School of Mathematics & Statistics

MMathStat Project

Bayesian Inference for Environmental Extremes

Author:
Katie Connor

Supervisor:
Dr. Lee Fawcett

May 2014

Abstract

Due to the world's changing climate, extreme storms are becoming more likely to occur with increased severity, making research into the application of extreme value theory to environmental data more valuable than ever. In an extreme value analysis data are scarce so a Bayesian framework can adopted to include extra information through informative priors, as well as for the use of the predictive density in order to estimate high quantiles - for example, the sea surge event we would expect to see once every hundred years. This dissertation first compares a maximum likelihood approach with a simplistic Bayesian analysis. Within the Bayesian framework we can then increase the complexity of the analysis by incorporating the built-in variability of the data by using a hierarchical model. This structure is used to identify seasonal and site effects for hourly sea surge data off the southern U.S. coast in the Gulf of Mexico. The hierarchical model is used in order to increase precision for inferences made at individual sites and for individual seasons.

Contents

Chapter 1

Background and Motivation

1.1 Introduction

Extreme value theory is an area of statistics used for the analysis of the largest (or smallest) values occurring in a stochastic process. This field of research has become increasingly popular over the last 20 years with the theory being applied in a range of different fields. However, a relatively new field of research is the use of the Bayesian framework in an extreme value analysis. Bayesian inference allows for the inclusion of extra information through the use of prior distributions which is helpful in an extreme analysis as extreme data are rare. (See Coles and Powell [1996] for a review of Bayesian techniques in an extreme value analysis). As we shall see in this project, a Bayesian analysis of extremes can result in increased precision of estimates of key quantities of interest.

Extreme value theory has many practical applications, most notably in fields such as

- environmental extremes;
- financial risk, eg stock market risk;
- structural engineering, and
- traffic predictions.

In this project we will be focussing on environmental extremes, in particular sea surges. The data we have used are heights of sea surges along the southern US coast in the Gulf of Mexico. We have data available for 7 sites, with the sites shown in figure 1.1. The sites of interest are Rockport, TX, Sabine Pass, TX, Calcasieu Pass, LA, New Canal Station, LA, Shell Beach, LA, Bay Waveland Yacht Club, LA and St Petersburg FL, shown from left to right in figure 1.1. The data were collected over 5 years, (2009-2013) and we have hourly sea heights above the mean high water (MHW) level. MHW is the average of all high water heights taken from each site from 1983 onwards.[*Data recorded by National Oceanic and Atmospheric Administration, http://tidesandcurrents.noaa.gov/stations*]

Figure 1.1: Data collection sites (*Image obtained from Google Maps*)

	Shell Beach	Calcasieu Pass	Sabine Pass	New Canal Station	Bay Waveland Yacht Club	St Petersburg	Rockport
MHW	32.67	28.69	14.98	4.78	4.02	5.35	6.81

Table 1.1: Mean High Water level (feet)

Table 1.1 shows the MHW levels for each site. We aim to estimate how high above the MHW the sea level will rise. As an illustration, figures 1.2 and 1.3 show (monthly) MHW boxplots, and autocorrelations (respectively), to investigate seasonal variability and temporal dependence within our data. For Shell Beach, the month of September contains the most extreme values observed during 2011. The Atlantic hurricane season runs from July-November. To avoid any issues of seasonality we could examine extremes from one month which, based on figure 1.2, would be September as we can see the most extreme values lie above the upper limit of the box plot.

In order to check for dependence within the data we can use partial autocorrelation plots and lagged time series plots to examine the autocorrelation between observations. Figure 1.3(i) shows the autocorrelation plot for Shell Beach. As we can see there is strong autocorrelation between observations, with a lag 1 autocorrelation of $r_1 = 0.977$. Figure 1.3(ii) shows the time series plot of observations for Shell Beach against points at lag 1. As we can see from this plot there is a strong correlation between x_i and x_{i+1}, where x_i is the sea level observation at time i.

