Modelling Environmental Extremes

Lee Fawcett and Dave Walshaw

School of Mathematics & Statistics

Newcastle University, UK

Short course for the 19th annual conference of

The International Environmetrics Society

Sunday, June 8th 2008

The University of British Columbia Okanagan,

Kelowna, Canada

1 Classical models and threshold models

1.1 Introduction

Statistical modelling of extreme weather has a very practical motivation: reliability — anything we build needs to have a good chance of surviving the weather/environment for the whole of its working life. This has obvious implications for civil engineers and planners. They need to know:

- how strong to make buildings;
- how high to build sea walls;
- how tall to build reservoir dams;
- how much fuel to stockpile;

etc.

This motivates the need to estimate what the:

- strongest wind;
- highest tide;
- heaviest rainfall;
- most severe cold-spell;

etc. will be over some fixed period of future time. The only sensible way to do this is to use data on the variable of interest (wind, rain etc.) and fit an appropriate statistical model. The models themselves are motivated by asymptotic theory, and this is our starting point.

1.2 Classical models

Extreme value modelling has a central theoretical result, analogous to the Central Limit Theorem. Suppose $X_1, X_2, \ldots,$ is an independent and identically distributed sequence of random variables. Define

$$M_n = \max\{X_1, \ldots, X_n\}.$$

We are interested in the limiting distribution of M_n as $n \to \infty$. As with the mean, $\bar{X}$, of $\{X_1, \ldots, X_n\}$, the limiting distribution of M_n as $n \to \infty$ is *degenerate*, and we need to work with a normalised version.

1.2.1 The Extremal Types Theorem (Fisher and Tippett, 1928)

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \quad \text{as} \quad n \to \infty,$$

where G is a non–degenerate distribution function, then G belongs to one of the following families:

$$\begin{aligned}
I : G(z) &= \exp\left\{-\exp\left[-\left(\frac{z-\beta}{\gamma}\right)\right]\right\}, \quad -\infty < z < \infty; \\
II : G(z) &= \exp\left\{-\left(\frac{z-\beta}{\gamma}\right)^{-\alpha}\right\}, \quad z > \beta; \quad [G(z) = 0, z \leq \beta]; \\
III : G(z) &= \exp\left\{-\left[-\left(\frac{z-\beta}{\gamma}\right)^{\alpha}\right]\right\}, \quad z < \beta; \quad [G(z) = 1, z \geq \beta],
\end{aligned}$$

for parameters $\gamma > 0, \beta$, and $\alpha > 0$.

1.2.2 The Generalised Extreme Value Distribution (GEV)

Families I, II and III are widely referred to as Gumbel, Frechet and Weibull (or Extreme Value Types I, II and III) respectively.

Fortunately they can be combined into a single family, known as the Generalised Extreme Value Distribution (GEV), with c.d.f.

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{1}$$

defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, and where μ, $\sigma > 0$ and ξ are *location*, *scale* and *shape* parameters respectively.

Note that the Extreme Value Types I, II and III correspond to the cases $\xi = 0$, $\xi > 0$ and $\xi < 0$ respectively.

For Type I, we need to take the limiting form of Equation (1) as $\xi \to 0$, which gives

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \tag{2}$$

defined for all z.

So the Extremal Types Theorem can be restated with (1) as the limiting form, and this provides the basis for our first modelling approach.

Approach 1: "Block maxima"

Break up our sequence $X_1, X_2, \ldots$ into blocks of size n (with n reasonably large), and extract only the maximum observation from each block.

Now fit Model (1) to the sequence of extracted maxima $M_{(1)}, M_{(2)}, \ldots, M_{(N)}$ and use this as the basis for statistical inference. The most common implementation of this approach for weather data is to take our block size to be one year. This rough and ready approach has shown itself to be surprisingly robust!

1.2.3 Example: Annual maximum rainfall

Consider the annual maxima of daily rainfall accumulations (*mm*) at a location in SW England, from 1914 to 1961.

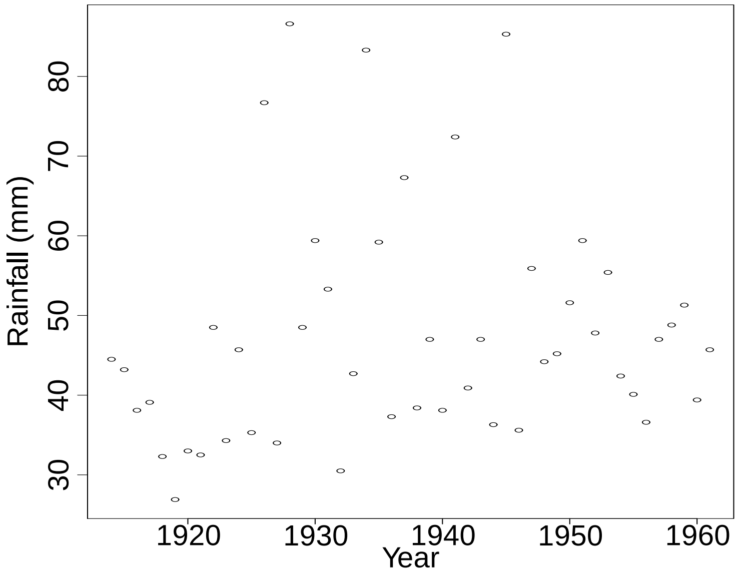

Figure 1: Annual maxima of daily rainfall totals at a location in South West England

1.2.4 Inferences for the block maxima approach

Here our blocks have $n = 365$, which is reasonably large, so we fit Model (1) to the $N = 48$ annual maxima (e.g. using *maximum likelihood estimation*). We obtain fitted parameter values (standard errors in parentheses):

$$\mu = 40.7(1.5) \qquad \sigma = 9.4(1.2) \qquad \xi = 0.14(0.12).$$

More importantly, we can make inferences on the quantities most useful to practitioners $\ldots$. For example, the 99th percentile in the distribution of annual maxima is known as the *100 year return level*. The fitted value of this is easily obtained on inversion of Model (1):

$$q_{100} = 101.3(18.9).$$

1.2.5 Remarks about the block maxima approach

- We don't need to deal explicitly with normalisation constants. We don't even need to know n!
- The assumption of n independent and identically distributed variables in each block is cavalier, but inferences are surprisingly robust.
- The inferences on return levels are crucial for designers and engineers, to the extent they are built into legally binding codes of practice.

- In actual fact, the existing codes of practice are usually based on a very primitive version of the methods just described. Fits are often based on restricting to one of the Fisher–Tippett types, ignoring estimation uncertainty, and using an *ad hoc* interpolation of return levels across a network of sites.

- In any case the block–maxima approach is often *very* wasteful of data, leading to large uncertainties on return level estimates. This motivates a different approach ……

1.2.6 Diagnostics for the block maxima approach

The goodness–of–fit of the GEV model is most easily assessed using various diagnostic plots. Here we consider four plots:

1. **Probability plot:** the fitted value of the c.d.f. is plotted against the empirical value of the c.d.f. for each data point.

2. **Quantile plot:** the empirical quantile is plotted against the fitted quantile for each data point.

3. **Return level plot:** the return level (with error bars) is plotted against the return period. Each data point defines a sample point.

4. **Density plot:** the fitted p.d.f. is superimposed on a histogram of the data.

For our rainfall example, the diagnostic plots look like this …

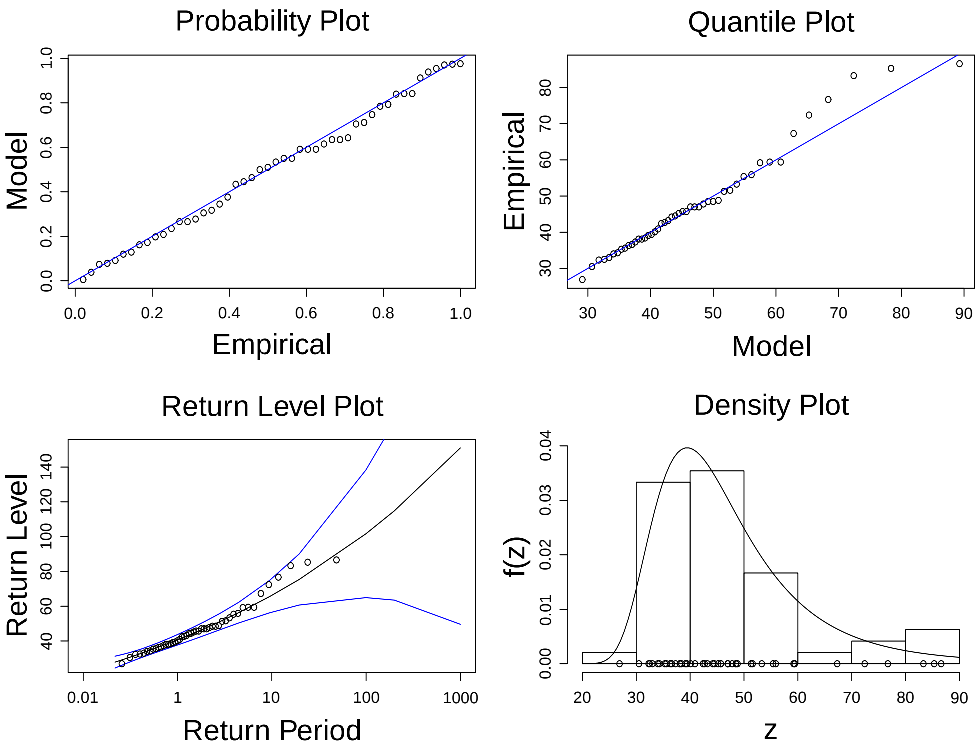

Figure 2: Diagnostic plots for GEV fit to rainfall annual maxima

1.2.7 Confidence intervals for return levels

Although we could construct a symmetrical confidence interval for the r–year return–level using classical likelihood theory ($\hat{q}_r \pm 1.96 \times$ standard error), this is not recommended. This practice assumes the limiting quadratic behaviour of the likelihood surface near the maximum, whereas in fact the surface is usually very asymmetrical.

We recommend using the method of *profile likelihood* to take this into account: by reparameterisation of Equation (1) to replace one of the parameters by q_r, we can maximise the likelihood *conditional* on q_r taking each possible value. We plot this constrained value against $q_r \ldots$

1.2.8 Profile likelihood confidence interval for q_{100}

For the rainfall example we get:

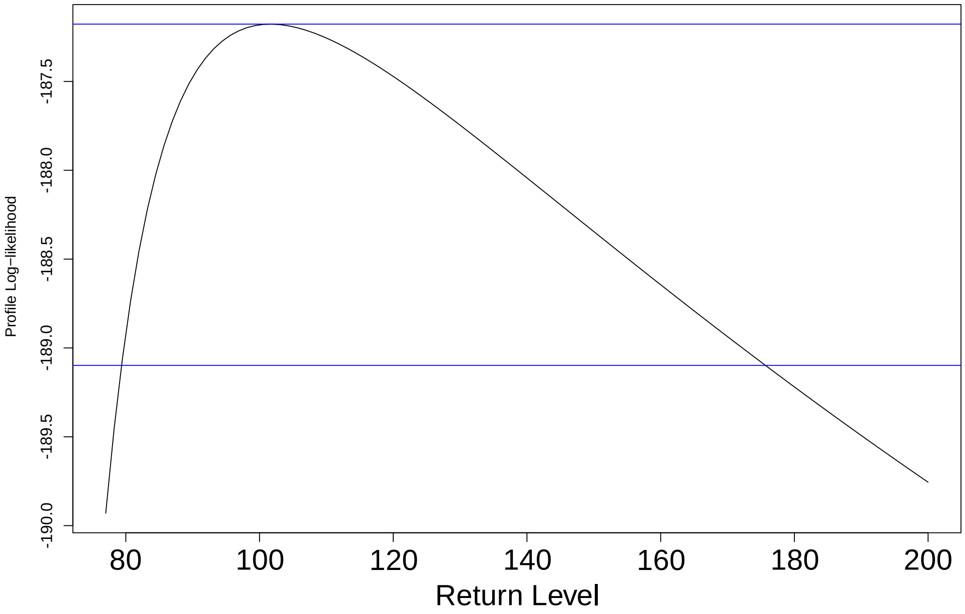

Figure 3: Profile log-likelihood for 100 year return level

The likelihood–ratio test can be applied directly to this likelihood surface by using a cut–off equal to $0.5 \times \chi^2_1(.)$. Here we see that the 95% confidence interval is approximately (78,176).

1.3 Threshold methods

Threshold methods use a more natural way of determining whether an observation is extreme - *all* values greater than some high value (*threshold*) are considered. This allows more efficient use of data, but brings its own problems. We must first go back and consider the asymptotic theory appropriate for this new situation.

1.3.1 The Generalised Pareto Distribution (GPD)

The appropriate limit theorem can be stated as follows:

Under very broad conditions, if it exists, any limiting distribution as $u \to \infty$ of $(X - u|X > u)$ is of Generalised Pareto Distribution (GPD) form (setting $Y = X - u$):

$$H(y) \;=\; 1 - \left(1 + \frac{\xi y}{\sigma}\right)_{+}^{-1/\xi}, \tag{3}$$

where $a_+ = \max(0, a)$ and σ $(\sigma > 0)$ and ξ $(-\infty < \xi < \infty)$ are scale and shape parameters respectively. Once again the GPD exists for $\xi = 0$, and is given by taking the limit of (3) as $\xi \to 0$. This time we get

$$H(y) \;=\; 1 - \exp\left(\frac{-y}{\sigma}\right), \tag{4}$$

defined for $y > 0$. This shows that when $\xi = 0$, the GPD is in fact the Exponential Distribution with mean equal to the scale parameter σ ($\sigma > 0$).

1.3.2 Return levels for the threshold excesses approach

If the GPD is a suitable model for exceedances of a threshold u by a random variable X, then for $x > u$,

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

It follows that

$$\Pr\{X > x\} = \lambda_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}. \tag{5}$$

where $\lambda_u = \Pr\{X > u\}$. So the level x_m that is exceeded once every m observations is the solution of

$$\lambda_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}.$$

Rearranging this we obtain

$$x_m = u + \frac{\sigma}{\xi}[(m\lambda_u)^{\xi} - 1],$$

so long as m is large enough to ensure that $x_m > u$. Now if there are n_y observations per year, then by setting $m = N \times n_y$, the N–year return level is obtained as

$$z_N = \mu + \frac{\sigma}{\xi}[(N n_y \lambda_u)^{\xi} - 1] \tag{6}$$

or when $\xi = 0$,

$$z_N = u + \sigma \log(N n_y \lambda_u),$$

and standard errors can be obtained using the delta method.

Approach 2: “Exceedances over thresholds”

In practice, modelling might typically proceed as follows:

1. Choose some threshold u_0 which is high enough so that the GPD (3) is a good model for $(X - u_0|X > u_0)$.

2. Fit the GPD to the observed excesses $x - u_0$.

3. Use the fitted GPD, together with some model for the rate of exceedances $X > u_0$, to provide estimates for *return levels* using (6).

1.3.3 Example: daily rainfall totals

For the rainfall data we used before, now consider the daily totals themselves.

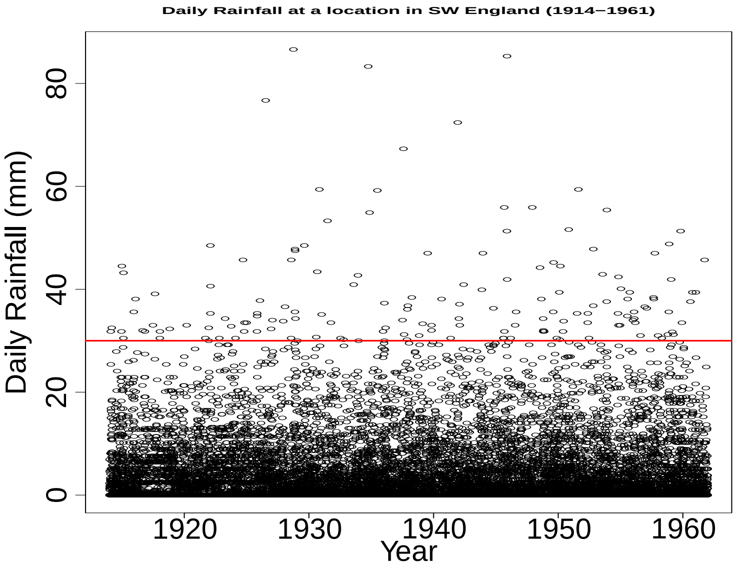

Figure 4: Daily Rainfall (1914-1961)

1.3.4 Threshold choice: Mean residual life plot

We make use of the fact that if the GPD is the correct model for all the exceedances x_i above some high threshold u_0, then the *mean excess*, i.e. the mean value of $(x_i - u)$, plotted against $u > u_0$, should give a linear plot (Davison and Smith, 1990) [Because $E[X_i - u_0]$ is a linear function of $u : u > u_0$]. By producing such a plot for values of u starting at zero, we can select reasonable candidate values for u_0.

