362 (2.565)	83.056 (4.890)	94.010 (11.239)	107.718 (14.564)

Table 7.4: Return levels for Nottingham

7.3.6 Prediction

As in section 6.3.1, we now exploit the natural extension to predictive densities under the Bayesian framework. For each site j , we can solve equation (6.15) for various values of r , and plot these predictive return levels against $-\log\{-\log(1 - 1/r)\}$. Here, the number of iterations after the removal of “burn-in”, B , is 4800. Figure 7.9 shows a plot of the predictive distribution of a future threshold exceedance (results shown for Bradfield and Nottingham) on the usual return period scale. Table 7.5 summarises these results by displaying the 10, 50, 200 and 1000 year predictive return levels. Comparing these results to figure 6.6 and table 6.6 (for Bradfield), we see that, unlike the curve under the standard Bayesian approach, the predictive return level curve under the hierarchical model does not rise steeply for the longer-period return levels, suggesting extrapolation to levels beyond the range of the observed data is subject to less uncertainty in the latter. As we would expect, predictive return levels estimated from the hierarchical model are still greater than any estimative return levels (Bayesian or non-Bayesian – see table 6.6) – since they account for parameter uncertainty and randomness in future observations –

but, for longer periods, these are not as large as predictive return levels under a standard Bayesian approach. The lower return level curve for Nottingham reflects the lower wind speeds observed here.

Figure 7.9: Predictive return level curves for Bradfield (solid line) and Nottingham (dashed line)

	Return period (years)			
	10	50	200	1000
Bradfield	104.392	113.089	119.957	128.128
Nottingham	68.389	75.504	81.655	88.782

Table 7.5: Predictive return levels for Bradfield and Nottingham

7.4 Simulation study

Here, we consider a stationary first-order Markov chain with joint distribution of consecutive pairs of the process given by a bivariate extreme value distribution, with logistic dependence function and GPD marginals (as used in the simulation study of chapter 3). For simplicity, we simulate two sites of data, each with two distinct seasons. Table 7.6 shows the parameter values used in the simulation for each site and each season; the logistic dependence in both seasons within each site was set at $\alpha = 0.2$ to resemble the strong short-term dependence observed in the wind speed data. Seasonal thresholds are chosen such that 5% of the data within each season exceed them. Figure 7.10 shows time series plots of the generated data at both sites, with the two seasons clearly identifiable within each site. Also shown in figure 7.10 are box-plots of the simulated data, by site and season separately. For both sites j ($j = 1, 2$), the t -observation return level q_t (see section 2.7) is then found by substituting the seasonal parameters into (7.27) (where the extremal index for a process with $\alpha = 0.2$ is known to be 0.0616 – see table 5.4) and solving numerically for q_t . Such solutions (for various t) are the *true* t -observation return levels.

	Site1		Site 2	
	Season 1	Season 2	Season 1	Season 2
σ	10	2	15	12
ξ	-0.15	-0.25	-0.1	-0.3

Table 7.6: Parameters used to simulate data

The hierarchical model can then be used to estimate site and seasonal effects, and from these we can obtain posterior distributions for the GPD parameters. Return levels can then be estimated from the hierarchical model by substitution of the posterior draws into (7.27), to obtain posterior distributions for the t -observation return levels q_t . Similarly, sampling distributions for the q_t based on maximum likelihood estimates can be obtained by repeatedly simulating seasonal data with the required marginals and finding MLEs for the GPD parameters (and logistic dependence parameter α). In the same way that each draw from the posterior distribution is used to solve (7.27) to obtain a posterior distribution for q_t , each draw from the sampling distribution can also be substituted into (7.27) to obtain a maximum likelihood sampling distribution for q_t . Credible intervals can be found from both posterior distributions and sampling distributions, and compared to each other (and the true values).

Figure 7.10: Time series plots and box-plots of simulated data

Tables 7.7 and 7.8 show the results of the simulation study for the 1000, 2000, 5000 and 10000-observation return levels. Since both posterior distributions and sampling distributions were heavily skewed, the medians (and 95% credible intervals in parentheses) for each return level are reported. To some extent, the results of the simulation study support the results from the real-life study; both tables 7.7 and 7.8 show the increase in precision under the hierarchical model, with tighter confidence intervals than those found using maximum likelihood. Though in all cases the confidence intervals for both approaches contain the true return levels, the medians from the hierarchical model are closer to the target values than are their maximum likelihood counterparts. The discrepancy between both approaches, as in the real-life study, is most noticeable for the longer-period return levels – especially for the 10000-observation return level (at both sites). In the real-life study, return level estimates for some sites were lower in the hierarchical analysis than under the maximum likelihood approach – for example, Nottingham and Linton on Ouse. This phenomenon was not repeated in the simulation study; for both sites of simulated data, the maximum likelihood approach under-estimated all return levels relative to the hierarchical model. However, only two sites of data were generated. Had a larger-scale simulation study been undertaken, patterns of lower estimates from the hierarchical model in sites that observe lower gusts may have been found. That is not to say that at such sites the hierarchical model performs poorly relative to the maximum likelihood approach; indeed, it may be the case that here the MLEs for various return levels are *over-estimating*.