A sea surge is the difference between the current tide height and the average tide height. During a hurricane strong winds push against the surface of the ocean causing the sea level to rise. The subsequent rise in sea level is known as the sea surge. During a particularly powerful storm sea surges can be high, sometimes overpowering a sea wall leading to flooding within a town or city. The purpose of an extreme value analysis in this situation might be to estimate how high the sea level will rise during a storm once, say, every 100 years. Thus, analyses might estimate the *100 year return level*. Estimates of such return levels are often

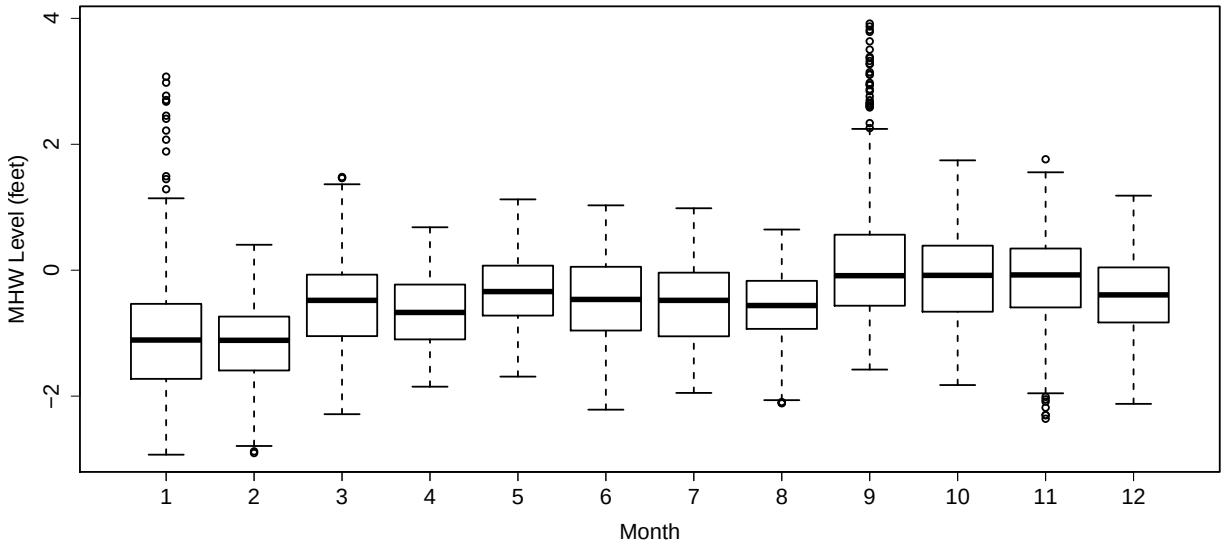

Figure 1.2: Shell Beach monthly MHW level (feet) in 2011

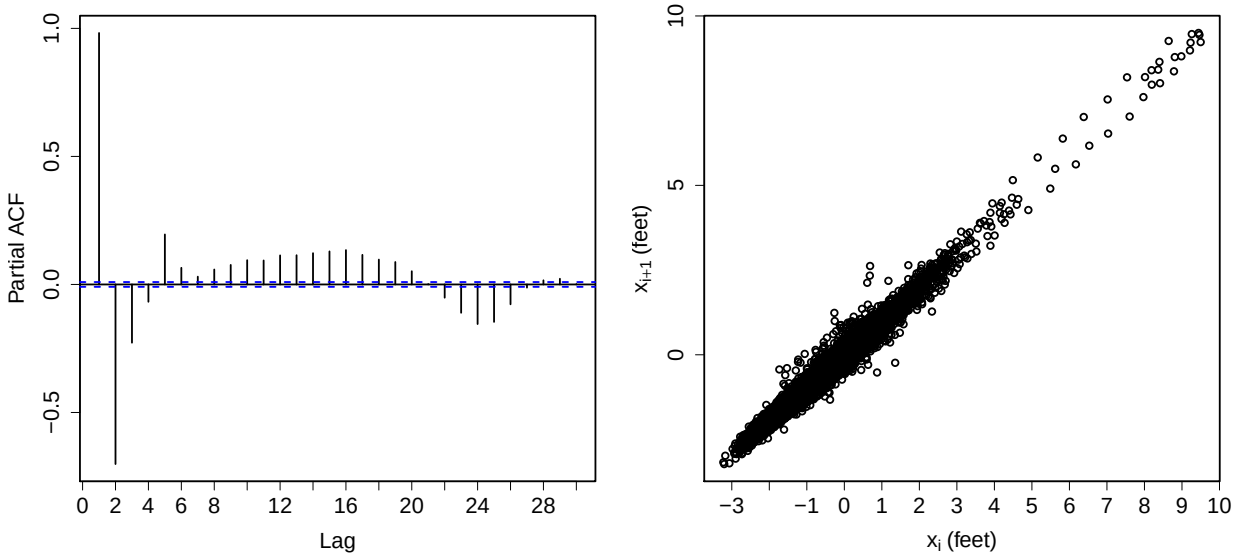

Figure 1.3: i) Autocorrelation plot, ii) Time series against lagged points for MHW at Shell Beach

used as design requirements for structures such as sea walls and other flood defence systems to provide protection up to a given level of safety.

In August 2005 New Orleans was one of the worst affected coastal cities hit by hurricane Katrina, in which over 1800 people lost their lives. The city was catastrophically affected