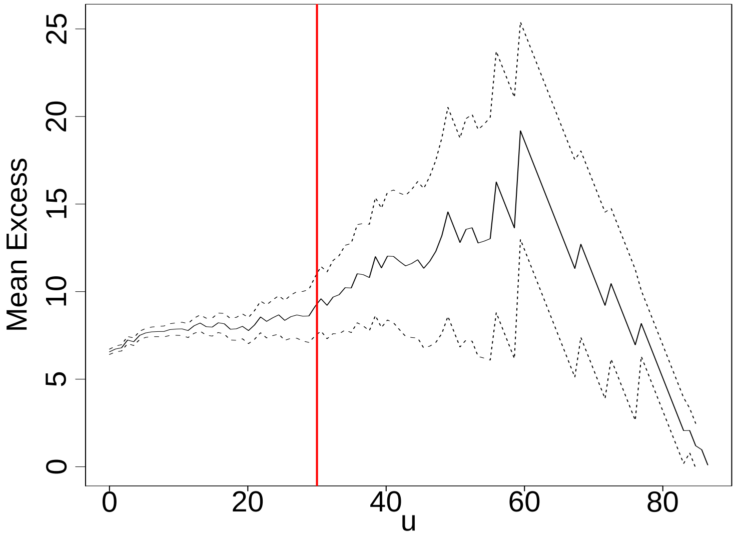

Figure 5: Mean residual life plot for daily rainfall

1.3.5 Inferences for the rainfall threshold excesses

Model (3) turns out to work reasonably well for all the excesses above $u_0 = 30$*mm*. This gives 152 exceedances $x_i; i = 1, \ldots, 152$, and Model (3) is fitted to the values $(x_i - u)$, again using maximum likelihood. We get

$$\sigma = 7.44(0.96) \qquad \xi = 0.18(0.10).$$

Assuming a uniform rate of exceedances, we estimate the 100–year return level: $q_{100} = 106.3(20.8)$.

1.3.6 Diagnostics for the rainfall threshold excesses

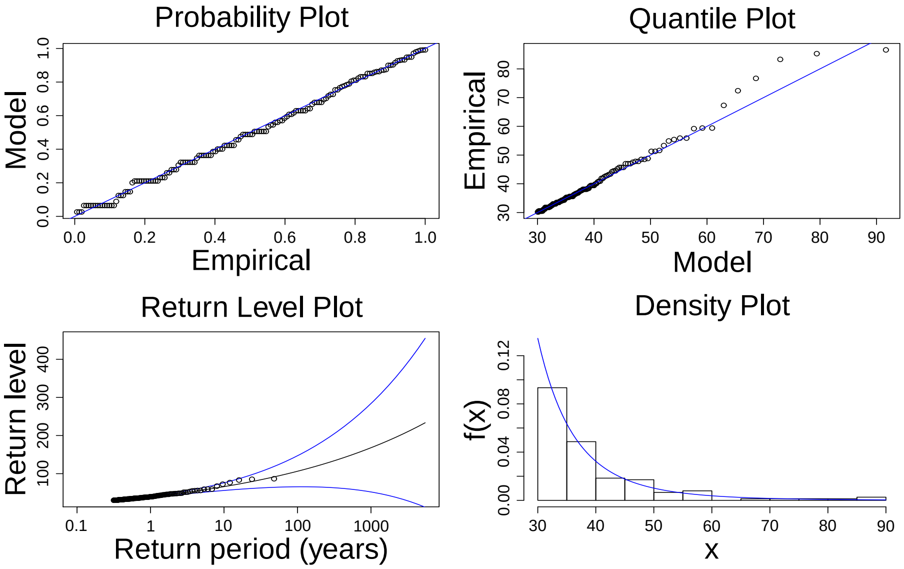

Figure 6: Diagnostic plots for the the threshold exceedance model for rainfall

1.3.7 Profile likelihood confidence interval for q_{100}

From the graph below, the 95% confidence interval is approximately (81,184).

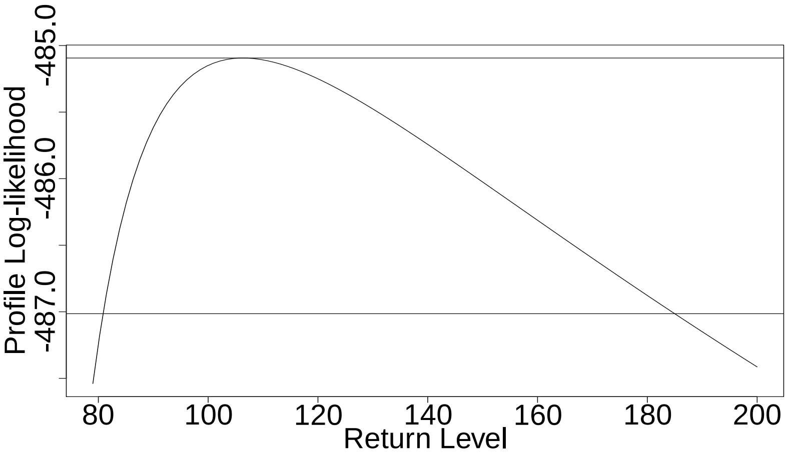

Figure 7: Profile log–likelihood for q_{100} based on threshold excess model

1.3.8 Threshold choice revisited

If the GPD with shape parameter ξ and scale parameter σ_{u_0} is the correct model for excesses over u_0, then for any threshold $u > u_0$, the excesses will be GPD with shape parameter ξ, and scale parameter

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0).$$

If we now use a modified version of the scale parameter,

$$\sigma^* = \sigma_u - \xi u,$$

we can see that both σ^* and ξ should be constant over thresholds greater than u_0 if we model excesses $x_i - u$ for $u > u_0$ using the GPD. This provides us with a further tool for assessing our original choice of threshold u_0.

1.3.9 Parameter stability plots

We refit the GPD for a range of thresholds upwards of u_0, and investigate the stability of our estimates of ξ and σ^*. 95% confidence intervals are shown by vertical lines, and help us assess the significance of any variation we see.

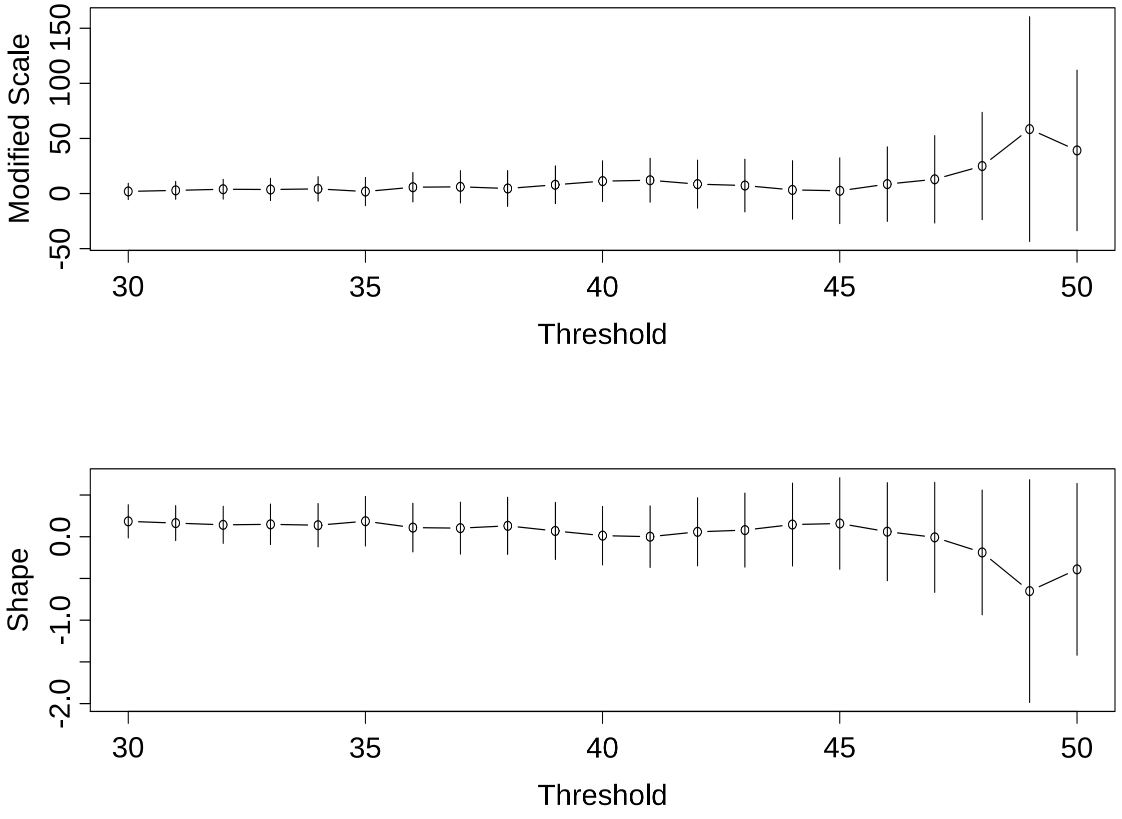

Figure 8: Parameter stability plots for the threshold model for rainfall

We can be reassured about our original choice of $u_0 = 30$!

2 Dependence and non–stationarity

The asymptotic results introduced in Part 1 have assumed the underlying process to be independent and identically distributed (i.i.d.). 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. In Part 2, we will consider some of these in detail.

2.1 Extremes of dependent sequences

For the types of data to which extreme value models are commonly applied, temporal independence is usually an unrealistic assumption. In particular, extreme conditions often persist over several consecutive observations, bringing into question the appropriateness of models such as the GEV. A detailed investigation of this requires mathematical treatment at a level of sophistication beyond which we have time to capitulate in this short course; however, the general ideas are not difficult and the main result offers a simple, practical, interpretation. For the remainder of this section on dependent sequences, we shall assume that our process is *stationary*, corresponding to a series whose variables may be mutually dependent, but whose stochastic properties are homogeneous throughout time.

Dependence in stationary sequences can take many different forms. With practical applications in mind, it is common to assume a condition that limits the extent of dependence to short–range temporal dependence so that, for example, events X_i and X_j, both of which are extreme, are independent provided time points i and j are far enough apart. Indeed, many stationary sequences satisfy this property. By excluding the possibility of long–range dependence in this way, we focus our attention on dependence at a much shorter range. Effects of such short–range dependence, it turns out, can be quantified within the standard extreme value limits discussed in Part 1.

2.1.1 Maxima of stationary 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 the Definition below.

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

$$\begin{aligned} \Big| \text{Pr}\left\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\right\} & \\ -\text{Pr}\left\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\right\} \text{Pr}\left\{X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\right\} \Big| & \leq \alpha(n, l), \end{aligned} \qquad (7)$$

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

For sequences of independent variables, the difference in probabilities in the above expression 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 (7), while not zero, is sufficiently close to zero to have no effect on the limit laws for extremes.

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

$$\text{Pr}\left\{(\tilde{M}_n - b_n)/a_n \leq x\right\} \to G^\theta(x) \qquad (8)$$

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

The parameter θ is known as the *extremal index*, and quantifies the extent of extremal dependence: $\theta = 1$ for a completely independent process, and $\theta \to 0$ with increasing levels of (extremal) dependence. Since G in the above theorem is necessarily an extreme value distribution, and due to the *max–stability* property (see Leadbetter *et al.*, 1983), 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.

The above theorem implies that if maxima of a stationary series converge – which, from Part 1, we know they will do – then, provided an appropriate $D(u_n)$ condition is satisfied, the limit distribution is related to the limit distribution of an independent series. The effect of dependence, as seen in expression (8), is just a replacement of G as the limit distribution with G^θ. In fact, if G corresponds to the GEV distribution with parameters (μ, σ, ξ), then

$$\begin{aligned} G^\theta(z) &= \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}^\theta \\ &= \exp\left\{-\left[1 + \xi\left(\frac{z-\mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\}, \end{aligned}$$

where $\mu^* = \mu - \frac{\sigma}{\xi}\left(1 - \theta^{-\xi}\right)$ and $\sigma^* = \sigma\theta^\xi$. Thus, if the (approximate) distribution of M_n is GEV with parameters (μ, σ, ξ), then the (approximate) distribution of $\tilde{M}_n$ is GEV with parameters (μ^*, σ^*, ξ).

2.1.2 Modelling block maxima

Provided long–range dependence is weak, we can proceed to model block maxima from series with short–range extremal dependence as outlined in Part 1, since the distribution of block maxima falls within the same family of distributions as would be appropriate if the series were truly independent. This is fantastic news! Short–range temporal dependence is a much more plausible assumption than complete independence, and our modelling approach is still valid! However, the main difference – excluding the change in parameters from (μ, σ, ξ) to (μ^*, σ^*, ξ) – is that our implied n (the number we are taking the maxima over) is now effectively reduced due to the dependence, so convergence of maxima to the limit distribution will be slower. And shouldn't we be using threshold methods anyway, which use information on *all* extremes and not just those that are the maximum within their block?

2.1.3 Modelling threshold exceedances

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, including:

1. filtering out an (approximately) independent set of threshold exceedances
2. fitting the GPD to *all* exceedances, ignoring dependence, but then appropriately adjusting the inference to take into account the reduction in information
3. Explicitly modelling the temporal dependence in the process

Though the first approach above is by far the most widely–used, our research has focussed on the relative merits of the other two approaches. The third approach makes use of multivariate extreme value theory, and so we shall re–visit this idea in more detail in Parts 4 and 5 this afternoon. For now, let us consider the first two approaches, which we will call *removing* dependence and *ignoring* dependence, respectively.

2.1.4 Example: Cluster peaks or all excesses?

Figure 9 shows a series of 3–hourly measurements of sea–surge heights at Newlyn, a coastal town in the southwest of England, collected over a three year period. The sea–surge is the meteorologically induced non–tidal component of the still–water level of the sea. The practical motivation for the study of such data is that structural failure — probably a sea–wall in this case — is likely under the condition of extreme surges. Also shown in Figure 9 is a plot of the time series against the lag 1 time series.

A natural way of modelling extremes such time series is to use the Generalised Pareto Distribution (GPD) as a model for excesses over a high threshold. As already discussed in Part 1, this approach might be preferable to the block maxima approach which is highly wasteful of data (and precious extremes!). Figure 9 also shows the presence of substantial temporal dependence in the sequence of three–hourly surges. We will now consider approaches **1** and **2**, outlined above, to circumvent this problem.

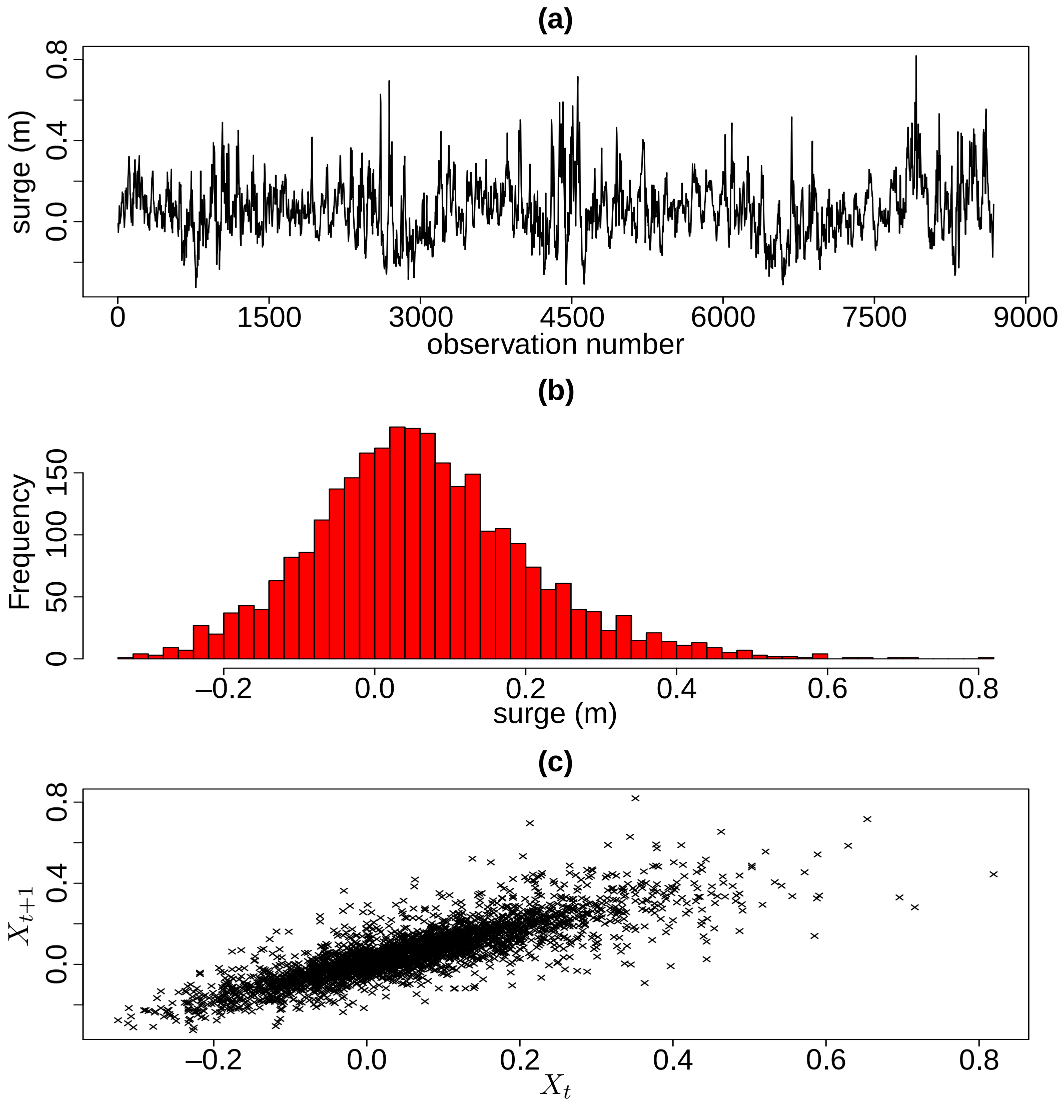

Figure 9: Newlyn sea–surge data: (a) Time series plot; (b) histogram; (c) plot of the time series against the series at lag 1.