	Return Period (observations)			
	1000	2000	5000	10000
True value	43.138	45.460	48.184	50.009
Hierarchical model	42.224 (40.755, 43.978)	45.183 (43.276, 47.766)	48.249 (45.668, 51.915)	50.284 (47.200, 54.918)
Maximum likelihood	42.556 (39.876, 45.701)	44.772 (41.788, 48.728)	46.472 (43.195, 51.747)	48.514 (45.080, 54.871)

Table 7.7: Return level inference for simulation study: Site 1

	Return Period (observations)			
	1000	2000	5000	10000
True value	74.821	79.855	85.998	90.284
Hierarchical model	74.622 (69.976, 78.201)	79.797 (74.322, 83.398)	85.923 (79.799, 90.870)	89.949 (82.439, 96.667)
Maximum likelihood	74.191 (67.735, 81.538)	78.863 (70.285, 88.965)	84.534 (73.080, 98.763)	88.410 (74.743, 106.832)

Table 7.8: Return level inference for simulation study: Site 2

7.5 Discussion

The aim of this chapter was to investigate the use of Bayesian hierarchical models in the context of extremes. Until quite recently, extreme value analyses have been fairly rudimentary, almost always using simple model structures and procedures for estimation. Through the hierarchical model presented in this chapter, we aimed to make full use of the available data by more structured statistical modelling, with the identification of site effects and seasonal effects. Further, the temporal evolution of the process was accounted for by imposing a Markov assumption, allowing the quantification of temporal dependence at each site in the study. Estimating the model via Markov chain Monte Carlo, posterior distributions were obtained for the fixed effects, leading to posterior distributions for the GPD parameters (and logistic dependence parameters) themselves. However, as throughout much of this thesis, our aim was to improve return level inference.

The Bayesian fixed effects analysis of the U.K. hourly gust maximum wind speed data has proved entirely satisfactory, the main advantage being an increase in precision owing to the sharing of information between sites and seasons. Moreover, return level inference seems to have been improved still further in this chapter, with return levels estimated from the hierarchical model being more accurate and precise than any other estimative method used in this thesis. The apparent discrepancy between return level estimates (for longer-period return levels) using the standard likelihood approach and the Bayesian fixed effects approach has been supported by the simulation study, which showed maximum likelihood sampling distributions for various return levels to be subject to far greater sampling variation and inaccuracy than posterior distributions for these return levels obtained through the hierarchical model. The extension to predictive return levels also proved to be more satisfactory under the hierarchical model than in the standard Bayesian analysis in chapter 6; though still accounting for parameter uncertainty and randomness in future observations, the predictive analysis in the former was subject to far less variation than in the latter, testimony to the more structured statistical modelling and pooling of information across different sources of variation.

The value of information in extreme value analyses is especially high, and the improvements in efficiency found here through structured statistical modelling are bound to lead to considerable improvements in any practical application. Indeed, though no expert prior information was available for this analysis, any such information which *may* be available could be handled within the model, and could only serve to improve any forms of inference still further.

Appendix A

Appendix to chapter 3

A.1 Simulation study results – first run, 15 studies

Tables A.1, A.2 and A.3 summarise the findings of the first full run of the simulation study in chapter 3. Table A.1 shows the results for studies 1, 4, 7, 10 and 13 (see table 3.1), i.e. the results for processes with strong temporal dependence (given by a logistic dependence parameter of $\alpha = 0.2$). Table A.2 shows corresponding results for studies 2, 5, 8, 11 and 14 ($\alpha = 0.5$) and A.3 for studies 3, 6, 9, 12 and 15 ($\alpha = 0.8$). All three tables show sampling distribution means, medians and standard deviations, as well as 95% and 99% credible intervals, for the marginal GPD scale and shape parameters σ^* and ξ , the 1000–observation return level q_{1000} , the 95% threshold exceedance rate λ and the 99% threshold exceedance rate p . Sampling distribution **means** are shown in bold type, medians in standard type, standard deviations are underlined, and (*confidence intervals*) are given in italics in parentheses. Results are based on sampling distributions of length 1000, and are given to 5 decimal places. All three tables show results which use all threshold exceedances (i.e. analyses which *ignore* temporal dependence) and results which are based on declustering exceedances using either cluster peaks or the first observed cluster values (see section 3.5.3). Estimates using cluster peak excesses are given a subscript p , and those using first observed cluster excesses are given a subscript f . True parameter values are given in the far-left box for each table A.1 – A.3.

Panels A.1, A.2 and A.3 show (unsmoothed) density plots of the sampling distributions for the GPD scale and shape parameters and the 1000–observation return level. In each plot, the vertical line represents the true parameter value; bold density plots indicate the use of all threshold excesses, thinner lines cluster peak excesses and dashed lines indicate the use of first observed threshold excesses. The plots in panel A.1 correspond to the results tabulated in *table* A.1, i.e. the results for which $\alpha = 0.2$. Similarly, table

A.2 summarises the density plots shown in panel A.2, and table A.3 the plots in panel A.3.