by the hurricane when the city's flood defences proved ineffective against the category 3 hurricane. The hurricane first formed as a tropical storm over the Bahamas before making landfall in southern Florida as a category 1 hurricane, causing flooding and numerous deaths, even as a category 1 storm. Eventually the hurricane strengthened over the Gulf of Mexico and before making landfall once again in New Orleans, shown in figure 1.4.

The city of New Orleans was built below sea level, leaving the city already vulnerable to flooding. When hurricane Katrina made landfall in Louisiana the levee systems in place throughout New Orleans failed within hours, leading to up to 80% of the city being flooded, with the coastal towns being the worst affected areas. The sea surges caused by the hurricane reached up to an estimated 28 feet above the average sea level causing irreparable damage to many homes and businesses. The failure of the levee system is considered to be one of the worst civil engineering disasters in U.S. history. The flooding caused $81.2 billion worth of damage, making it the costliest hurricane on record in America.

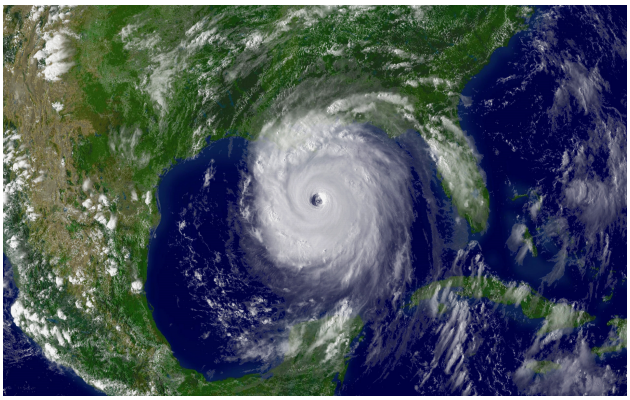

Figure 1.4: Sattelite image of Hurricane Katrina *http://web.mit.edu/12.000/www/m2010/images/katrina-08-28-2005.jpg*

We hope to show, using different techniques for extreme value analyses, that we can estimate the height the sea wall should be built to withstand an extreme storm like Katrina.