‘Removing’ dependence

The most commonly adopted approach to circumvent the problems caused by such temporal dependence is to employ a declustering scheme to filter out a set of approximately independent threshold excesses. One method, which is often considered to be the most ‘natural’ way of identifying ‘clusters’ of extremes, is ‘runs–declustering’. This is how it works:

1. Choose an auxiliary ‘declustering parameter’ (which we call κ)

2. A cluster of threshold excesses is then deemed to have terminated as soon as at least κ consecutive observations fall below the threshold

3. Go through the entire series identifying clusters in this way

4. The maximum (or ‘peak’) observation from each cluster is then extracted, and the GPD fitted to the set of cluster peak excesses.

This approach is often referred to as the *peaks over threshold* approach (POT, Davison and Smith, 1990) and is widely accepted as the main pragmatic approach for dealing with clustered extremes. Although this approach is quite easy to implement, there are issues surrounding the choice of κ; if

- κ is too small, the cluster peaks will not be far enough apart to safely assume independence
- κ is too large, there will be too few cluster exceedances on which to form our inference

It has also been shown that parameter estimates can be sensitive to the choice of κ. In this example, we use a separation interval of 60 hours (and so $\kappa = 20$) following the example of Coles and Tawn (1991), which should be large enough to safely assume independence between successively identified clusters allowing for wave propagation time. We used a mean residual life plot (see Part 1) to identify a suitably high threshold (0.3m).

The table below shows maximum likelihood estimates of the GPD scale and shape parameters σ and ξ, along with the associated 95% confidence intervals, fitted to the set of cluster peak excesses using $\kappa = 20$. Shown for comparison are the corresponding estimates using *all* threshold exceedances, ignoring temporal dependence. Note the discrepancy in the estimation of the two parameters under the two approaches; however, when allowing for sampling variability, these differences are not significant.

	$\hat{\sigma}$	$\hat{\xi}$
Cluster peaks	0.187	–0.259
95% confidence interval	(0.109, 0.265)	(–0.545, 0.027)
All excesses	0.104	–0.090
95% confidence interval	(0.084, 0.125)	(–0.215, 0.035)

Table 1: Maximum likelihood estimates, and associated 95% confidence intervals, for the GPD scale and shape parameters

‘Ignoring’ dependence

Table 1 above shows that, although there is a slight discrepancy in parameter estimation when using (i) cluster peak exceedances and (ii) *all* exceedances, these discrepancies are non–significant. Therefore, why bother declustering? Surely we’re better off using *all* excesses?

The confidence intervals for the estimates using all excesses are too narrow – fitting to all exceedances when there is clearly evidence of short–term temporal dependence will result in underestimated standard errors. Smith (1991) suggests a procedure in which the usual asymptotic likelihood calculations are supplemented by empirical information on dependence, in order 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, denote the observed information matrix by H. If independence were a valid assumption, then the covariance matrix of the maximum likelihood estimates (m.l.e.s) would be approximately H^{-1}. Smith (1991) shows that to account for dependence this approximation should be replaced by $H^{-1}VH^{-1}$, where V is the covariance matrix of the likelihood gradient vector. Furthermore, V can be estimated by decomposing the log–likelihood sum into its contributions by year (which should be independent up to a good approximation) and obtaining the appropriate covariance matrix empirically.

Similar arguments can be applied to modify the procedure for testing hypotheses. Specifically, denoting model parameters by $\psi = (\rho, \zeta)$ where ρ and ζ are of dimensions p and q respectively, suppose that a test of $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$ is required, ζ being a nuisance parameter. Assuming independence, test procedures are usually based on the asymptotic distribution of

$$2\{\ell(\hat{\psi}_1) - \ell(\hat{\psi}_0)\}, \tag{9}$$

which is χ^2_p. Here, $\ell(\hat{\psi}_0)$ and $\ell(\hat{\psi}_1)$ denote the log–likelihood evaluated at the maximum likelihood estimate under H_0 and H_1 (respectively). Now suppose we wish to account for dependence. Partitioning

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

where H_{11}, H_{12}, H_{21} and H_{22} are the appropriate sub–matrices of dimensions $p \times p$, $p \times q$, $q \times p$ and $q \times q$ respectively, then we partition the inverse of H as

$$H^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix},$$

where each sub–matrix $H^{\cdot\cdot}$ has the same dimensions as $H_{\cdot\cdot}$. Now let

$$C = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} - H_{22}^{-1} \end{pmatrix}.$$

Then Smith (1991) shows that the approximate distribution of expression (9) is given by

$$\sum_{i=1}^{p} \lambda_i z_i^2 \tag{10}$$

where the z_i, $i = 1, \ldots, p$, are standard normal variates and the λ_i are the non–zero eigenvalues of $V^{1/2}CV^{1/2}$. This replaces the usual χ^2_p–distribution, which is valid in the case of independence, and which would be recovered if all the λ_i were set equal to 1. It is then easy to simulate from the modified distribution (10) to estimate any required quantile of the test statistic. Profile likelihood confidence intervals then arise as the set of values of $\hat{\psi}_1$ such that the test statistic (9) is smaller than the quantile which represents the desired level of significance.

Table 2 reports maximum likelihood estimates for the GPD scale and shape parameters, along with their 95% confidence intervals, for analyses using *all excesses* and just *cluster peak excesses* (as before); in the analysis using information on all extremes, though, standard errors have now been inflated to account for temporal dependence via Smith's method (1991).

Table 3 shows maximum likelihood estimates for return levels for four return periods — $s = 10$, 50, 200 and 1000 years. The corresponding 95% confidence intervals have been obtained using the method of profile likelihood, where the appropriate cut–off for the test statistic (9) has been obtained using the modified distribution (10). In this way the profile likelihood confidence intervals have been inflated to account for the dependence in a way which is consistent with the modifications proposed by Smith (1991). Figure 10 shows a plot of the profile likelihood for one of these return levels — z_{50} — illustrating the severe asymmetry which is commonly observed for return levels. This plot is for the analysis using all threshold exceedances. The 95% profile likelihood confidence interval for z_{50}, after adjusting for dependence, is identified on the plot. Also shown is the much narrower interval which would have been obtained if dependence had been ignored.

Table 2 shows that, when the analysis is restricted to a set of cluster peak exceedances, the GPD scale parameter σ is overestimated, and the shape parameter ξ underestimated, relative to the approach which uses all exceedances. However, when we account for sampling variability, we see that these differences are not significant.

Of greater practical interest are the estimated return levels. Table 3 shows that estimates barely differ for the ten year return period, but are consistently smaller in the cluster peaks analysis for the other three periods studied — in fact, quite substantially so for the 200 and 1000 year return periods. Since estimates of such long–range return levels are often used as a design requirement in oceanographic situations (e.g. for the height of sea walls), designing to a level specified by an analysis based on cluster peak excesses could result in substantial under–protection.

	$\hat{\sigma}$	$\hat{\xi}$
Cluster peaks	0.187	–0.259
95% Confidence Interval	(0.109, 0.265)	(–0.545, 0.027)
All excesses	0.104	–0.090
95% Confidence Interval	(0.082, 0.126)	(–0.217, 0.037)

Table 2: Maximum likelihood estimates, and associated Wald 95% confidence intervals, for the GPD scale and shape parameters and the threshold exceedance rate when using all excesses, and just cluster peak excesses.

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{200}$	$\hat{z}_{1000}$
Cluster peaks	0.868	0.920	0.951	0.975
95% Confidence Interval	(0.770, 1.031)	(0.813, 1.099)	(0.838, 1.008)	(0.858, 1.063)
All excesses	0.867	0.947	1.007	1.068
95% Confidence Interval	(0.736, 1.067)	(0.790, 1.193)	(0.844, 1.257)	(0.891, 1.335)

Table 3: Maximum likelihood estimates, and associated 95% profile likelihood confidence intervals, for four return levels (units are in metres).

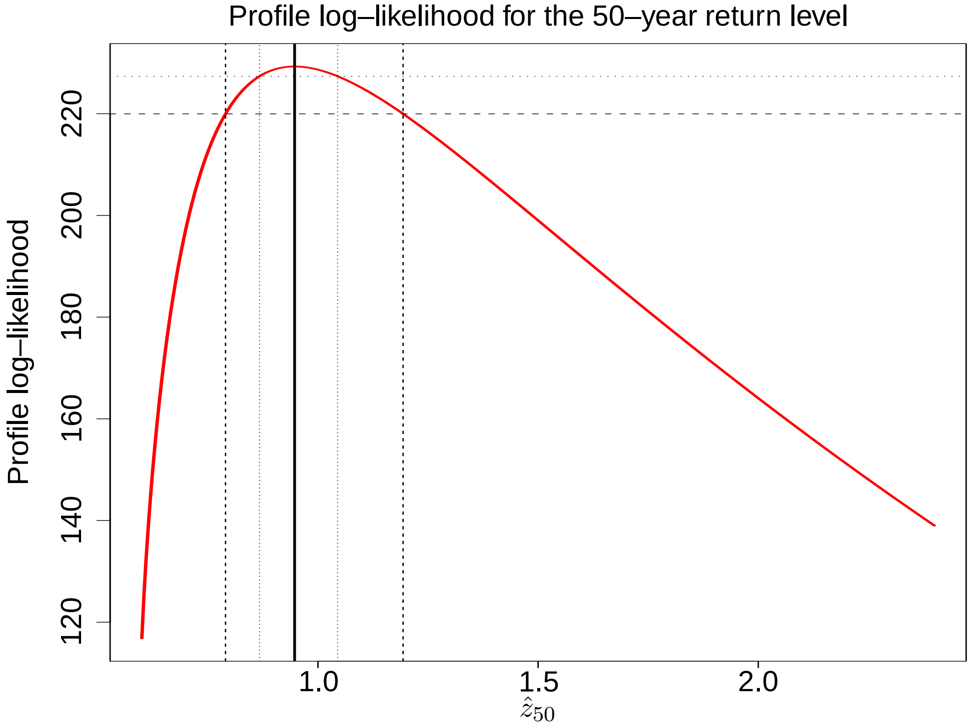

Figure 10: Profile log–likelihood surface, with corresponding 95% confidence intervals, for the 50 year return level $\hat{z}_{50}$. The dashed lines show the construction of the interval which has been inflated to account for temporal dependence in the sea–surge data (since in this example all threshold excesses were used). The dotted lines show how the interval would be constructed if dependence had been ignored.

Simulation study

So we know there are differences – some significant – in return level estimation when we use (i) cluster peak excesses and (ii) all threshold excesses. Which approach are we to trust?

— The usual approach is to use cluster peaks, then we have effectively removed temporal dependence

— However, return levels using this approach are underestimated relative to the procedure which uses all threshold excesses

— Using cluster peak excesses could result in substantial under–protection (i.e. not building a sea–wall high enough to protect against the 1 in 1000 year surge)

Figure 11 below shows some results of a simulation study undertaken by Fawcett and Walshaw (2007), in which the GPD was fitted to a simulated dataset for which the true values of σ, ξ and various return levels were *known*, and the strength of temporal dependence was similar to that of which is often observed in real–life environmental time series. The bold lines correspond to sampling distributions for the GPD parameters (and two return levels) using all threshold excesses, the thin lines correspond to the equivalent when using just cluster peak excesses. Clearly, for all parameters, the analysis using all threshold excesses outperforms that which uses just cluster peak excesses. Of most concern are the result shown for the two return levels; Fawcett and Walshaw (2007) found systematic underestimation of return levels when using cluster peak excesses (remember, this is the approach most commonly adopted to circumvent the problem of temporal dependence), whereas estimates of these return levels were much more accurate under the approach using all threshold excesses.

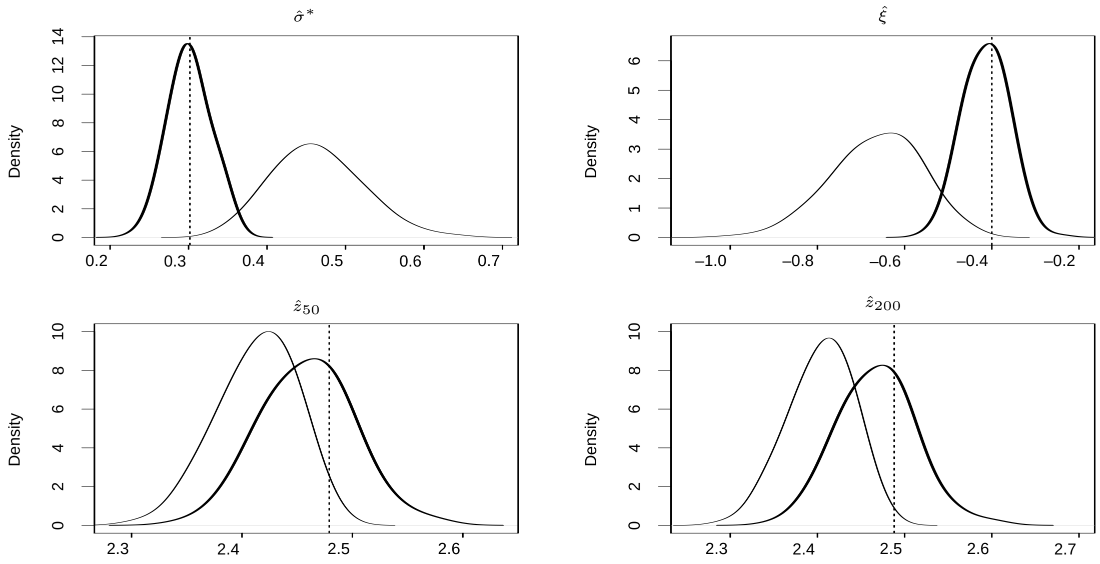

Figure 11: Sampling distributions of $\hat{\sigma}^*$, $\hat{\xi}$, $\hat{z}_{50}$ and $\hat{z}_{200}$ when $\alpha = 0.2$, and using all threshold excesses (heavy line) and cluster peak excesses (thin line). The *true* values for each parameter are shown by the vertical lines.

2.2 Non–stationarity: trend

In Section 2.1 we demonstrated that, subject to specified limitations, the usual extreme value limit models are still applicable in the presence of short–term temporal dependence. In fact, we can use the results for block maxima directly as they stand, though some thought is required when considering threshold models. The general theory cannot be extended for non–stationary series; instead, it is usual to adopt a pragmatic approach of using the standard extreme value models as basic templates that can be augmented by statistical modelling.

Figure 12 (over-leaf) below shows a time series plot of annual maximum sea levels observed at Fremantle, Western Australia, between 1900 and 1986; the right–hand–side plot shows these sea–levels plotted against the annual mean value of the *Southern Oscillation Index* (SOI), which is a proxy for meteorological volatility. There appears to be an increase in annual maximum sea levels through time, as well as an association between annual maximum sea levels and the mean SOI.

We can accommodate the time–trend shown in the plot on the left–hand–side of Figure 12 by fitting the GEV distribution (as we have annual maxima), but allowing for a linear trend in the underlying level of extreme behaviour. For example, if we define Z_t to be the annual maximum sea level at Fremantle in year t, then we might use

$$Z_t \quad \sim \quad \text{GEV}\,(\mu(t), \sigma, \xi)$$

where

$$\mu(t) \quad = \quad \beta_0 + \beta_1 t. \tag{11}$$

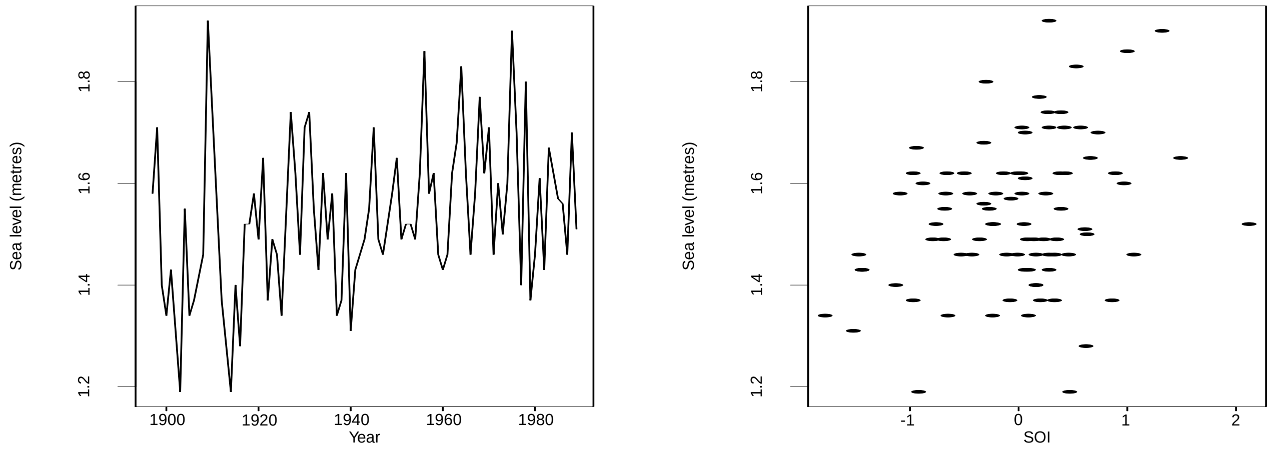

Figure 12: Time series plot of annual maximum sea levels observed at Fremantle (left), and a plot of the mean SOI against annual maximum sea level (right).