Tables A.4, A.5 and A.6 summarise the performance of the MLEs for the GPD scale and shape parameters and the 1000–observation return level when temporal dependence was both ignored and removed. These tables show the estimated bias (true parameter value minus the expected value found from the sampling distribution), **mean squared error** (MSE) (given in bold type) and the (percentage) *efficiency* of estimates (given in italics) based on using a filtered set of exceedances (i.e. using cluster peaks or first cluster values) relative to the method which uses *all* threshold excesses. Table A.4 shows results for studies 1, 4, 7, 10 and 13 ($\alpha = 0.2$), table A.5 for studies 2, 5, 8, 11 and 14 ($\alpha = 0.5$) and table A.6 for studies 3, 6, 9, 12 and 15 ($\alpha = 0.8$).

Tables A.7, A.8 and A.9 compare estimated standard errors of MLEs (obtained from standard likelihood theory) with the standard deviations of the sampling distributions for the GPD parameters σ^* and ξ and the 1000–observation return level q_{1000} . Each table reports the standard deviations of the sampling distributions and the mean of the distributions of the (estimated) standard errors attached to each estimate. As before, table A.7 summarises results for $\alpha = 0.2$ (studies 1, 4, 7, 10 and 13), table A.8 for $\alpha = 0.5$ (studies 2, 5, 8, 11 and 14) and table A.9 for $\alpha = 0.8$ (studies 3, 6, 9, 12 and 15).

A.2 Distribution of cluster peak excesses

Let $X_1, X_2, \dots$ be IID $\text{GPD}(\sigma, \xi)$ random variables.

For *any exceedance* X_i of a threshold u , $u > 0$,

$$\begin{aligned} \Pr(X_i > u + x | X_i > u) &= \left[1 + \frac{\xi x_i}{\sigma^*} \right]_+^{-1/\xi} \\ &= S(x_i; \sigma^*, \xi), \end{aligned}$$

where S is the survival function of the GPD and $\sigma^* = \sigma + \xi u$ and ξ are the scale and shape parameters (respectively).

Now consider the distribution of any *cluster peak*. We know that, in the limit, the distribution of excess for a randomly chosen exceedance *has exactly the same distribution as that of a cluster peak*. However, in practice, it might be unsurprising to find that the distributions of a randomly chosen excess and a cluster peak excess differ.

For simplicity, define X_i to be a cluster peak above u if, and only if, $X_i > u$ and $X_i > X_{i-\kappa}, X_i > X_{i-\kappa+1}, \dots, X_i > X_{i+\kappa-1}, X_i > X_{i+\kappa}$ for some declustering parameter κ . We call this method of declustering “top-down” declustering. We need to evaluate the distribution of cluster peaks found in this way. Now

$$\Pr(X_i > u + x | X_i \text{ is a cluster peak above } u)$$

$$= \Pr(X_i > u + x | X_i > u \cap X_i \text{ largest observation in } X_{i-\kappa}, \dots, X_{i+\kappa})$$

$$\begin{aligned} &= \Pr(X_i > u + x | X_i \text{ largest of exactly **one** exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa}) \\ &\quad \times \Pr(\text{exactly **one** exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u) \end{aligned} \tag{A.1}$$

$$\begin{aligned} &+ \Pr(X_i > u + x | X_i \text{ largest of exactly **two** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa}) \\ &\quad \times \Pr(\text{exactly **two** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u) \end{aligned} \tag{A.2}$$

+

⋮

$$\begin{aligned} &+ \Pr(X_i > u + x | X_i \text{ largest of exactly **2}\kappa + \mathbf{1}** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa}) \\ &\quad \times \Pr(\text{exactly **2}\kappa + \mathbf{1}** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u). \end{aligned}$$

Look at (A.1). Using the definition of conditional probability, we can write

$$\Pr(X_i > u + x | X_i \text{ largest of exactly **one** exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})$$

as

$$\begin{aligned} & \frac{\Pr(X_i > u + x \cap X_i > u, X_i \text{ largest of exactly one exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})}{\Pr(X_i > u, X_i \text{ largest of exactly one exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})} \\ &= \frac{\Pr(X_i > u + x)}{\Pr(X_i > u)} \\ & \times \frac{\Pr(X_i \text{ largest of exactly one exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u + x)}{\Pr(X_i \text{ largest of exactly one exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u)} \\ &= \frac{\lambda_{u+x}}{\lambda_u} \times \frac{\Pr(X_i \text{ largest of exactly one exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u + x)}{\Pr(X_i \text{ largest of exactly one exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u)} \\ &= \frac{\lambda_{u+x}}{\lambda_u} \times \frac{A_1}{B_1}, \end{aligned}$$

where λ_u and λ_{u+x} are the probabilities of exceeding thresholds u and $u + x$ respectively. Given $X_i > u + x$, then X_i is an exceedance of u . Thus, the probability of this value being the largest of exactly one exceedance of u is 1, since this is the *only* exceedance. So $A_1 = 1$. Similarly, given $X_i > u$, the probability of exactly one exceedance of u is also 1, since this is the only exceedance. Thus, $B_1 = 1$. So, (A.1) is equal to

$$\frac{\lambda_{u+x}}{\lambda_u} \times 1 \times \Pr(\text{exactly **one** exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u).$$