There are two main approaches used to analyse data on extremes based on two very different ways for characterising data as extreme:

1. The block maxima approach.
2. The threshold exceedance method.

The block maxima approach involves specifying a block of time or values, such as a year, and extracting the maximum value from each block, M_n, where

$$M_n = \max\{X_1, X_2, \ldots, X_n\}. \tag{1.1}$$

The Extremal Types Theorem provides a limiting distribution for extremes characterised in this way. The theorem was first published by Fisher and Tippett [1928] and later proved by Gnedenko [1943], and is given in Theorem 1.1.

Theorem 1.1 *(Extremal Types Theorem) If there exists sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$Pr\left\{\frac{(M_n - b_n)}{a_n} \leq z\right\} \to G(z) \quad as \quad n \to \infty$$

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

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

for parameters $a > 0, b$ *and, for families II and III,* $\alpha > 0$.

This theorem states that the maxima, M_n, will follow one of the three distributions known as the *extreme value distributions*, which are the *Gumbel*, *Weibull* and *Fréchet* distributions for types I, II and III respectively. In practice, working with three different models is inconvenient; the Generalised Extreme Value distribution unifies the extreme value distributions, giving us a single limiting distribution for our maxima M_n.

1.2 Generalised Extreme Value Distribution

The Generalised Extreme Value (GEV) distribution is defined to be:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \tag{1.2}$$

where $a_+ = max(0, a)$ and $\mu\,(-\infty < \mu < \infty)$, $\sigma > 0$ and $\xi\,(-\infty < \xi < \infty)$ are the location, scale and shape parameters respectively.

The GEV is linked to the three distributions through the shape parameter ξ. When $\xi < 0$ this corresponds to a type II extreme value distribution. When $\xi > 0$ we have a type III extreme value distribution. Equation (1.2) is not defined when $\xi = 0$ so we take the limit as $\xi \to 0$ which gives

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right\}\right]. \tag{1.3}$$

This corresponds to a type I extreme value distribution. The uncertainty in the choice of extreme value distribution (type I, II and III) is now captured by our uncertainty in the

value of the shape parameter ξ.

In order to carry out an extreme value analysis we have to estimate the GEV parameters μ, σ and ξ. To do this there are various methods such as maximum likelihood estimation, method of moments and Bayesian inference. In order to use maximum likelihood estimation we would extract the block (often annual) maxima, and using standard likelihood techniques obtain estimates for the parameters in the usual way - that is, by maximising the likelihood function. If we were to use a Bayesian approach we could use MCMC algorithms to estimate the three parameters of the GEV distribution by using an appropriate summary from the inferred posterior distribution.

1.2.1 Return Levels

Extreme value analyses have some very practical applications. For example, with our data the aim might be to estimate how high a new section of sea wall in New Orleans should be in order to withstand a storm that you would see once in every, say, 100 years. We do this by estimating *return levels*. A return level, z_r, is the value you would expect to see on average once in every r observations. For annual maximum sea surges at a particular site the return level for $r = 100$ would correspond to the sea-surge you would expect to see, on average, once in every r years. In order to estimate the return level we equate the fitted distribution function from equation (1.2), to $1 - {}^1\!/_r$, that is,

$$\exp\left\{-\left[1+\hat{\xi}\left(\frac{\hat{z_r}-\hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}\right\} = 1 - \frac{1}{r} \tag{1.4}$$

where $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ are, typically, the maximum likelihood estimates for μ, σ and ξ respectively. We then rearrange this expression to find $\hat{z}_r$, giving

$$\hat{z}_r = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(-\log\left(1-\frac{1}{r}\right)\right)^{-\hat{\xi}} - 1\right]. \tag{1.5}$$

We can change the value of r to estimate return levels for different periods. In a maximum likelihood analysis, standard errors for $\hat{z}_r$ can be obtained via the delta method, where

$$\mathrm{Var}\left(\hat{z}_r\right) \simeq \nabla z_r^T V \nabla z_r. \tag{1.6}$$

Here V is the variance covariance matrix of $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ obtained on inversion of Fisher's information matrix, and

$$\nabla z_r^T = \left[\frac{\partial z_r}{\partial \mu}, \frac{\partial z_r}{\partial \sigma}, \frac{\partial z_r}{\partial \xi}\right],$$

evaluated at $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$. In a Bayesian analysis we can use posterior summaries such as the posterior standard deviation or posterior credible intervals.