In this way, variations through time in the observed process are modelled as a linear trend in the location parameter of the appropriate extreme value model (the GEV in this case). We might choose to adopt the following model for $\mu(t)$:

$$\mu(t) \quad = \quad \beta_0 + \beta_1 \text{SOI}(t)$$

to allow for a linear association between the maximum sea level in year t and the SOI in year t. Or perhaps a textitmultiple linear regression model for $\mu(t)$, whereby

$$\mu(t) \quad = \quad \beta_0 + \beta_1 t + \beta_2 \text{SOI}(t); \tag{12}$$

we can then assess our preferences between the stationary model ($\mu(t) = \beta_0$), the models which allow for a dependence in time (alone), a dependence on SOI through time (alone), and the model which allows the underlying extremal behaviour to be determined by *both* a change in time *and* SOI, by referring to the usual likelihood ratio tests (since these models are nested). For example, fitting a stationary GEV distribution to these data, we get:

$$\hat{\mu} = 1.482(0.017) \qquad \hat{\sigma} = 0.141(0.011) \qquad \hat{\xi} = -0.217(0.064),$$

with a maximised log–likelihood of 43.6. Fitting the model which allows for a trend in time (the model shown in 11), we get:

$$\hat{\beta}_0 = 1.387(0.027) \qquad \hat{\beta}_1 = 0.002(0.0005) \qquad \hat{\sigma} = 0.124(0.010) \qquad \hat{\xi} = -0.128(0.068)$$

with a maximised log–likelihood of 49.79. Referring

$$\begin{aligned} D &= 2\{49.79 - 43.6\} \\ &= 12.38 \end{aligned}$$

to χ^2_1 tables, we have a significant result, suggesting that the model which includes a linear trend in time for μ explains substantially more of the variation in the data than the stationary model. Figure 13 shows the time series plot of the Fremantle sea level data with fitted estimates for μ superimposed. Also shown, for comparison, is the fitted estimate for μ under the stationary model.

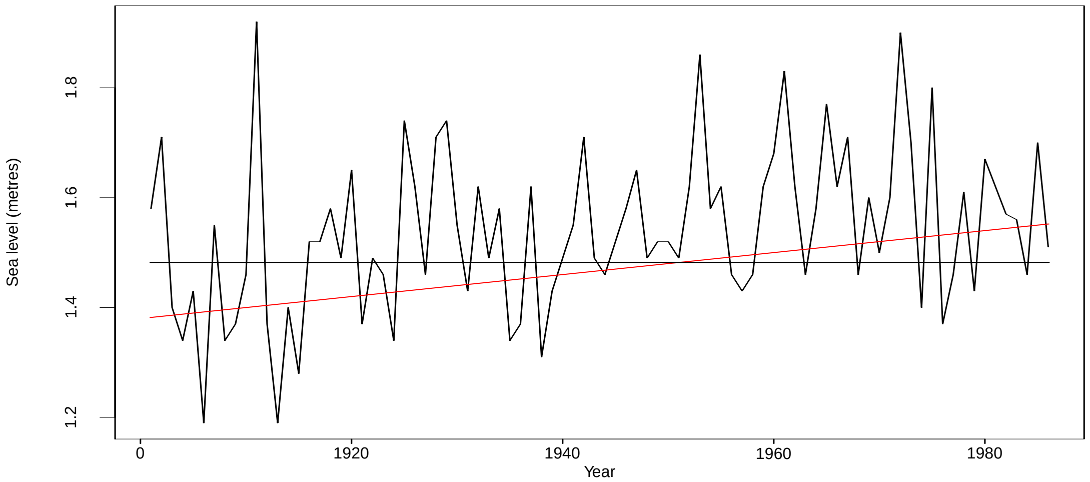

Figure 13: Time series plot of annual maximum sea levels observed at Fremantle, with fitted estimates for μ based on the stationary GEV model and the model which allows for a linear trend in time.

Similar methodology actually suggests that the model in equation 12 is the best model to use here, i.e. that which allows for a trend in μ depending on both time and SOI. In fact, we get:

$$\hat{\beta}_0 = 1.389(0.027) \qquad \hat{\beta}_1 = 0.002(0.0005) \qquad \hat{\beta}_2 = 0.055(0.020)$$
$$\hat{\sigma} = 0.121(0.010) \qquad \hat{\xi} = -0.154(0.064)$$

giving

$$\hat{\mu} \quad = \quad 1.389 + 0.002t + 0.055\text{SOI}(t).$$

Of course, more exotic model structures can be incorporated into this framework, including quadratic models, higher–order polynomial models, and models which allow for non–normal error structures. Trend can also be incorporated into the other GEV/GPD model parameters.

2.3 Non–stationarity: 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, 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.

3 R session: Weather extremes

To get started, you will need to be seated at a computer with R installed, and initiate R, which is usually done through menus selected from the Start menu, or an icon. In addition the libraries ismev and evd should be installed. We will connect these, and install some of our own supplementary routines, using the commands

```
> library(ismev)
> library(evd)
> source('Rstufflee.r')
```

Provided these all go through without a hitch, we are ready to go!

1. In this question, we will do a simple analysis of annual maximum wind speeds recorded at Boston, Massachusetts, for 50 years from 1936 to 1985.

(a) Provided you have the file boston.txt in your working directory, this can be loaded into Rusing the command:

```
> boston<-scan('boston.txt')
```

We have now created an R object called boston which is a single column containing consecutive years with annual maximum wind speeds in mph. We can have a look at this by simply typing:

```
> boston
```

(b) We now wish to separate out the year and maximum components into separate vectors. This can be done using the commands:

```
bosyear<-as.numeric(boston[seq(1,length(boston),2)])
bosmax<-as.numeric(boston[seq(2,length(boston),2)])
```

which has the effect of creating vectors bosyear and bosmax containing the years and maxima respectively. [Note that when entering consecutive similar commands in R, it is convenient to use the up arrow to bring up the previous command and then edit it!] We can check the vectors by simply typing:

```
> bosyear
> bosmax
```

(c) Now we can have a look at the annual maxima over time using the command:

```
> plot(bosyear,bosmax)
```

If you like you can give your plot some nice labels:

```
> plot(bosyear,bosmax,xlab='Year',ylab='Wind speed (mph)',
    main='Annual maximum wind speeds at Boston MA')
```

(d) We are now ready to carry out an extreme value analysis on the annual maxima. Since these are regarded as observations on i.i.d. random variables, we can forget about the vector `bosyear`. We fit the GEV to the data in `bosmax`:

```
>gev.fit(bosmax)
```

Notice the output:

- `$conv` gives a value of zero (in row [1] of the output), which indicates successful convergence, i.e. no errors in fitting;
- `$nllh` shows the negative (maximised) log–likelihood;
- `$mle` shows the maximum likelihood estimates for μ, σ and ξ respectively;
- `$se` gives the associated standard errors for these parameters.

(e) We can investigate the model performance using the in–built diagnostics. First we must store the relevant information from the fit in an object we name ourselves, e.g.

```
> fit1<-gev.fit(bosmax) > gev.diag(fit1)
```

creates the 'fit' object `fit1` and then runs the diagnostic routines on the stored object. Make sure you interpret the four plots in the context of Section 1.2.6.

(f) We can obtain inference on return levels using the additional command which we have supplied in `Rstufflee.R`, which is called `gev.ret(data,period)`. This command refits the GEV model, and then provides us with the inference on the specified return level. E.g. for the 100–year level q_{100}, we would type:

```
> gev.ret(bosmax,100)
```

In addition to the information we obtained earlier, we get the 100–year return level estimate with associated standard error. Notice how this matches up with the return level plot in the diagnostic plots.

(g) If we want to construct a confidence interval for q_{100}, we are better off using the method of profile–likelihood as described in Section 1.2.8. We can use the function `gev.prof(fit,period,lower-bound, upper-bound)`. This command is slightly unstable, and relies on an appropriate choice of the bounds for the profile–likelihood. For the Boston annual maxima, the following works well for the 100–year level:

```
> gev.prof(fit1,100,75,130)
```

Note that this enables us to read off the 95% confidence interval (the default) for q_{100}. Suppose we wanted a 99% interval we would use:

```
> gev.prof(fit1,100,<lower>,<upper>,conf=0.99)
```

for appropriate choices of `<lower>` and `<upper>`. You may like to experiment. Note how asymmetrical these intervals are, and how misleading it would be to base the confidence intervals on $\pm 1.96($s.e.$)$!

2. In this question, we will analyse annual maximum sea levels (in cm) observed at Venice, Italy, between the years 1931 and 1981 (inclusive).

(a) Load the data into R by typing:

```
> data(venice)
```

Now look at the data by typing

```
> venice
```

You should see a matrix with 51 rows (one for each of the years 1931–1981) and 11 columns. The values in each column correspond to the year, and the *ten* largest sea levels observed in each of these years (in descending order) For example, in 1979, the ten largest sea levels were: 166, 140, 131, 130, 122, 118, 116, 115, 115, 112, the largest being 166cm.

(b) We intend to fit the Generalised Extreme Value distribution to the set of annual maxima – i.e. the largest sea levels only (166cm in 1979, for example). Extract the set of annual maxima in the following way:

(i) Create a new vector to store the set of annual sea level maxima by typing:

```
> maxima<-vector('numeric', length=51)
```

(ii) Now type:

```
> maxima<-venice[,2]
```

which will store the observations from column 2 in `venice` – i.e. the largest sea levels from each year – in the vector `maxima`.

We can fit the Generalised Extreme Value distribution to the set of annual maxima using the function `gev.fit`. Type

```
> gev.fit(maxima)
```

Write down the maximum likelihood estimates of μ, σ and ξ, along with their estimated standard errors. Also make a note of the value of the maximised log–likelihood.

(c) Now produce a time series plot of the set of annual maxima by typing

```
> plot(maxima∼venice[,1],type='l',xlab='Year',ylab='Sea level
(cm)')
```

which will plot the annual maxima against the first column in `venice`, which corresponds to the year. This will also provide convenient labels for both the x and y axes in the plot. Does the time series plot of annual maxima look stationary?

(d) We will now attempt to model variations through time in the sequence of annual sea level maxima by modelling a linear trend in the location parameter μ, i.e. $\mu(t) =$

$\beta_0 + \beta_1(t)$, where t represents the time–point (so $t = 1$ corresponds to 1931, etc.) Set up a time matrix by typing:

```
> time<-matrix(1:51,ncol=1)
```

Now type

```
> gev.fit(maxima, ydat=time, mul=1)
```

which tells R to use the matrix `ydat` as a matrix of covariates, and `mul=1` tells R which column in that matrix to use (as well as which parameter to use it for – μ!). Write down the maximum likelihood estimates for β_0, β_1, σ and ξ, along with their estimated standard errors, and make a note of the maximised log–likelihood.

(e) Use the maximised log–likelihood values from parts (b) and (d) to perform a likelihood ratio test to see if the model which allows for a trend provides a significant improvement over the stationary fit (Hint: $\chi^2_1(5\%) = 3.84$).

(f) Write down the simple linear regression equation for μ found from the fit in part (d), i.e. $\mu(t) = \beta_0 + \beta_1(t)$. We will now write an R function to calculate the fitted trend at each time point, and then superimpose this on the plot produced in part (c). Type

```
> trend.plot<-vector('numeric',51)
```

The vector `trend.plot` will take the fitted values of the trend for μ obtained from the equation. Now write

```
> for(i in 1:51)
+ {
+ trend.plot[i]<-beta0+beta1*time[i,1]
+ }
```

where `beta0` and `beta1` should be replaced with the estimated values found in the fit in part (d). Now type

```
> lines(trend.plot~venice[,1])
```

which should superimpose a plot of the trend line against the year on the original time series plot.

3. In this question we will investigate the use of “Peaks Over Threshold” to circumvent the problems of serial dependence when modelling threshold exceedances. We will do this by examining hourly gust maximum wind speeds observed at High Bradfield, a location in the Peak District in central northern England.

(a) These data were collected by the U.K. Meteorological Office, and are not included with any of the standard R packages. Thus, to load the data, type

```
> gusts<-scan('bradfield.txt')
```

which will store the data in a vector called `gusts`. Now produce a time series plot of these data, by typing

```
> plot(ts(gusts))
```

The data you see correspond to the hourly gust maximum wind speeds (in knots) collected over a ten–year period (1975–1984 inclusive) in the month of January; thus, the first observation is the maximum gust wind speed observed between midnight and 01:00 on the 1st January 1975, etc. We restrict our analysis to January because the U.K. has a seasonally varying wind climate, and the strongest wind speeds are usually observed in the month of January (i.e. in January we observe ’genuine’ extremes of wind speed). Comment on the nature of this time series.

(b) We will now investigate the extent of temporal dependence in the series.

(i) Type

`> acf(gusts)` and
`> pacf(gusts)`

These commands will produce plots of the autocorrelation, and *partial* autocorrelation function.

(ii) Now type

```
> plot(gusts[1:7259]~gusts[2:7260])
```

This will produce a plot of the time series against the series at lag 1 (the length of this dataset is 7260).

Using your plots in (i) and (ii) above, comment on the degree of short–term temporal dependence present in the series.

(c) We now intend to fit the Generalised Pareto Distribution (GPD) to a set of threshold exceedances. Use the command

```
> mrl.plot(gusts)
```

to produce a mean residual life plot for the gust data, and use this to choose an appropriate threshold for identifying extremes.

(d) Now fit the GPD to the set of threshold exceedances, by using

```
> gpd.fit(gusts,threshold)
```

where `threshold` is your chosen threshold from the mean residual life plot in part (c). Make a note of the estimates for σ and ξ (as well as their estimated standard errors).

(e) Now *decluster* the series of gusts and employ a *Peaks Over Threshold* analysis. Type

```
> cluster.peaks<-cluster10(gusts,threshold)
```

again, where `threshold` is the threshold identified in part (c). The function `cluster10` uses a value of $\kappa = 10$ observations to identify clusters of extremes, i.e. a cluster of extremes is deemed to have terminated as soon as at least 10 observations fall below the threshold. Now fit the GPD to the set of cluster peak excesses, and make a note of the parameter estimates and estimated standard errors.*[Note: you can vary the declustering interval κ by using different functions, e.g.* `cluster20` *or* `cluster30`*]*

(f) We will now calculate the *threshold exceedance rate* for each of the approaches in parts (d) and (e). Typing

```
> length(gusts[gusts>threshold])/length(gusts) and
> length(cluster.peaks)/length(gusts)
```

where `threshold` is as before, will work out the threshold exceedance rate λ_u for *all* excesses, and *cluster peak* excesses, respectively. Write down these threshold exceedance rates.

(g) You should now compare estimates of the 1000–observation return level using (i) all threshold excesses and (ii) cluster peak excesses. Typing

```
> gpd.ret(data,threshold,1000)
```

but replacing `data` with `gusts` and then `cluster.peaks` (and the `threshold` is that identified in part (c)) will estimate this value for *all* excesses and *cluster peak* excesses, respectively. The output produced will be the same as before – i.e. you will get estimates of the GPD parameters and their standard errors, but now you will also get an estimate of the specified return level (and its standard error via the delta method).

(h) Comment on your estimates of the 1000–observation return level in part (g) and your GPD parameter estimates in parts (d) and (e). Which approach to inference do you trust most?

4 Multivariate extremes

4.1 Introduction

In this section we consider the problems we face if we wish to model the extremal behaviour of two or more (dependent) processes simultaneously. There are several reasons why we may wish to do this:

- to model the extreme behaviour of a particular variable over several nearby locations (e.g. rainfall over a network of sites);
- to model the joint extremes of two or more different variables at a particular location (e.g. wind and rain at a site);
- to model the joint behaviour of extremes which occur as consecutive observations in a time–series (e.g. consecutive hourly maximum wind gusts during a storm).

All of these problems suggest fitting an appropriate limiting multivariate distribution to the relevant data. However, as we shall see, the derivation of such a multivariate distribution is not as easy as we might hope. The analogy with the Normal distribution as a model for means breaks down as we move into n dimensions! It is not even clear what the 'relevant data' should be! Most of the increased complexity is apparent in the move from 1 to 2 dimensions, so we will focus largely on bivariate problems.

4.2 Componentwise maxima models

4.2.1 Example: network of rainfall measurements

Suppose we want to study the joint extremes of daily rainfall accumulations at the network of 8 sites shown in Figure 14.

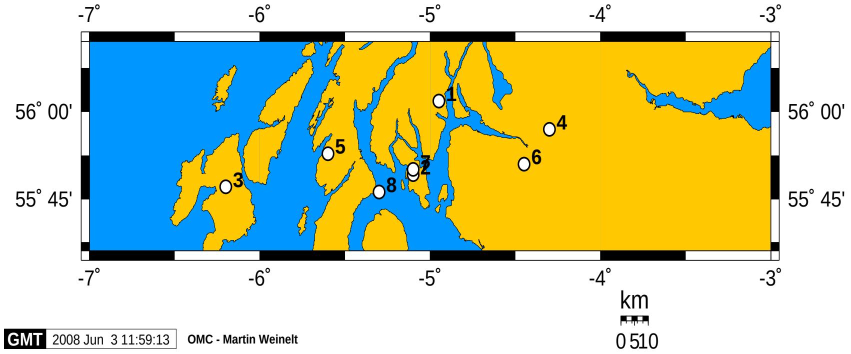

Figure 14: Eight rainfall recording stations in southern Scotland

Such issues are of great interest, especially currently, e.g. given the severe flooding experienced in the UK recently. Suppose we have sequences of daily total rainfall at each location. There is liable to be strong inter–site dependence in extremes, in the sense that days with heavy rain are liable to occur simultaneously across locations. The raw multivariate observations are 8–dimensional vectors of the daily rainfall over the eight sites.

Now suppose we wish to take a block–maxima approach, with 'blocks' being years. For any given year, the 8–dimensional vector of annual maxima is unlikely to be one of the raw multivariate observations. Let's simplify to the bivariate case. Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. vectors with distribution function $F(x, y)$. Now consider the componentwise block maxima

$$M_{x,n} = \max_{i=1,\ldots,n} \{X_i\} \qquad \text{and} \qquad M_{y,n} = \max_{i=1,\ldots,n} \{Y_i\}.$$

We define the *vector of componentwise maxima* to be

$$\mathbf{M_n} = (M_{x,n}, M_{y,n}).$$

$\mathbf{M_n}$ is not necessarily one of the original observations (X_i, Y_i). Nevertheless, we are interested in the limiting behaviour of $\mathbf{M_n}$ as $n \to \infty$. The first point to note is that standard univariate extreme value results apply in each margin. When considering the dependence, this allows us to make a simplifying assumption.