Now look at (A.2). Again, using the definition of conditional probability, we can write

$$\Pr(X_i > u + x | X_i \text{ largest of exactly **two** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})$$

as

$$\begin{aligned} & \frac{\Pr(X_i > u + x \cap X_i > u, X_i \text{ largest of exactly two exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})}{\Pr(X_i > u, X_i \text{ largest of exactly two exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})} \\ &= \frac{\Pr(X_i > u + x)}{\Pr(X_i > u)} \\ & \times \frac{\Pr(X_i \text{ largest of exactly two exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u + x)}{\Pr(X_i \text{ largest of exactly two exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u)} \\ &= \frac{\lambda_{u+x}}{\lambda_u} \times \frac{\Pr(X_i \text{ largest of exactly two exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u + x)}{\Pr(X_i \text{ largest of exactly two exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u)} \\ &= \frac{\lambda_{u+x}}{\lambda_u} \times \frac{A_2}{B_2}. \end{aligned}$$

Given $X_i > u$, the probability of X_i being the largest of exactly two exceedances is simply $\frac{1}{2}$, since both X_i and the other exceedance have an equal chance of being the largest. So $B_2 = \frac{1}{2}$. Given $X_i > u + x$, it is more likely that X_i is the largest of exactly two exceedances, since we only know that the other exceedance is greater than u . Let X_j be the other exceedance. Then A_2 can be written as

$$\begin{aligned} & \Pr(X_i \text{ largest} | X_j > u + x) \times \Pr(X_j > u + x | X_j > u) \\ & + \Pr(X_i \text{ largest} | u < X_j < u + x) \times \Pr(u < X_j < u + x | X_j > u) \\ = & \frac{1}{2} \times \frac{\lambda_{u+x}}{\lambda_u} + 1 \times \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right) \\ = & 1 - \frac{\lambda_{u+x}}{2\lambda_u}. \end{aligned}$$

So, for (A.2) we have

$$\frac{\lambda_{u+x}}{\lambda_u} \times 2 \left\{1 - \frac{\lambda_{u+x}}{2\lambda_u}\right\} \times \Pr(\text{exactly **two** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u).$$

The next contribution to the sum given on page 185 will be

$$\begin{aligned} & \Pr(X_i > u + x | X_i \text{ largest of exactly **three** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa}) \\ & \times \Pr(\text{exactly **three** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u). \end{aligned} \quad (\text{A.3})$$

As before, using the definition of conditional probability, we can write

$$\Pr(X_i > u + x | X_i \text{ largest of exactly **three** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})$$

as

$$\frac{\Pr(X_i > u + x \cap X_i > u, X_i \text{ largest of exactly three exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})}{\Pr(X_i > u, X_i \text{ largest of exactly three exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa})}$$

which, through similar arguments as before, is equal to

$$\begin{aligned} & \frac{\lambda_{u+x}}{\lambda_u} \times \frac{\Pr(X_i \text{ largest of exactly three exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u + x)}{\Pr(X_i \text{ largest of exactly three exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u)} \\ = & \frac{\lambda_{u+x}}{\lambda_u} \times \frac{A_3}{B_3}. \end{aligned}$$

Given $X_i > u$, and we know we have two other exceedances of u – say X_j and X_k – the probability that X_i is the largest exceedance is simply $\frac{1}{3}$. Thus, $B_3 = \frac{1}{3}$. Now given

$X_i > u + x$ and $X_j, X_k > u$, we can write A_3 as

$$\begin{aligned}
 & \Pr(X_i \text{ largest} | u < X_j, X_k < u + x) \times \Pr(u < X_j, X_k < u + x | X_j, X_k > u) \\
 & + \Pr(X_i \text{ largest} | X_j > u + x, u < X_k < u + x) \times \Pr(X_j > u + x, u < X_k < u + x | X_j, X_k > u) \\
 & + \Pr(X_i \text{ largest} | X_k > u + x, u < X_j < u + x) \times \Pr(X_k > u + x, u < X_j < u + x | X_j, X_k > u) \\
 & + \Pr(X_i \text{ largest} | X_j, X_k > u + x) \times \Pr(X_j, X_k > u + x | X_j, X_k > u) \\
 \\
 & = 1 \times \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^2 \\
 & \quad + \frac{1}{2} \times \left(\frac{\lambda_{u+x}}{\lambda_u}\right) \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right) \\
 & \quad + \frac{1}{2} \times \left(\frac{\lambda_{u+x}}{\lambda_u}\right) \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right) \\
 & \quad + \frac{1}{3} \times \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^2 \\
 \\
 & = 1 \times \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^2 \\
 & \quad + 2 \left[\frac{1}{2} \left(\frac{\lambda_{u+x}}{\lambda_u}\right) \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right) \right] \\
 & \quad + 1 \times \frac{1}{3} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^2.
 \end{aligned}$$

So for (A.3) we have

$$\begin{aligned}
 & \frac{\lambda_{u+x}}{\lambda_u} \times 3 \left\{ 1 \times \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^2 + 2 \times \left[\frac{1}{2} \left(\frac{\lambda_{u+x}}{\lambda_u}\right) \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right) \right] + 1 \times \frac{1}{3} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^2 \right\} \\
 & \times \Pr(\text{exactly **three** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u).
 \end{aligned}$$