We assume that the X_i and Y_i variables have a known marginal distribution. It is convenient to assume this is the GEV(0,1,1) distribution, also known as the unit Fréchet distribution, which has c.d.f.

$$F(z) = \exp(-1/z), \qquad z > 0.$$

This gives rise to a very simple normalization of maxima:

$$\Pr(X_i < x) = \Pr(M_{x,n}/n < x) = \exp(-1/x), \qquad x > 0,$$

(and similarly for Y_i). So if we consider the re–scaled vector

$$\mathbf{M_n^*} = \left(\max_{i=1,\ldots,n} \{X_i\}/n, \max_{i=1,\ldots,n} \{Y_i\}/n \right),$$

the margins are unit Fréchet for all n, and hence we can characterize the limiting joint behaviour of $\mathbf{M_n^*}$ without having to worry about the margins. Unfortunately no limiting parametric family exists! (for bivariate extremes, or multivariate extremes in general).

4.2.2 Theorem: limiting distributions for bivariate extremes

Let $\mathbf{M_n^*} = (M_{x,n}^*, M_{y,n}^*)$ be the normalized maxima as above, where the (X_i, Y_i) are i.i.d. with standard Fréchet marginal distributions. Then if

$$\Pr(M_{x,n}^*, M_{y,n}^*) \to G(x, y),$$

where G is non–degenerate, then G has the form

$$G\left(x, y\right) = \exp\left\{-V\left(x, y\right)\right\}; \quad x > 0, \;\; y > 0 \tag{13}$$

where:

$$V\left(x, y\right) = 2 \int_0^1 \max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right) dH\left(\omega\right) \tag{14}$$

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

$$\int_0^1 \omega \; dH\left(\omega\right) = 0.5. \tag{15}$$

Hence the class of bivariate extreme value distributions is in one-to-one correspondence with distribution functions H satisfying the constraint (15). If H is differentiable with density h, then (14) becomes

$$V\left(x,y\right)=2\int_0^1 \max\left(\frac{\omega}{x},\frac{1-\omega}{y}\right)h(\omega)d\omega.$$

However some simple models arise when H is not differentiable. E.g. if H places mass 0.5 on each of $\omega=0$ and $\omega=1$, then we get

$$G(x,y)=\exp\{-(x^{-1}+y^{-1})\},\qquad x>0,y>0,$$

corresponding to independent x and y.

Since the GEV provides the complete class of marginal limit distributions, then the complete class of bivariate extreme value distributions is obtained as follows. If we suppose X and Y are GEV with parameters (μ_x,σ_x,ξ_x) and (μ_y,σ_y,ξ_y) respectively, then the transformations

$$\tilde{x}=\left[1+\xi_x\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^{1/\xi_x}\quad\text{and}\quad\tilde{y}=\left[1+\xi_y\left(\frac{y-\mu_y}{\sigma_y}\right)\right]^{1/\xi_y}$$

obtain unit Fréchet margins. Hence

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

is a bivariate extreme value distribution with the appropriate margins for valid $V(.)$, and provided $[1+\xi_x(x-\mu_x)/\sigma_x]>0]$ and $[1+\xi_y(x-\mu_y)/\sigma_y]>0]$.

4.2.3 Modelling bivariate extremes in practice

In practice, modelling usually involves identifying a parametric sub–family with appropriate flexibility to handle the structure inherent in the data. Models can be fitted, e.g. by maximum–likelihood estimation, either in two steps (marginal components followed by dependence function), or in a single sweep. All of these procedures, including the choice of models, are handled in a very similar way when dealing with threshold exceedances. We consider the details in the next section.

4.3 Threshold excess models

We want to define our bivariate extremes in those observations which exceed a threshold in one or other margin. For our bivariate observation (X,Y), let's focus on X. We have already seen that the distribution function for the exceedances of a threshold u by a variable X, conditional on $X>u$ for large enough u, is given by:

$$G(x)=1-\lambda\left\{1+\frac{\xi\left(x-u\right)}{\sigma}\right\}^{-1/\xi}$$

defined on $\{x-u: x-u>0 \text{ and } \left(1+\xi\left(x-u\right)/\sigma\right)>0\}$, where $\xi\neq 0$, $\sigma>0$, and $\lambda=Pr\left(X>u\right)$. Now we can obtain a unit Fréchet margin with the transformation:

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

If we apply the analogous transformation to in the Y margin, we obtain

$$\tilde{F}\left(\tilde{x},\tilde{y}\right)=\exp\left\{-V\left(\tilde{x},\tilde{y}\right)\right\};\quad x>u_x,\;\; y>u_y,$$

where:

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

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

$$\int_0^1\omega\;dH\left(\omega\right)=0.5.$$

4.3.1 Example: wave–surge data

Here we choose a different type of example of dependence to the rainfall problem considered in Section 4.2. Here we consider two variables recorded concurrently at the same site. A series of 3-hourly measurements on sea–surge were obtained from Newlyn, southwest England. For suitably high thresholds, we can identify which observations are extreme.

4.3.2 Threshold representation

Bivariate threshold models are complicated by the possibility that a bivariate pair (x,y) may be an ‘exceedance’ and yet exceed the specified threshold in only one of the two components.

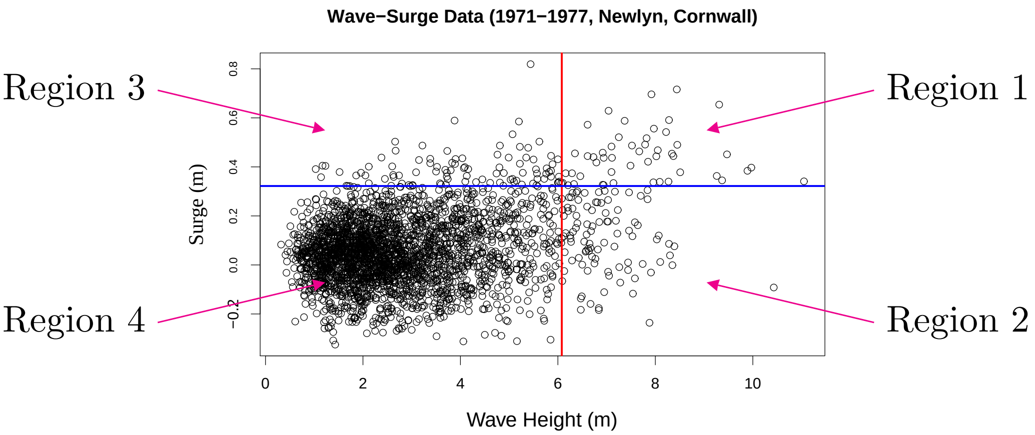

Figure 15: Threshold classification of bivariate data

4.3.3 Modelling the dependence structure

The class of bivariate extreme value models contains many families of distributions which can be used to model the dependence structure in the data. The dependence structure must satisfy the conditions on $H\left(\omega\right)$. Possible choices are:

- Logistic Model — symmetric
- Negative Logistic Model
- Bilogistic Model — asymmetric
- Dirichlet Model

Here we will focus on the logistic model and the bilogistic model as two commonly used but contrasting choices.

4.3.4 The Logistic model

$$G(x,y) = \exp\left\{-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}\right\}$$

where $x > 0$, $y > 0$ and $\alpha \in (0,1)$.

- $\alpha \to 1$ corresponds to independent variables.
- $\alpha \to 0$ corresponds to perfectly dependent variables.
- This model is symmetric — the variables are exchangeable.

4.3.5 The Bilogistic model

$$G(x,y) = \exp\left\{x\gamma^{1-\alpha} + y(1-\gamma)^{1-\beta}\right\}$$

where $0 < \alpha < 1$, $0 < \beta < 1$ and $\gamma = \gamma(x, y; \alpha, \beta)$ is the solution of:

$$(1-\alpha)\,x\,(1-\gamma)^{\beta} = (1-\beta)\,y\gamma^{\alpha}$$

- Independence is obtained when $\alpha = \beta \to 1$ and when one of α or β is fixed and the other approaches 1.
- When $\alpha = \beta$ the model reduces to the logistic model.
- The value of $\alpha - \beta$ determines the extent of asymmetry in the dependence structure.

4.3.6 Likelihood calculations

- For points in Region 1, the bivariate model structure shown applies, and the density of $\tilde{F}(\tilde{x}, \tilde{y})$ gives the appropriate likelihood component.
- In other regions, the likelihood component for the points must be censored.

4.3.7 The likelihood function

The likelihood function can be written as:

$$L(\theta; (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{n} \psi(\theta; (x_i, y_i))$$

where θ gives the parameters of F and

$$\psi(\theta; (x,y)) = \begin{cases} \left.\frac{\partial^2 F}{\partial x \partial y}\right|_{(x,y)} & \text{if } (x,y) \in \text{Region 1} \\ \left.\frac{\partial F}{\partial x}\right|_{(x,u_y)} & \text{if } (x,y) \in \text{Region 2} \\ \left.\frac{\partial F}{\partial y}\right|_{(u_x,y)} & \text{if } (x,y) \in \text{Region 3} \\ F(u_x, u_y) & \text{if } (x,y) \in \text{Region 4} \end{cases}$$

The various models can be fitted to data by maximum likelihood estimation using routines available in the R package evd. We will explore this in the second R practical.

4.4 Point process representation

It helps our understanding of bivariate (and hence multivariate) extremes to think in terms of a point process model as follows. Let $(x_1, y_1), (x_2, y_2), \ldots$ be a sequence of independent bivariate observations form a distribution with standard Fréchet margins such that

$$\Pr\{M^*_{x,n} \leq x, M^*_{y,n} \leq y\} \rightarrow G(x, y).$$

Let N_n be a sequence of point processes defined by

$$N_n = \{(n^{-1}x_1, n^{-1}y_1), \ldots, (n^{-1}x_n, n^{-1}y_n)\}.$$

Then

$$N_n \rightarrow N$$

on regions bounded away from $(0, 0)$, where N is a non–homogeneous Poisson process on $(0, \infty) \times (0, \infty)$. Moreover, if we change our coordinates to an angular-radial form ('pseudo-polar') by setting

$$r = x \quad \text{and} \quad \omega = \frac{x}{x + y},$$

then the intensity function of N is

$$\lambda(r, \omega) = 2\frac{dH(\omega)}{r^2},$$

where H is related to G in the usual way (Equations (13) — (15)). This is helpful because r and ω are measures of distance (from the origin) and angle (from the x-axis) respectively, and the dependence function H determines the angular spread of points of N, *and is independent of radial distance*. If H is differentiable, then since ω measures the relative size of x to y in the pair (x, y), then $h(.)$ determines the density of events of different relative size. It is fairly easy now to picture what different densities $h(.)$ will look like it terms of the scatter of points in the limiting point process N.

4.4.1 The point process representation in practice

We assume the Poisson limit to be a reasonable approximation to N_n on an appropriate region. Convergence is guaranteed on any region bounded from the origin, and things are especially simple if we choose a region of from $A = \{(x, y) : x/n + y/n > r_0\}$ for suitably large r_0, since then

$$\Lambda(A) = 2\int_A \frac{dr}{r^2} dH(\omega) = 2\int_{r=r_0}^{\infty} \frac{dr}{r^2} \int_{\omega=0}^{1} dH(\omega) = 2/r_0,$$

which is constant with respect to the parameters of H. If we assume H has density h, then the likelihood is given by

$$\begin{aligned} L(\theta; (x_1, y_1), \ldots, (x_n, y_n)) &= \exp\{\Lambda(A)\} \prod_{i=1}^{N_A} \lambda(x_{(i)}/n, y_{(i)}/n) \\ &\propto \prod_{i=1}^{N_A} h(\omega_i), \end{aligned}$$

where $\omega_i = x_{(i)}/(x_{(i)} + y_{(i)})$ for the N_A points $(x_{(i)}, y_{(i)})$ which are in A. [This is based on assuming that we have already transformed the margins so that $(x_1, y_1), \ldots, (x_n, y_n)$ have standard Fréchet distributions.] Now we can fit the model using maximum–likelihood estimation.

4.4.2 Point process model for wave–surge data

A point process model was fitted to the wave–surge data after transformation to unit Fréchet margins, and using a threshold of the form $X + Y = r_0$, where r_0 was chosen so that the marginal thresholds are both at the 95th percentile. Fitting the two dependence models (logistic and bilogistic) to the wave–surge data we obtain the following results:

Model	log–lik.	α	β
Logistic	227.2	0.659 (0.013)	
Bilogistic	230.2	0.704 (0.024)	0.603 (0.032)

These results suggest a fairly weak, while clearly significant, dependence. The logistic and bilogistic models can be compared using a likelihood ratio test, and significant asymmetry is suggested. It is also possible to produce graphs of the fitted $h(\omega)$ functions, with the histograms of the empirical ω values super–imposed. Here we just show some dependence functions for the logistic model.

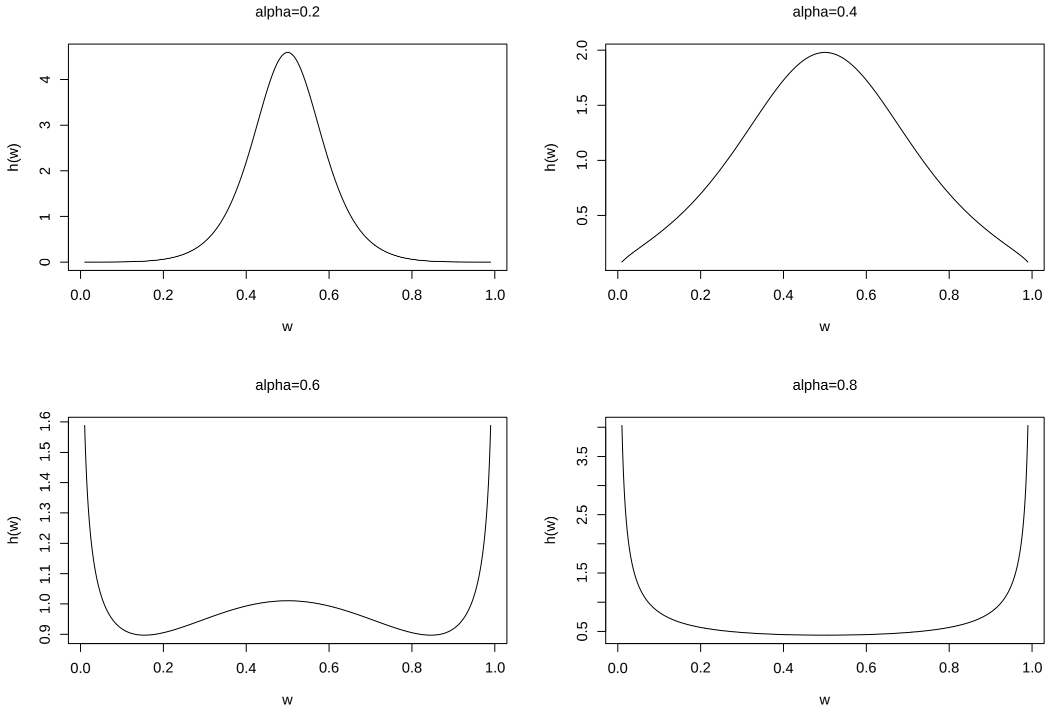

Figure 16: Some dependence functions for the logisitc model

4.5 Asymptotic dependence and independence

One key problem with using limit distributions for multivariate extremes is that they force one of two possibilities:

1. extremes occur independently in the different margins;

2. extremes occur with a dependence structure which conforms to an asymptotic extreme value distribution.

In practice this imposition is not helpful ... it is often the case that asymptotic independence is suggested by the data, and yet quite strong dependence is present, even at high levels. Data that seem to be dependent at ordinary levels may not necessarily be dependent in the limiting distribution. Consider the region $A = \left\{\frac{X}{n} > u, \frac{Y}{n} > v\right\}$. Then:

$$\Pr\left[\left(\frac{X}{n}, \frac{Y}{n}\right) \in A\right] = \begin{cases} C/n, & \textit{Asymptotic Dependence} \\ C/n^2, & \textit{Exact Independence} \end{cases}$$

where C is a constant term that does not depend on n.

4.5.1 The coefficient of tail dependence

Consider the variable:

$$T = \min\left(X, Y\right).$$

The distribution function of T is given by:

$$Pr\left(T \leq t\right) = 1 - \frac{K}{t^{1/\boldsymbol{\delta}}}, \qquad t > u,$$

where u is a threshold above which the data are regarded as extreme and K is a (almost) constant term with respect to t. $\boldsymbol{\delta}$ gives a measure of extremal dependence between X and Y and is known as the "**coefficient of tail dependence**".

4.5.2 Inference for $\boldsymbol{\delta}$

The likelihood function for T is:

$$L\left(K, \boldsymbol{\delta}; t\right) = \left(1 - \frac{K}{u^{1/\boldsymbol{\delta}}}\right)^{n-n_u} \left(\frac{K}{\boldsymbol{\delta}}\right)^{n_u} \prod_{i=1}^{n_u} t_i^{-(1+1/\boldsymbol{\delta})}$$

where n_u is the number of observations that satisfy $T > u$. Maximum likelihood estimation gives the estimate:

$$\hat{\boldsymbol{\delta}} = \frac{1}{n_u} \sum_{i=1}^{n_u} \log\left(\frac{t_i}{u}\right)$$

evaluated for the n_u points in the data set above u. $\boldsymbol{\delta}$ describes the limiting dependence structure:

- $\boldsymbol{\delta} = 1$ implies asymptotic dependence.
- $\frac{1}{2} < \boldsymbol{\delta} < 1$ implies positive association.
- $\boldsymbol{\delta} = \frac{1}{2}$ implies near independence.
- $0 < \boldsymbol{\delta} < \frac{1}{2}$ implies negative association.

4.5.3 Wave–surge data

Plots of $\hat{\boldsymbol{\delta}}$ against increasing u give an indication of the level of dependence present between two processes in the limiting distribution.

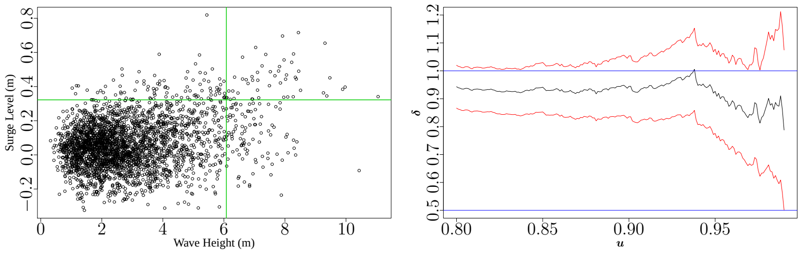

Figure 17: Wave-Surge data with 95% quantiles; $\boldsymbol{\delta}$–plot with 95% confidence bounds.

$\boldsymbol{\delta} = 1$ is within the 95% confidence bounds for all u as u increases, suggesting that wave–height and surge are **asymptotically independent**.

Research into modelling data in such instances, i.e. where there is still dependence within the 'extremes' in the data set, but yet asymptotic independence is suggested, is all fairly recent. The most prominent work is the article by Heffernan and Tawn (JRSS B, 2004). Here they develop semiparametric models based on assuming observations are extreme in at least one component, and then conditioning on this. This approach can be quite messy in implementation, combining as it does a range of different estimation procedures, and some *ad hoc* assumptions concerning the parametric forms of the key normalising constants. Here we briefly consider another approach, suggested by (Bortot *et al.*, 2000), and currently the subject of ongoing work by Atyeo and Walshaw.

4.5.4 The multivariate Gaussian tail model

The multivariate Gaussian tail model for the multivariate distribution function F is defined on the **joint tail region** (Bortot *et al.*, 2000):

$$R(\mathbf{u}) = (u_1, \infty) \times \ldots \times (u_p, \infty)$$

where $\mathbf{u} = (u_1, \ldots, u_p)$. (e.g. Region 1 in Figure 15). For each observation in the joint tail region $R(\mathbf{u})$ we transform each marginal observation to have a standard Normal marginal distribution, and then apply the p–dimensional standard Normal distribution function. We then transform *back* to extreme value margins. This provides a more realistic representation of the dependence, while retaining the asymptotic arguments for the marginal extremes.

We have been able to fit such models to the 8-dimensional rainfall problem associated with Figure 14, however inference for this problem was much simplified by adopting a Bayesian approach.

5 Bayesian inference for extremes

Throughout this short course, 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).

5.1 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 $\boldsymbol{x} = (x_1, \ldots, x_n)$ 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 $\boldsymbol{x}$ using the probability density function $f(\boldsymbol{x}; \boldsymbol{\psi})$. The likelihood function for $\boldsymbol{\psi}$ is therefore $L(\boldsymbol{\psi}|\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\psi})$. 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

$$\pi(\boldsymbol{\psi}|\boldsymbol{x}) \quad = \quad \frac{\pi(\boldsymbol{\psi})L(\boldsymbol{\psi}|\boldsymbol{x})}{f(\boldsymbol{x})}, \tag{16}$$

where

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

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

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

In equation (16), $\pi(\boldsymbol{\psi}|\boldsymbol{x})$ is the *posterior* distribution of the parameter vector $\boldsymbol{\psi}$, $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, 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 (16) can be problematic, and usually analytically intractable. There is nothing particularly special about the fact that equation (16) 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 (16). 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.

5.2 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.

5.2.1 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

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

are available for simulating from (ψ_{-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

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

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.

5.2.2 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
$$A(\boldsymbol{\psi}, \boldsymbol{\psi}') \quad = \quad \min\left\{1, \frac{\pi(\boldsymbol{\psi}')L(\boldsymbol{\psi}'|\boldsymbol{x})p(\boldsymbol{\psi}', \boldsymbol{\psi})}{\pi(\boldsymbol{\psi})L(\boldsymbol{\psi}|\boldsymbol{x})p(\boldsymbol{\psi}, \boldsymbol{\psi}')}\right\};$$
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.

5.2.3 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.

5.3 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

$$f(y|\boldsymbol{x}) \;=\; \int_{\boldsymbol{\Psi}} f(y|\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\boldsymbol{x})d\boldsymbol{\psi} \tag{17}$$

when $\boldsymbol{\psi}$ is continuous. From equation (17), 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 $\boldsymbol{x} = (x_1, \ldots, x_n)$. Thus, in the Bayesian framework, we would have

$$\Pr\left\{Y \leq y|x_1, \ldots, x_n\right\} \;=\; \int_{\boldsymbol{\Psi}} \Pr\left\{Y \leq y|\boldsymbol{\psi}\right\}\pi(\boldsymbol{\psi}|\boldsymbol{x})d\boldsymbol{\psi}. \tag{18}$$

Equation (18) gives the distribution of a future threshold excess, allowing for both parameter uncertainty and randomness in future observations. Solving

$$\Pr\left\{Y \leq q_{r,\text{pred}}|x_1, \ldots, x_n\right\} \;=\; 1 - \frac{1}{r}$$

for $q_{r,\text{pred}}$ therefore gives an estimate of the r–year return level that incorporates uncertainty due to model estimation. Though (17) 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 $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_B$ that can be regarded as realisations from the stationary distribution $\pi(\boldsymbol{\psi}|\boldsymbol{x})$. Thus

$$\Pr\left\{Y \leq q_{r,\text{pred}}|x_1, \ldots, x_n\right\} \;\approx\; \frac{1}{B}\sum_{i=1}^{B}\Pr\left\{Y \leq q_{r,\text{pred}}|\boldsymbol{\psi}_i\right\},$$

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 alternative.

5.3.1 Example: Annual maximum sea levels: Port Pirie, South Australia

Figure 18 shows a time series plot of annual maximum sea levels at another Australian location – Port Pirie, in South Australia. Notice that, unlike the corresponding data from Fremantle in Wester Australia, there doesn't appear to be any trend in this series; in fact, the series appear stationary.

We use the GEV as a model for the annual maximum sea levels at Port Pirie Z_i in year i, i.e.

$$Z_i \;\sim\; \text{GEV}\left(\mu, \sigma, \xi\right), \qquad i = 1, \ldots, 65.$$

When employing MCMC methods it is common to re–parameterise the GEV scale parameter and work with $\eta = \log(\sigma)$ to retain the positivity of this parameter. In the absence of any expert

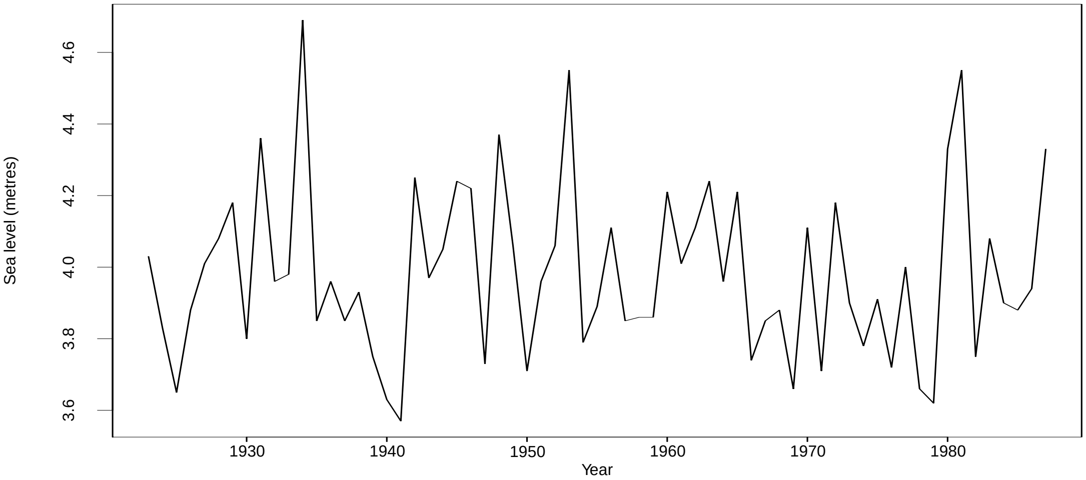

Figure 18: Time series plot of annual maximum sea levels observed at Port Pirie.

prior information regarding the three parameters of the GEV distribution, we adopt a 'naive' approach and use largely non–informative, independent priors for these, namely

$$\begin{aligned} \pi(\mu) &\sim N(0, 10000), \\ \pi(\eta) &\sim N(0, 10000) \qquad \text{and} \\ \pi(\xi) &\sim N(0, 100), \end{aligned}$$

the large variances of these distributions imposing near–flat priors.

We use a Metropolis–Hastings MCMC sampling scheme; after setting initial starting values for $\boldsymbol{\psi} = (\mu, \eta, \xi)$, we use an arbitrary probability rule $p(\boldsymbol{\psi}_{i+1}|\boldsymbol{\psi})$ for iterative simulation of successive values in the chain. Once this rule has been used to generate a candidate value $\boldsymbol{\psi}'$ for $\boldsymbol{\psi}_{i+1}$, we accept this with probability A (see 5.2.2); otherwise, $\boldsymbol{\psi}_{i+1} = \boldsymbol{\psi}_i$. Here, we use a *random walk* procedure to generate candidate values, i.e.

$$\begin{aligned} \mu' &= \mu_i + \epsilon_\mu \\ \eta' &= \eta_i + \epsilon_\eta \qquad \text{and} \\ \xi' &= \xi_i + \epsilon_\xi, \end{aligned}$$

with the ϵ being normally distributed with zero mean and variances v_μ, v_η and v_ξ respectively. In fact, the choice of algorithm and its 'tuning parameters' (v_μ, v_η and v_ξ) does not affect the model. It does, however, affect the efficiency of the algorithm. Some believe there is a 'fine art' to tuning the algorithm used, but it is common to aim for an overall acceptance rate of around 30%.

Initialising with $\boldsymbol{\psi}^{(0)} = (5, 0.5, 0.1)$, we get the following values generated by 5000 iterations of the MCMC scheme (see Figure 19). The settling–in period seems to take around 300 iterations, after which the chain seems to have converged. This settling–in period is often known as the *burn–in*. Thus, after deleting the first 300 simulations, the remaining 4700 simulated values can be treated as dependent realisations whose marginal distribution is the target posterior. Over-leaf, in Figure 20, is a panel of plots corresponding to the sampling distributions of the three GEV parameters (after the removal of burn–in), as well as the 100–year return level. The sampling distribution for the posterior of the return level has been obtained by inversion of the distribution function for the GEV (Equation 1) and then by repeated substitution of $\mu^{(301)}, \sigma^{(301)}, \xi^{(301)}, \ldots, \mu^{(5000)}, \sigma^{(5000)}, \xi^{(5000)}$.

The posterior means, standard deviations and 95% credible intervals are shown in Table 4, along with the corresponding maximum likelihood estimates for comparison.

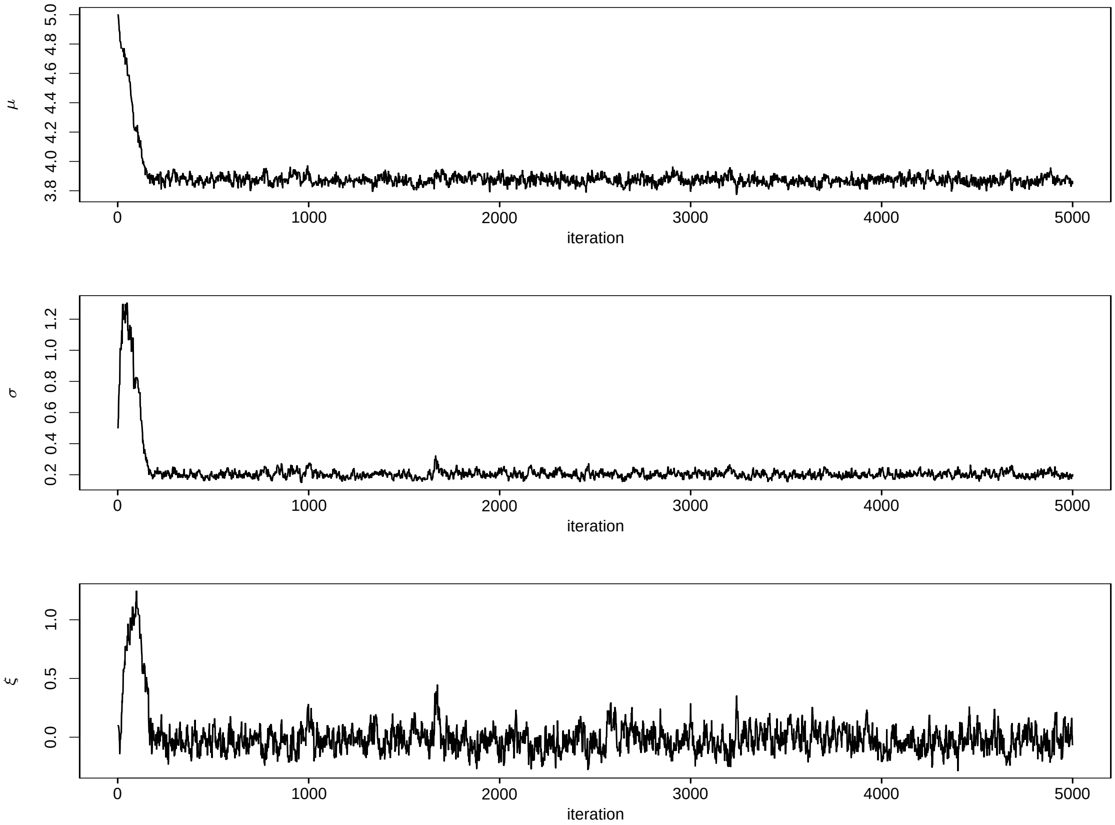

Figure 19: MCMC realisations of the GEV parameters in a Bayesian analysis of the Port Pirie sea level maxima.

		μ	σ	ξ	q_{100}
Posterior	mean (st. dev.)	3.874 (0.028)	0.203 (0.021)	–0.024 (0.098)	4.788 (0.255)
distribution	95% CI	(3.819, 3.932)	(0.166, 0.249)	(–0.196, 0.182)	(4.516, 5.375)
Maximum	m.l.e. (s.e.)	3.872 (0.028)	0.198 (0.020)	-0.040 (0.098)	4.692 (0.158)
likelihood	95% CI	(3.821, 3.930)	(0.158, 0.238)	(–0.242, 0.142)	(4.501, 5.270)

Table 4: Summary statistics for the posterior location, scale and shape, and the 100–year return level. Shown also, for comparison, are the corresponding m.l.e.s, the confidence interval for the return level being found via profile likelihood.

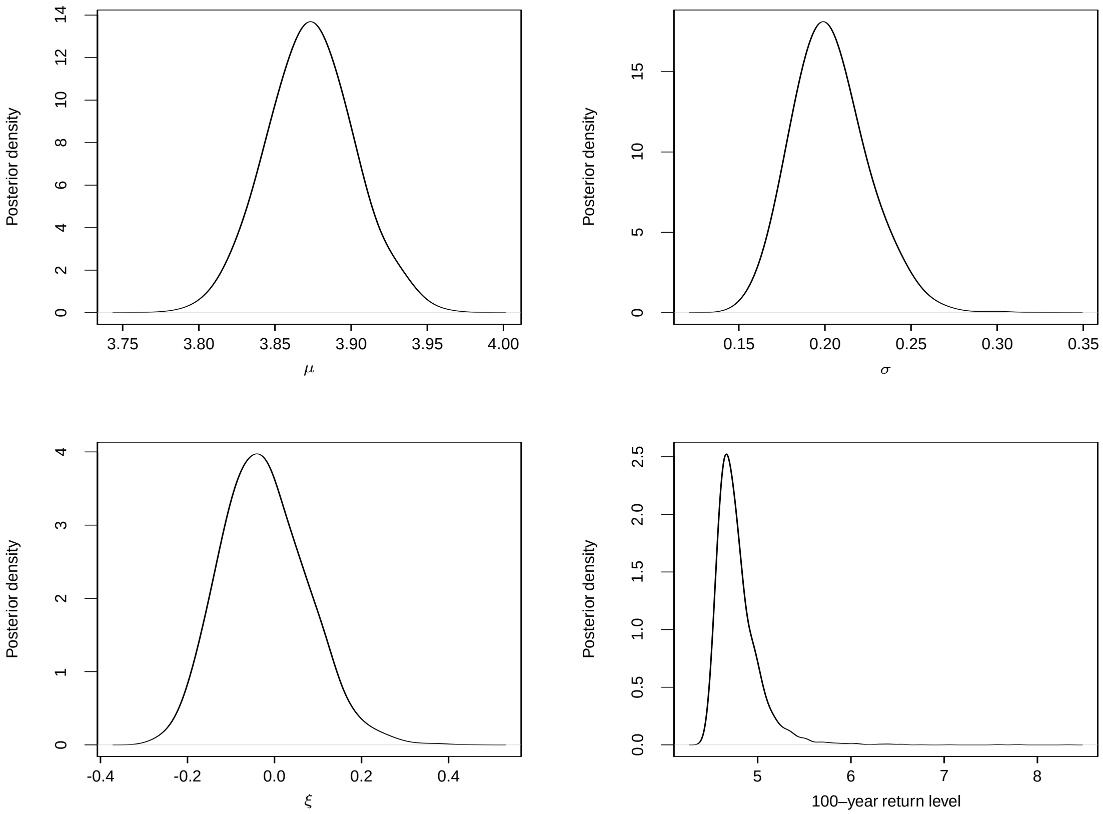

Figure 20: Sampling distributions for the posterior densities of μ, σ, ξ and the 100–year return level.