For the fourth contribution we will get

$$\begin{aligned}
 & \frac{\lambda_{u+x}}{\lambda_u} \times 4 \left\{ 1 \times \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^3 + 3 \times \left[\frac{1}{2} \left(\frac{\lambda_{u+x}}{\lambda_u}\right) \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^2 \right] \right. \\
 & \quad \left. + 3 \times \left[\frac{1}{3} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^2 \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right) \right] + 1 \times \frac{1}{4} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^3 \right\} \\
 & \times \Pr(\text{exactly **four** exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u).
 \end{aligned}$$

For the T -th contribution we will get

$$\begin{aligned}
\frac{\lambda_{u+x}}{\lambda_u} \times T \left\{ \begin{aligned} & \frac{\binom{T-1}{0}}{1} \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^{T-1} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^0 \\ & + \frac{\binom{T-1}{1}}{2} \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^{T-2} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^1 \\ & + \frac{\binom{T-1}{2}}{3} \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^{T-3} \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^2 \\ & + \vdots \\ & + \frac{\binom{T-1}{T-1}}{T} \left(1 - \frac{\lambda_{u+x}}{\lambda_u}\right)^0 \left(\frac{\lambda_{u+x}}{\lambda_u}\right)^{T-1} \end{aligned} \right\} \\
\times \Pr(\text{exactly } \mathbf{T} \text{ exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u).
\end{aligned}$$

Now, $\Pr(\text{exactly } \mathbf{one} \text{ exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u)$ is equivalent to having *no* exceedances of u in $X_{i-\kappa}, \dots, X_{i+\kappa}$ (excluding X_i) since we already have one exceedance (X_i itself). So let W : Number of exceedances of u (excluding X_i); W can be assumed binomially distributed with 2κ trials and probability of ‘success’ λ_u , i.e. $W \sim \text{Bin}(2\kappa, \lambda_u)$. Thus,

$$\begin{aligned}
& \Pr(\text{exactly } \mathbf{one} \text{ exceedance of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u) \\
&= \Pr(W = 0) \\
&= \binom{2\kappa}{0} \lambda_u^0 (1 - \lambda_u)^{2\kappa}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \Pr(\text{exactly } \mathbf{two} \text{ exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u) \\
&= \Pr(W = 1) \\
&= \binom{2\kappa}{1} \lambda_u^1 (1 - \lambda_u)^{2\kappa-1}.
\end{aligned}$$

Generally,

$$\begin{aligned}
& \Pr(\text{exactly } \mathbf{T} \text{ exceedances of } u \text{ in } X_{i-\kappa}, \dots, X_{i+\kappa} | X_i > u) \\
&= \Pr(W = T - 1) \\
&= \binom{2\kappa}{T-1} \lambda_u^{T-1} (1 - \lambda_u)^{2\kappa-(T-1)}.
\end{aligned}$$

For the distribution of cluster peak excesses, we now have

$$\begin{aligned}
 & \Pr(X_i > u + x | X_i \text{ is a cluster peak above } u) \\
 &= \frac{\lambda_{u+x}}{\lambda_u} \times 1 \times \binom{2\kappa}{0} \lambda_u^0 (1 - \lambda_u)^{2\kappa} \\
 &+ \frac{\lambda_{u+x}}{\lambda_u} \times 2 \left\{ 1 - \frac{\lambda_{u+x}}{2\lambda_u} \right\} \times \binom{2\kappa}{1} \lambda_u^1 (1 - \lambda_u)^{2\kappa-1} \\
 &+ \vdots \\
 &+ \frac{\lambda_{u+x}}{\lambda_u} \\
 &\times \left[(2\kappa + 1) \left\{ \frac{\binom{2\kappa}{0}}{1} \left(1 - \frac{\lambda_{u+x}}{\lambda_u} \right)^{2\kappa} \left(\frac{\lambda_{u+x}}{\lambda_u} \right)^0 \right. \right. \\
 &+ \vdots \\
 &\left. \left. + \frac{\binom{2\kappa}{2\kappa}}{2\kappa + 1} \left(1 - \frac{\lambda_{u+x}}{\lambda_u} \right)^0 \left(\frac{\lambda_{u+x}}{\lambda_u} \right)^{2\kappa} \right\} \right] \\
 &\times \binom{2\kappa}{2\kappa} \lambda_u^{2\kappa} (1 - \lambda_u)^0 \\
 &= \frac{\lambda_{u+x}}{\lambda_u} \sum_{T=1}^{2\kappa+1} T \left\{ \sum_{n=0}^{T-1} \frac{\binom{T-1}{n}}{n+1} \left(1 - \frac{\lambda_{u+x}}{\lambda_u} \right)^{T-n-1} \left(\frac{\lambda_{u+x}}{\lambda_u} \right)^n \right\} \binom{2\kappa}{T-1} \lambda_u^{T-1} (1 - \lambda_u)^{2\kappa+1-T},
 \end{aligned} \tag{A.4}$$

where

$$\begin{aligned}
 \frac{\lambda_{u+x}}{\lambda_u} &= \frac{\Pr(X_i > u + x)}{\Pr(X_i > u)} \\
 &= \frac{\Pr(X_i > u + x \cap X_i > u)}{\Pr(X_i > u)} \\
 &= \Pr(X_i > u + x | X_i > u) \\
 &= S(x_i; \sigma^*, \xi).
 \end{aligned} \tag{A.5}$$

The expression in (A.4) gives the distribution of *cluster peak excesses* based on filtering out a set of threshold excesses using “top–down” declustering with declustering parameter κ ; (A.5) gives the distribution of *all threshold excesses* which is the GPD with parameters σ^* and ξ . The result is valid for independent observations only.