5.3.2 More complex structures: A random effects model for extreme wind speeds

In this section we briefly discuss the work which Lee Fawcett will present at next week's TIES conference. In this work, we develop a hierarchical model for hourly maximum wind speeds over a region of central and northern England. The data used consist of hourly gust maximum wind speeds recorded for the British Meteorological Office at twelve locations (see Figure 21). We construct a model which is based on a standard limiting extreme value distribution, but incorporates random effects for the sites, for seasonal variation, and for the serial dependence inherent in the time series of hourly maximum speeds obtained at each site. The Bayesian paradigm provides the most feasible modelling approach to capture the rich meteorological structure present in these data. Figure 22 illustrates an exploratory analysis of data from two contrasting sites, Nottingham and Bradfield. Shown are time series plots of the hourly maxima, histograms, and a plot of the time series against the version at lag 1. The first three years of data only are used in each case, to best illustrate the relevant data characteristics. We now (very briefly) outline the model structure used.

Modelling threshold exceedances

We will start with the Generalised Pareto Distribution as a model for threshold excesses; by doing so, we can incorporate more extreme data in our analysis than if we were to select "block maxima", and so increase the precision of our analysis. Thus, wind speed excesses over a high threshold will be modelled with a GPD(σ, ξ).

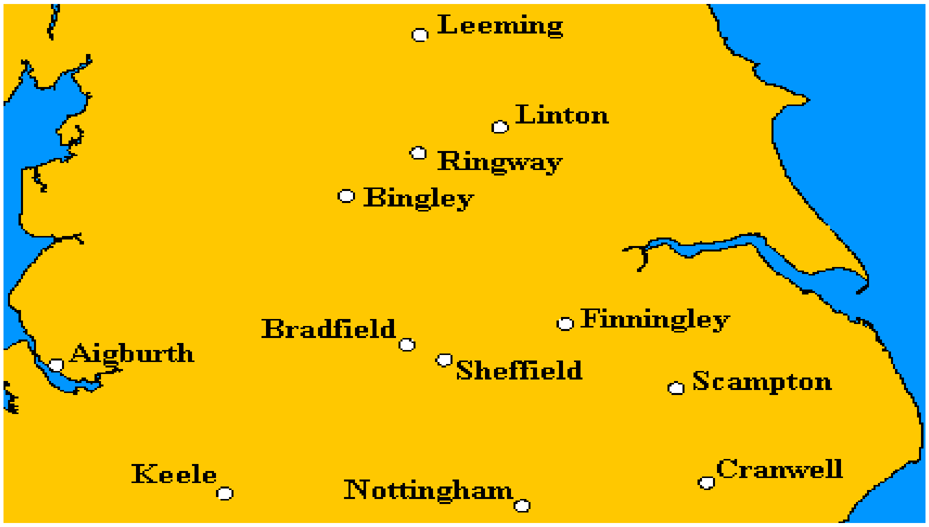

Figure 21: Location of wind speed stations.

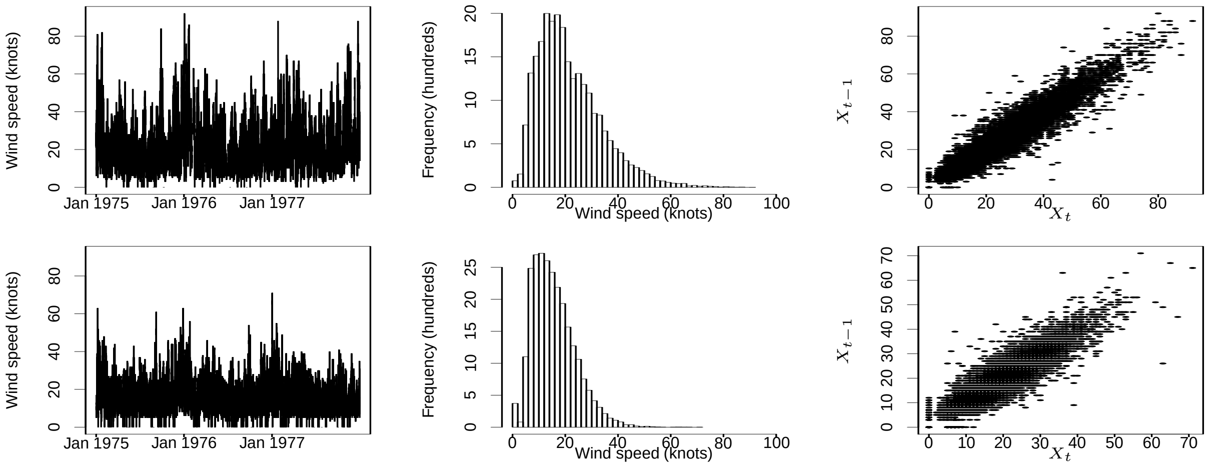

Figure 22: Time series plots and histograms of hourly gusts observed at Bradfield (top row) and Nottingham (bottom row) over a three year period (1975–1977 inclusive). Also shown are plots of the time series against the lagged series.

Site and seasonal variability

For our purposes, we need the GPD parameters to vary across sites, and seasonally. We take a pragmatic approach to seasonality, partitioning the annual cycle into twelve 'months'. Thus our hierarchical model will need to yield parameter pairs $(\sigma_{m,j}, \xi_{m,j})$ for $m = 1, \ldots, 12$ and $j = 1, \ldots, 12$, where m and j are indices of season and site respectively. It is also necessary to allow the threshold u used for excesses modelled by the GPD to vary, since different criteria about what constitutes an extreme value will be in play for each combination of season and site. We will denote by $u_{m,j}$ the value of the exceedance threshold for month m and site j.

Temporal dependence

To account for the presence of temporal dependence within each season and site, we now adopt approach 3 outlined in Section 2.1.3; specifically, we use bivariate extreme value theory discussed in Part 4 of this short course to formulate a simple first–order Markov chain structure for successive extreme wind speeds. As with fitting to all threshold exceedances and then adjusting the inferences accordingly (as we recommended with the Newlyn sea–surge data in Section 2.1.4), this approach avoids the need to arbitrarily identify clusters of extremes and filter out a set of independent extreme values (thus discarding many precious extremes!), but also quantifies the extent of extremal dependence at each site. Put simply, at each site, the logistic model with parameter α_j (discussed in Part 4) is used to model each successive pair of threshold exceedances (say (x_i, x_{i+1})) at site j. The parameter $\alpha_j \in (0, 1]$ measures the strength of dependence between consecutive extremes, smaller values indicating stronger dependence. Independence and complete dependence are obtained when $\alpha_j = 1$ and $\alpha_j \searrow 0$ respectively. Following work in Fawcett (2005), which suggests that the serial dependence in extremes is fairly constant across all seasons, we assume that the Markov chain model describes the dependence over all seasons at site j.

Threshold stability property

In order to ensure a threshold stability property in our models, we use $\tilde{\sigma}_{m,j} = \sigma_{m,j} - \xi_{m,j} u_{m,j}$ in place of the usual scale parameter $\sigma_{m,j}$. With this parameterisation, if $(X - u^*_{m,j})$ is distributed GPD($\tilde{\sigma}_{m,j}, \xi_{m,j}$), then for all values $u_{m,j} > u^*_{m,j}$, we have that $(X - u_{m,j})$ is also GPD($\tilde{\sigma}_{m,j}, \xi_{m,j}$) distributed (e.g. see Coles (2001)). This is useful here, because it allows comparisons of the GPD scale and shape parameters across seasons and sites. It also allows us to specify prior information for both parameters without having to worry about the additional complications that would arise for parameters which were threshold dependent.

The model

We then specify the following random effects model for our extreme wind speeds:

$$\begin{aligned} \log(\tilde{\sigma}_{m,j}) &= \gamma^{(m)}_{\tilde{\sigma}} + \epsilon^{(j)}_{\tilde{\sigma}}, \\ \xi_{m,j} &= \gamma^{(m)}_{\xi} + \epsilon^{(j)}_{\xi} \qquad \text{and} \\ \alpha_j &= \epsilon^{(j)}_{\alpha}, \end{aligned}$$

where, generically, 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 random effects for $\log(\tilde{\sigma}_{m,j})$ and $\xi_{m,j}$ are taken to be normally and independently distributed:

$$\gamma^{(m)}_{\tilde{\sigma}} \sim N_0(0, \tau_{\tilde{\sigma}}) \qquad \text{and} \tag{19}$$

$$\gamma^{(m)}_{\xi} \sim N_0(0, \tau_{\xi}), \qquad m = 1, \ldots, 12, \tag{20}$$

for the seasonal effects, and

$$\begin{aligned} \epsilon^{(j)}_{\tilde{\sigma}} &\sim N_0(a_{\tilde{\sigma}}, \zeta_{\tilde{\sigma}}) \qquad \text{and} \\ \epsilon^{(j)}_{\xi} &\sim N_0(a_{\xi}, \zeta_{\xi}), \qquad j = 1, \ldots, 12, \end{aligned}$$

for the site effects, where $N_0(\eta, \rho)$ is the normal distribution with mean η and *precision* ρ (used for notational convenience). We choose the mean of the normal distribution of the seasonal effects to be fixed at zero in (19) and (20) in order to avoid over–parameterisation and problems of model identifiability, although we could equally well have fixed the mean for the distribution of the *site* effects to achieve this. In the absence of any prior knowledge about α_j, we set the prior by specifying

$$\epsilon^{(j)}_{\alpha} \sim U(0, 1).$$

The final layer of the model is the specification of prior distributions for the random effect distribution parameters. Here we adopt conjugacy wherever possible to simplify computations, specifying:

$$\begin{aligned} a_{\tilde{\sigma}} &\sim N_0(b_{\tilde{\sigma}}, c_{\tilde{\sigma}}), \qquad & a_{\xi} &\sim N_0(b_{\xi}, c_{\xi}); \\ \tau_{\tilde{\sigma}} &\sim Ga(d_{\tilde{\sigma}}, e_{\tilde{\sigma}}), \qquad & \tau_{\xi} &\sim Ga(d_{\xi}, e_{\xi}); \\ \zeta_{\tilde{\sigma}} &\sim Ga(f_{\tilde{\sigma}}, g_{\tilde{\sigma}}), \qquad & \zeta_{\xi} &\sim Ga(f_{\xi}, g_{\xi}); \end{aligned}$$

subject to the choice of arguments for these functions, i.e. the hyper–parameters which determine the precise Normal and Gamma distributions.

MCMC algorithm

We use a hybrid scheme (see Section 5.2.3) – specifically 'Metropolis with Gibbs' – to sample form the posteriors. This means we update each component singly using a Gibbs sampler where the conjugacy allows straightforward sampling from the full conditionals, and a Metropolis step elsewhere.

Some results

Some results are shown in Figures 23–26 and in Table 5. The main points to notice are listed below:

— Advantage of the hierarchical model over a standard likelihood–based analysis: a reduction in sampling variation (posterior standard deviations in the bottom portion of Table 5 are substantially smaller than the corresponding standard errors) due to the pooling of information across sites and seasons

— Figure 25 further highlights this reduction in variability – notice the *shrinkage* in estimates of the GPD shape parameter ξ in the Bayesian analysis relative to the standard likelihood–based analysis

— Separate seasonal parameters are recombined for each site to obtain site–by–site estimates of return levels (see Figure 25, bottom right); notice that estimates of extreme quantiles using maximum likelihood estimation can be very unstable, whereas the hierarchical model achieves a greater degree of stability through the pooling of information across sites

— Figure 26 shows an extension to *predictive return levels*, which cannot be achieved under the classical approach to inference

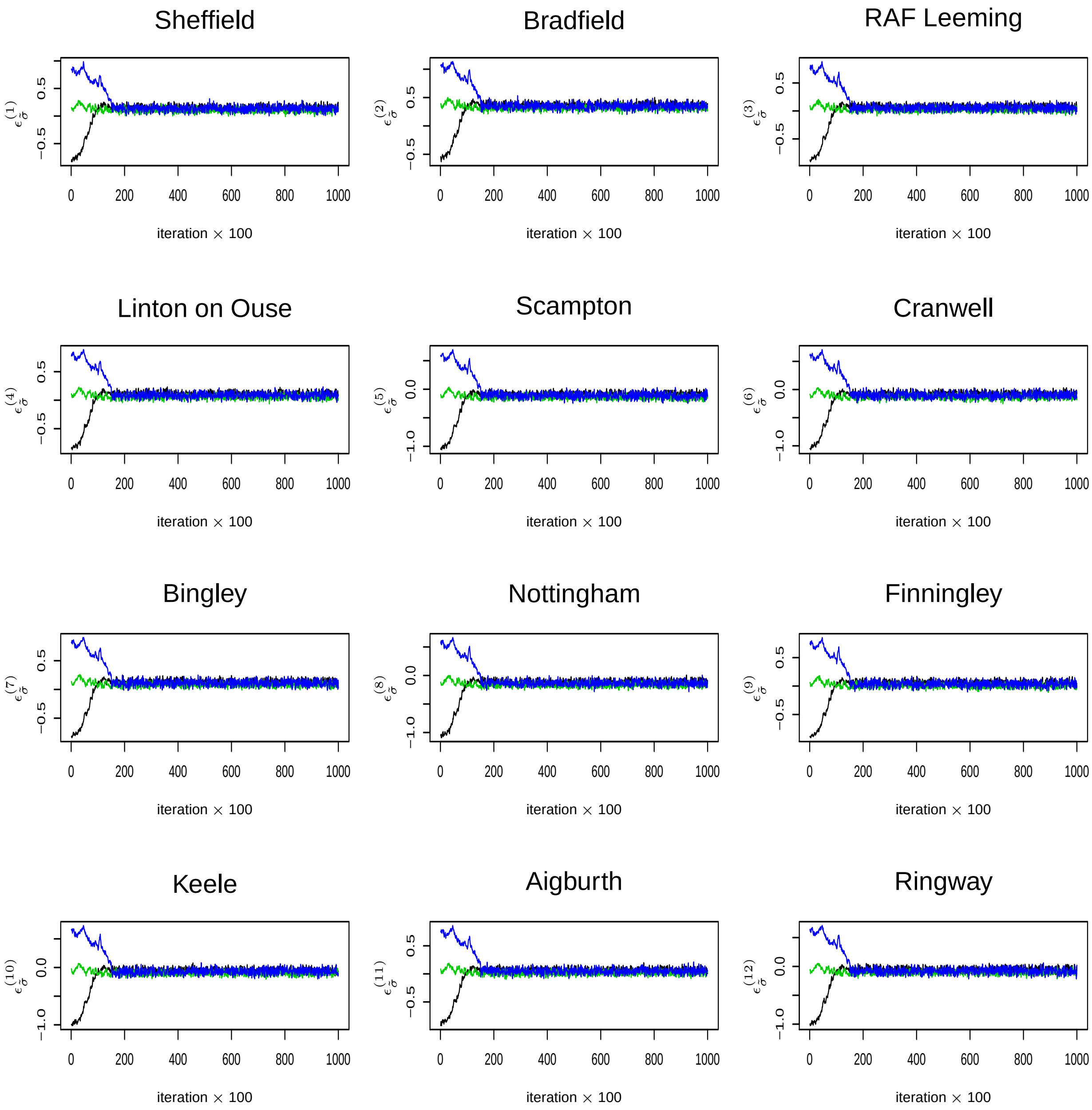

Figure 23: Trace plots of the site effects for $\log(\tilde{\sigma})$ for each site in the study