Bibliography

- Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.
- Ancona–Navarrete, M.A. and Tawn, J.A. (2000). A Comparison of Methods for Estimating the Extremal Index. *Extremes*, **3**, 5—38.
- Anderson, C.W. (1990). Discussion of paper by A.C. Davison and R.L. Smith. *J. R. Statist. Soc., B*, **52**, 425—426.
- BADC (2003). Met Office Surface Data Users Guide.
badc.nerc.ac.uk/data/surface/ukmo_guide.html. (12 February 2004).
- Barnett, V. and Turkman, K.F. (1993). *Statistics for the environment: SPRUCE conference*. Commission of the European Communities, Lisbon.
- Barnett, V. and Turkman, K.F. (1994). *Statistics for the environment: water-related issues*. Wiley, Chichester.
- Barnett, V. and Turkman, K.F. (1997). *Statistics for the environment 3: pollution assessment and control*. Wiley, Chichester.
- Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The (NEW) S Language*. Chapman & Hall, London.
- de Blij, H.J. and Müller, P.O. (1996). *Physical Geography of the Global Environment*. Wiley, New York.
- Bortot, P. and Coles, S.G. (2001). Extremes of Markov Chains with Tail-Switching Potential. *J. R. Statist. Soc., B*, **65**, 851—867.
- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice Hall, New Jersey.
- Brabson, B.B. and Palutikof, J.P. (2000). Tests of the Generalized Pareto Distribution for Predicting Extreme Wind Speeds. *J. App. Met.*, **39**, 1627—1640.

- Brainerd, W.S., Goldberg, C.H. and Adams, J.C. (1996). *Programmer's Guide to Fortran 90*. Springer, New York.
- British Standards Institution (1997). *Code of Basic Data for the Design of Buildings; CP 3, ch. V, Loading; part 2, Wind Loads*. British Standards Institution, London.
- Cartwright, D.E. (1958). On estimating the mean energy of sea waves from the highest waves in a record. *Proc. R. Soc. Lond. A*, **247**, 22—48.
- Coles, S.G. (1991). Statistical Methodology for the Multivariate Analysis of Environmental Extremes. *PhD thesis*, University of Sheffield, Sheffield.
- Coles, S.G. and Tawn, J.A. (1991). Modelling Extreme Multivariate Events. *J. R. Statist. Soc., B*, **53**, 377—392.
- Coles, S.G., Smith, R.L. and Tawn, J.A. (1994). A seasonal Markov model for extremely low temperatures. *Environmetrics*, **5**, 221—239.
- Coles, S.G. and Walshaw, D. (1994). Directional Modelling of Extreme Wind Speeds. *Appl. Statist.*, **43**, 139—157.
- Coles, S.G. and Powell, E.A. (1996). Bayesian methods in extreme value modelling: a review and new developments. *Internat. Statist. Rev.*, **64**, 119—136.
- Coles, S.G. and Tawn, J.A. (1996). A Bayesian Analysis of Extreme Rainfall Data. *Appl. Statist.*, **45**, 463—478.
- Coles, S.G. (1999). *Extreme value theory and applications*.
www.maths.lancs.ac.uk/coless/notes/. (30 Sep. 2000).
- Coles, S.G. and Dixon, M.J. (1999). Likelihood-Based Inference for Extreme Value Models. *Extremes*, **2**, 5—23.
- Coles, S.G., Heffernan, J.E. and Tawn, J.A. (1999). Dependence Measures for Extreme Value Analyses. *Extremes*, **2**, 339—365.
- Coles, S.G. (2001). *An introduction to statistical modeling of extreme values*. Springer, London.
- Coles, S.G. (2001b). A Random Effects Analysis of Extreme Wind Speed Data. Preprint.
- Cook, N.J. (1982). Towards better estimation of extreme winds. *J. Wind Engng. and Ind. Aerodynamics*, **9**, 295—323.

- Cook, N.J. (1985). *The designer's guide to wind loading of building structures*. Dept. of the Env., London.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, London.
- Davison, A.C. and Smith, R.L. (1990). Models for Exceedances over High Thresholds (with discussion). *J. R. Statist. Soc., B*, **52**, 393—442.
- DEFRA (2003). A Satellite-Based Atlas of Rainfall and Wind Speeds over the Eastern North Atlantic and North Sea.
www.defra.gov.uk/environment/marine/mlis/summaries/. (5 Oct. 2003).
- Dixon, M.J. (1998). *A Course in Extreme Value Statistics incorporating Environmental Statistics*. Course notes, Newcastle University, Newcastle upon Tyne.
- Fawcett, L. and Walshaw, D. (2005). A Hierarchical Model for Extreme Wind Speeds. Submitted.
- Fawcett, L. and Walshaw, D. (2005). Bayesian Inference for the Extremal Index. Submitted.
- Fawcett, L. and Walshaw, D. (2005). Estimation of Extremes Occurring in Clusters. Submitted.
- Ferro, C.A.T. and Segers, J. (2003). Inference for clusters of extreme values. *J. R. Statist. Soc., B*, **65**, 545—556.
- Fisher, R.A. and Tippett, L.H.C. (1928). On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.*, **24**, 180—190.
- Galambos, J., Lechner, J. and Simiu, E. (1993). *Extreme Value Theory and Applications*. Kluwer, Dordrecht.
- Gelfand, A.E. and Trevisani, M. (2002). Bayesian Measures of Model Complexity and Fit. *J. R. Statist. Soc., B*, **87**, 631.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.
- Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, **44**, 423—453.

- Gomes, M.I., (1993). On the estimation of parameters of rare events in environmental time series. In *Statistics for the environment 2: Water Related Issues* (V. Barnett and K.F. Turkman), 225—241.
- Gross, J.L., Heckert, N.A., Lechner, J.A. and Simiu, E. (1994). Extreme Winds Estimation by ‘Peaks Over Threshold’ and Epochal Methods. *Structures Congress XII*, **2**, 1472—1477.
- Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.
- Gumbel, E.J. (1960). Bivariate Exponential Distributions. *J. Am. Statist. Ass.*, **55**, 698—707.
- Guttman, N.B., Hosking, J.R.M. and Wallis, J.R. (1993). Regional precipitation quantile values for the continental U.S. computed from L -moments. *Journal of Climate*, **6**, 2326—2340.
- Heffernan, J.E. and Tawn, J.A. (2004). A conditional approach for multivariate extreme values. *J. R. Statist. Soc., B, Part 2*, **66**, 1—34.
- Holmes, J.D. (2001). *Wind Loading of Structures*. Spon Press, London.
- Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985). Estimation of the Generalized Extreme-Value Distribution by the method of Probability-Weighted Moments. *Technometrics*, **27**, 251—261.
- Hosking, J.R.M. and Wallis, J.R. (1987). Parameter and quantile estimation for the generalised Pareto distribution. *Technometrics*, **29**, 339—349.
- Hosking, J.R.M. (1990). L -moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Statist. Soc., B*, **52**, 105—124.
- Hosking, J.R.M. (1992). Moments or L -moments? An example comparing two measures of distributional shape. *The American Statistician*, **46**, 186—189.
- Hosking, J.R.M. and Wallis, J.R. (1993). Some statistics useful in regional frequency analysis. *Water Resources Research*, **29**, 271—281.
- Hsing, T. (1991). Estimating parameters of rare events. *Stoch. Process. Applic.*, **37**, 117—139.
- Hsing, T. (1993). Extremal index estimation for a weakly dependent stationary sequence. *Ann. Statist.*, **21**, 2043—2071.

- Hsing, T., Hüsler, J. and Leadbetter, M.R. (1988). On the exceedance point process for a stationary sequence. *Prob. Theory Rel. Fields* **78**, 97—112.
- Hüsler, J. (1986). Extreme values of non-stationary random sequences. *J. Appl. Probab.*, **23**, 937—950.
- Jenkinson, A.F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J. Roy. Met. Soc.*, **81**, 158—171.
- Joe, H., Smith, R.L. and Weissman, I. (1992). Bivariate threshold models for extremes. *J. R. Statist. Soc., B*, **54**, 171—183.
- Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.
- Krause, A. and Olsen, M. (1997). *The Basics of S and S-Plus*. Springer-Verlag, New York.
- Leadbetter, M.R. (1983). Extremes and local dependence in stationary sequences. *Z. Wahrscheinlichkeitsth*, **65**, 291—306.
- Leadbetter, M.R. (1995). On high level exceedance modeling and tail inference. *J. Statist. Plan. and Inf.*, **45**, 247—260.
- 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. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.*, **16**, 431—476.
- Lindgren, G. and Rootzén, H. (1987). Extreme values: theory and technical applications. *Scand. J. Statist.*, **14**, 241—279.
- Loynes, R.M. (1965). Extreme values in uniformly mixing stationary stochastic processes. *Ann. Math. Statist.*, **36**, 993—999.
- Maindonald, J.H. (2000). *Using R for Data Analysis and Graphics – An Introduction*. Australian National University, Canberra.
- von Mises, R. (1954). La distribution de la plus grande de n valeurs. In *Selected papers*, Vol. II, pp. 271—294. Amer. Math. Soc., Providence R.I.
- Newell, G.F. (1964). Asymptotic extremes for m -dependent random variables. *Ann. Math. Statist.*, **35**, 1322—1325.