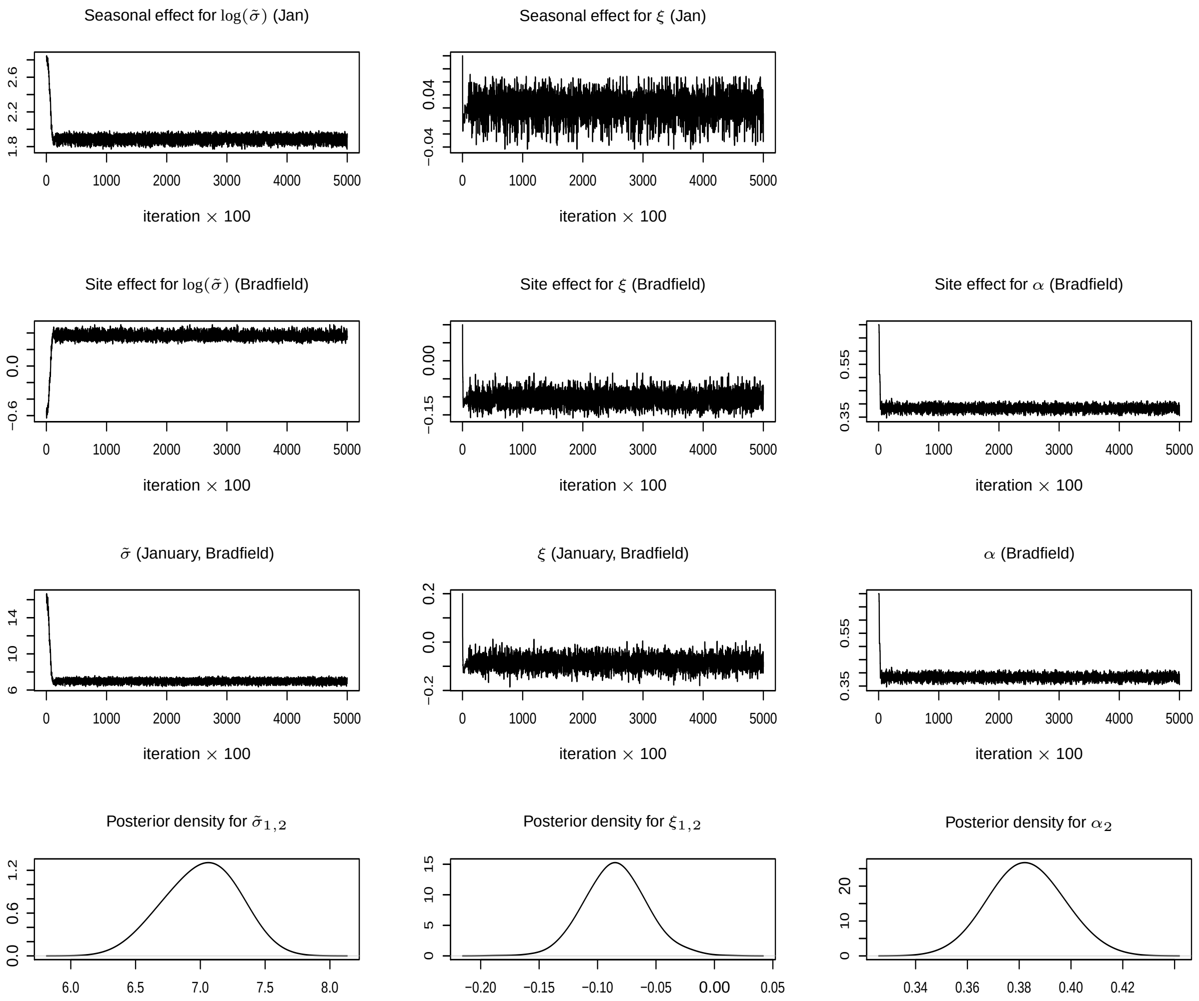

Figure 24: MCMC output for Bradfield in January

	Bradfield, January	**Nottingham, July**
	Mean (st. dev.) *MLE (asymp. s.e.)*	Mean (st. dev.) *MLE (asymp. 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 5: Bayesian random effects analysis of extreme wind speeds – Bradfield (January) and Nottingham (July)

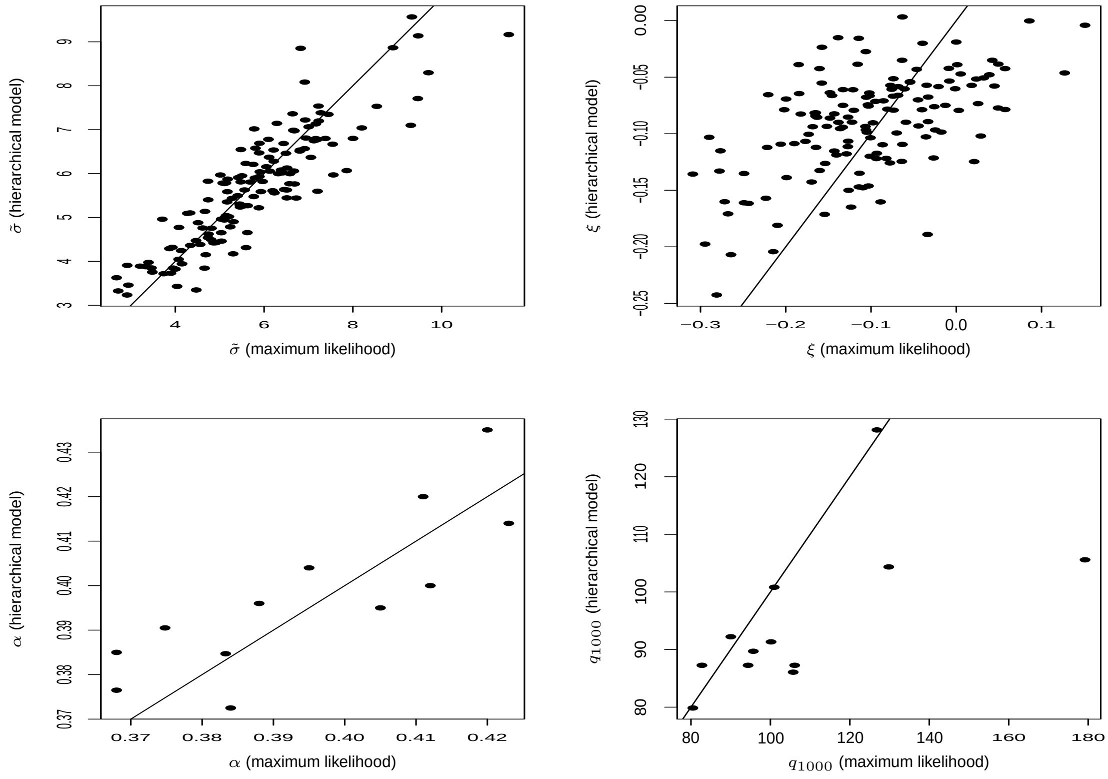

Figure 25: Posterior means against maximum likelihood estimates of GPD parameters, logistic dependence parameter and 1000–year return level

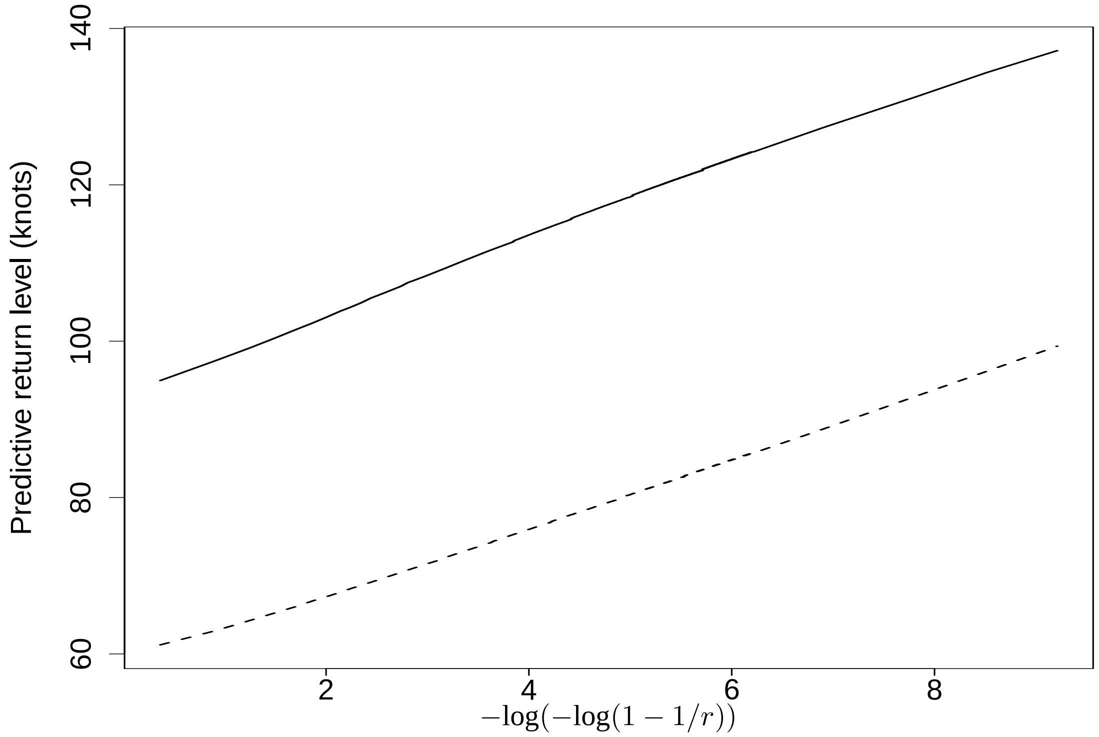

Figure 26: Predictive return level curves for Bradfield (—) and Nottingham (- - - -).

6 R session: Multivariate extremes and Bayesian inference

As before, you will need to start R, and then attach the libraries `ismev` and `evd`, and the supplementary R routines we have provided, using the commands:

```
> library(ismev)
> library(evd)
> source('Rstufflee.r')
```

1. In this question we carry out a simple bivariate analysis using the block–maxima approach. The dataset `wind` has 40 rows and 3 columns; the second and third columns contain annual maximum wind speeds at Albany, New York and Hartford, Connecticut (respectively) over the period 1944 to 1983.

(a) Load the data into `R` using:

```
> data(wind)
```

and have a look at it by typing

```
> wind
```

(b) The data we want are the annual maxima for Hartford and Albany respectively, stored in columns 2 and 3. We extract them using

```
> hartford<-wind[,2]
> albany<-wind[,3]
```

and then recombine them into a vector of bivariate annual maxima using

```
> blockmax<-cbind(hartford,albany)
```

(c) We can now fit a bivariate extreme value distribution using the logistic model:

```
> fbvevd(blockmax)
```

since the logistic model is the default. You may like to experiment with other models, e.g.

```
> fbvevd(blockmax,model="bilog",std.err = FALSE)
```

although note that this model is too complex to calculate standard errors, hence the need to switch this facility off (to avoid an error!). You may like to experiment with other models.

(d) If we want to produce diagnostic plots we must first create an object containing the fits, e.g.

```
> fit1<-fbvevd(blockmax)
```

and then run the plots command

```
> plot.bvevd(fit1)
```

You may like to think about what these plots are telling us, and investigate how well different models fit these data.

2. The data set `wavesurge` contains the data on which the bivariate example in Section 4.3 was based. The data has 2894 rows and 2 columns; corresponding to the wave height and sea surge in consecutive measurements taken at Newlyn, Cornwall, between 1971 and 1977.

(a) Load the data into `R` using:

```
> data(wavesurge)
```

Now separate wave and surge using:

```
> wave<-wavesurge[,1]
> surge<-wavesurge[,2]
```

(b) You can check this has worked by plotting surge against wave height using:

```
> plot(wave,surge)
```

At this stage it would be possible to carry out univariate threshold–based analyses of each of `wave` and `surge` separately, and you may like to do this in your own time. However we will proceed directly to a bivariate analysis in the exercises below.

(c) We will first identify appropriate thresholds for the analysis. We decide to identify the empirical 95% quantile in each margin, and we can do this using:

```
> quantile(wave,0.95)
> quantile(surge,0.95)
```

We can now create an appropriate bivariate threshold vector, e.g. using:

```
> thresh<-c(6.080,0.322)
```

(d) We are now in a position to fit various bivariate models to the bivariate object `wavesurge`:

```
> fbvpot(wavesurge,thresh)
```

fits the logistic model (the default). Check you understand all of the output, including identifying the relevant model parameters.

(e) To fit the bilogistic model, use:

```
> fbvpot(wavesurge,thresh,model=”bilog”)
```

You may like to experiment with other models.

(f) For any particular model fit, we can explore the model fit, and various aspects of the inference, using the graphical routine `plot.bvpot()` applied to an object generated from a fit. For example, to investigate the fitted logistic model, use:

```
> fitlogistic<-fbvpot(wavesurge,thresh)
> plot.bvpot(fitlogistic)
```

3. We return to the dataset `wind` containing annual maximum wind speeds at Albany, New York and Hartford, Connecticut over the period 1944 to 1983. The first column gives corresponding years. The data set should already be in R, but if you have not done Question 1 in this Rsession, reload it using:

```
> data(wind)
```

Now separate the two sets of wind speeds using

`> albany<-wind[,2]` and
`> hartford<-wind[,3]`
The function

```
> gev.bayes(n,dataset,mustart,sigmastart,xistart, ...
... errmu,errlogsigma,errxi,sdmu,sdlogsigma,sdxi)
```

produces (approximate) draws from the posterior distribution $\pi(\mu, \sigma, \xi|\underline{y})$, where μ, σ and ξ are the location, scale and shape parameters of the GEV distribution and $\underline{y} = (y_1, y_2, \ldots, y_{40})$ are the annual wind speed maxima in years 1944, 1945, . . ., 1983. This routine uses Metropolis–Hastings sampling with a random walk update scheme for each of the parameters. As in the notes, independent Normal priors are used for μ, log(σ) and ξ.

The arguments in the function are defined as follows:

`n`	The number of iterations in the Metropolis–Hastings sampler
`dataset`	A single vector containing the data
`mustart`	The starting value for μ in the chain
`sigmastart`	The starting value for σ in the chain
`xistart`	The starting value for ξ in the chain
`errmu`	The random walk innovation variance for μ
`errlogsigma`	The random walk innovation variance for log(σ)
`errxi`	The random walk innovation variance for ξ
`sdmu`	The Normal distribution prior standard deviation for μ
`sdlogsigma`	The Normal distribution prior standard deviation for log(σ)
`sdxi`	The Normal distribution prior standard deviation for ξ

(a) Run the Metropolis–Hastings sampler for the wind speed maxima observed at Albany, NY, for 10,000 iterations, using

- $(\mu^{(0)}, \sigma^{(0)}, \xi^{(0)}) = (20, 15, 0.1)$;
- $v_\mu = v_{\log(\sigma)} = v_\xi = 0.1$;
- Large Normal prior standard deviations for μ, log(σ) and ξ – 10000, 10000, 100 (respectively).

Make sure you store your results somewhere, e.g. use

```
> mcmc.results1<-gev.bayes( ...
```

and ignore the `warning` message that R returns. Then `mcmc.results1` will store the 10,000 draws from the posteriors of μ, log(σ) and ξ, as well as the corresponding acceptance probabilities – these can be accessed by typing, for example,

```
> mcmc.results1$mu
```

(b) Now examine your output using

```
> par(mfrow=c(3,1))
> plot(ts(mcmc.results1$mu))
> plot(ts(mcmc.results1$logsigma))
> plot(ts(mcmc.results1$xi))
```

(You may want to edit the labels for the axes as we did in Part 3 of this course, using, for example, `xlab='iteration'`.) Do you think your sampler is performing well? Does it converge? If so, what is the 'burn–in' period?

(c) Remember, an overall acceptance probability for each parameter of between 30%–50% is usually good enough. Look at your acceptance probabilities for μ, log(σ) and ξ by typing, for example

```
> mean(mcmc.results1$aprobmu)
```

Do you think your sampler is performing well?

(d) Now run the sampler again (maybe store your results in `mcmc.results2`) but choose more appropriate starting values based on your plots in part (b) and change the variances of your random walk innovations if necessary (if you increase `errmu`, `errlogsigma` or `errxi` the corresponding acceptance probabilities will decrease). Examine your output as you did in parts (b) and (c) and check for improvement.

(e) Once you are satisfied with your MCMC, you should summarise your posteriors (after the removal of burn–in). Typing

```
> mu.burn<-mcmc.results2$mu[2000:10000]
```

would, for example, discard the first 2000 iterations as 'burn–in' and store the remainder of the posterior draws for μ in the vector `mu.burn`. After identifying an appropriate burn–in period for *your* MCMC output, use commands similar to that above to obtain vectors containing posterior draws for μ, log(σ) and ξ after the removal of burn–in (and store them in `mu.burn`, `logsigma.burn` and `xi.burn`).

(f) We will now look at the posterior densities of our MCMC draws for μ, σ and ξ. Type

```
> par(mfrow=c(2,2))
> plot(density(mu.burn))
> plot(density(exp(logsigma.burn)))
> plot(density(xi.burn))
```

to produce density plots of the posterior draws for the parameters μ, σ and ξ (note the transformation back to σ by exponentiation of the log(σ) vector).

(g) Find the posterior mean and standard deviation for each of the three GEV parameters by typing, for example,

`> mean(mu.burn)` and
`> sd(mu.burn)`

(h) Now we can obtain the posterior distribution for, say, the 1000–year return level by using the function `ret.level.gev` on each of the draws for μ, σ and ξ. We can do this by typing:

```
> retlevel<-vector('numeric', length(mu.burn))
> for(i in 1:length(retlevel))
+ {
+ retlevel[i]<-ret.level.gev(mu.burn[i],...
...exp(logsigma.burn[i]),xi.burn[i],1000)
+ }
```

Now typing

```
plot(density(retlevel))
```

will add a density plot of the posterior for the 1000–year return level to your panel of plots produced in part (f). Numerical summaries can be obtained in a similar fashion to (g), though owing to the (often) severe asymmetry of the posterior surface for return levels, you may want to use `median()` and not `mean()` as a summary of posterior location here.

(i) Now find maximum likelihood estimates for μ, σ, ξ and the 1000–year return level (see Part 3) and compare these with the results from your Bayesian analysis (compare m.l.e.s with posterior means, for example, and estimated standard errors with posterior standard deviations).

(j) If you have time, and are interested in this stuff, you could re–run this type of analysis on the Hartford data.

Acknowledgement

Lee Fawcett and Dave Walshaw would like to thank Jill Johnson for her help in producing some of the `R` routines and plots used in this short course.

References

Bortot, P., Coles, S. and Tawn, J. (2000). The Multivariate Gaussian tail model: an application to oceanographic data. *Applied Statistics*, **49**, 1, 31—50.

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.

Davison, A.C. and Smith, R.L. (1990). Models for Exceedances over High Thresholds (with discussion). *J. R. Statist. Soc., B*, **52**, 393—442.

Fawcett, L. (2005). Statistical Methodology for the Estimation of Environmental Extremes. PhD Thesis, University of Newcastle-upon-Tyne.

Fawcett, L. and Walshaw, D. (2006). A Hierarchical Model for Extreme Wind Speeds. *Applied Statistics*, **55**, 5, 631—646.

Fawcett, L. and Walshaw, D. (2007). Improved Estimation for Temporally Clustered Extremes. *Environmetrics*, **18**, 2, pp. 173-188

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.

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.

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.

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

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. (1991). Regional Estimation from Spatially Dependent Data. Preprint.

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.