- Numerical Algorithms Group (1999). *NAG Fortran Library Manual*. Numerical Algorithms Group, Oxford.
- O'Brien, G.L. (1974). The maximum term of uniformly mixing stationary processes. *Z. Wahrscheinlichkeitsth*, **30**, 57—63.
- Pauli, F. and Coles, S.G. (2001). Penalized likelihood inference in extreme value analyses. *J. Appl. Stat.*, **28**, 547—560.
- Pickands, J. (1981). Multivariate extreme value distributions. *Bull. Int. Statist. Inst.*, XLIX (Book 2), 859—878.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd edn. Wiley, New York.
- Ross, W.H. (1987). A peaks-over-threshold analysis of extreme wind speeds. *Canadian Journal of Statistics*, **15**, 328—335.
- Sachs, P. (1978). *Wind Forces in Engineering*. Pergamon Press, New York.
- Shellard, H.C. (1976). Wind. In *The Climate of the British Isles* (eds. Chandler, T.J. and Gregory, S.), pp. 39—73. Longman, London.
- Simiu, E. and Lozier, D.W. (1975). *The buffeting of tall structures by strong winds*. U.S. Dept. of Commerce, National Bureau of Standards, Washington.
- Simiu, E. and Scanlan, R.H. (1996). *Wind effects on structures: fundamentals and applications to design*. Wiley, New York.
- Smith, E.L. and Walshaw, D. (2003). Modelling Bivariate Extremes in a Region. *Bayesian Statistics*, **7**, 681—690.
- Smith, R.L. (1984). Threshold methods for sample extremes. In *Statistical Extremes and Applications* (ed. J. Tiago de Oliveira), pp. 621—638. Reidel, Dordrecht.
- Smith, R.L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, **72**, 67—92.
- Smith, R.L. (1986). Extreme value theory based on the r -largest annual events. *J. Hydrol.*, **86**, 27—43.
- Smith, R.L. and Naylor, J.C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Appl. Statist.*, **36**, 358—369.

- Smith, R.L. (1988). Extreme values for dependent sequences via the Stein–Chen method of Poisson approximation. *Stochast. Process. Applicns*, **30**, 317–327.
- Smith, R.L. (1989). Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground–Level Ozone. *Statistical Science*, **4**, 367–393.
- Smith, R.L., Tawn, J.A. and Yuen, H.K. (1990). Statistics of multivariate extremes. *Internat. Statist. Rev.*, **58**, 47–58.
- Smith, R.L. (1991). Regional Estimation from Spatially Dependent Data. Preprint.
- Smith, R.L. (1992). The Extremal Index for a Markov Chain. *J. Appl. Prob*, **29**, 37–45.
- Smith, R.L. and Weissman, I. (1994). Estimating the Extremal Index. *J. R. Statist. Soc., B*, **56**, 515–528.
- Smith, R.L., Tawn, J.A. and Coles, S.G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, 249–268.
- Smith, S.G. (1984). A stochastic model to generate sequences of hourly mean wind speeds for different sites in the United Kingdom. *Journal of Climatology*, **4**, 133–148.
- Smithson, P.A. (1987). An analysis of wind speed and direction at a high–altitude site in the southern Pennines. *Meteorological Magazine*, **116**, 74–85.
- Stephenson, A. (2003). Simulating Multivariate Extreme Value Distributions of Logistic Type. Preprint.
- Stephenson, A. (2003). *A User’s Guide to the evd Package (Version 2.0)*.
www.maths.lancs.ac.uk/~stephena/. (20 Jul. 2003).
- Tawn, J.A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, **75**, 397–415.
- Tawn, J.A. (1988b). An extreme value theory model for dependent observations. *J. Hydrol.*, **101**, 227–250.
- Tawn, J.A. (1990). Discussion of paper by A.C. Davison and R.L. Smith. *J. R. Statist. Soc., B*, **52**, 428–429.
- Tawn, J.A. and Vassie, J.M. (1989). Extreme sea–levels: the joint probabilities method revisited and revised. *Proc. Inst. Civ. Engrs., Part 2*, **25**, 433–450.

- Tiago de Oliveira, J. (1984). Bivariate models for extremes; statistical decision. In *Statistical Extremes and Applications* (Ed. J. Tiago de Oliveira), pp. 131—153. Reidel, Dordrecht.
- Walshaw, D. (1991). Statistical Analysis of Extreme Wind Speeds. *PhD Thesis*. University of Sheffield, Sheffield.
- Walshaw, D. (1994). Getting the Most From Your Extreme Wind Data: A Step by Step Guide. *J. Res. Natl. Inst. Stand. Technol.*, **99**, 399—411.
- Walshaw, D. (1999). Extremes of Mixed Environmental Processes. *Bayesian Statistics*, **6**, 849—858.
- Walshaw, D. (2000). Modelling extreme wind speeds in regions prone to hurricanes. *Appl. Statist.*, **49**, 51—62.
- Walshaw, D. and Anderson, C.W. (2000). A model for extreme wind gusts. *Appl. Statist.*, **49**, 499—508.
- Weissman, I. and Novak, S.Y. (1998). On blocks and runs estimators of the extremal index. *J. Statist. Plan. and Inf.*, **66**, 281—288.
- Yue, S. (2000). The Gumbel Mixed Model Applied to Storm Frequency Analysis. *Water Resources Management*, **14**, 377—389.