1.2.2 Drawbacks of the Block Maxima Approach

The block maxima approach usually involves taking the maximum value observed each year. However for each of our sites in figure 1.1, this results in a very small sample size - just 5 observations from some 43,000 hourly sea levels. Some of the non-selected observations will still be extreme observations, just not as extreme as the *most* extreme value observed that year. So we can see that this method is particularly wasteful of data. We could reduce the block size to include more block maxima but this then begs the question, 'what is the optimal block size to use?' A block size of one year is often used for convenience, and can avoid issues of seasonal variability and temporal dependence, but if the block size is too large we may have too few extremes to work with. Conversely, if the block size is too small the limiting asymptotic arguments may not hold. In practical terms numerical optimisation procedures may struggle to converge when finding maximum likelihood estimates.

Finally extrapolation with small sample sizes can be difficult. We must interpret the return levels with care as extrapolation into the distant future such as the 100 or 1000 year return periods can be prone to huge uncertainty as we are basing these estimates on only 5 annual maxima. In order to have a reliable extrapolation we require more past data.

One way forward here is to use threshold methods which take all data points over a certain threshold u into account. This approach is generally better as we use more of the data we have collected and therefore avoid some of the issues that arise from using the block maxima approach.

1.3 Generalised Pareto Distribution

Threshold methods provide a more flexible approach for characterising extremes. We determine a threshold value u and use all observations that exceed this threshold in our analysis, as opposed to just a filtered set of yearly maxima. Firstly we need the Distribution of Excesses theorem, and this is given in Theorem 1.2.

Theorem 1.2 *(Distribution of Excesses)*
For large enough u, the distribution function of $(X - u)$*, conditional on* $X > u$*, is approximately:*

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)_+^{-1/\xi} \tag{1.7}$$

defined on $y > 0$*, where*

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{1.8}$$

Equation (1.7) is the distribution function of the Generalised Pareto Distribution (GPD). From section 1.2 we know that the block maxima have a distribution $G(z)$, (see equation

(1.2)), so from Theorem 1.2 we can say the threshold excesses have a corresponding approximate distribution within the Generalised Pareto family. This also implies that the parameters of the GPD are determined by those of the GEV. The shape parameter ξ is the same as that in the GEV and the scale parameter $\tilde{\sigma}$ is a linear combination of the scale and location parameters from the GEV. From this point we will refer to $\tilde{\sigma}$ in the GPD as σ.

Given the advantages over the block maxima approach - mainly the inclusion of more data on extremes, but more generally this being a more 'natural' way of identifying extremes - the threshold approach using the GPD will form the basis of our analysis of sea surges in the Gulf of Mexico throughout this report.

1.3.1 Threshold Selection

Firstly, before fitting the GPD we need to obtain a value for our threshold u. In this method extreme values are defined as those that are greater than the threshold value u. There are two methods used to select u, which are often used in combination. These are:

1. Mean residual life (MRL) plots, and
2. Parameter stability plots

The Mean Residual Life (MRL) Plot

The MRL plot shows the mean excesses over u for different threshold values u. For large values of u, the graph becomes unstable due to the small amount of data available above the threshold. In order to pick a suitable threshold we look to where the graph becomes linear as the function which describes the MRL plot,

$$E(X - u|X > u) = \frac{\sigma_{u_0} + \xi u}{1 - \xi},$$

is linear in u. The LHS of this expression denotes the expectation of a GPD random variable, see Theorem 1.2, and so the point above which this linear relationship holds in our data could indicate the lower bound for identifying extremes to be modelled with the GPD. This technique will be demonstrated in section 1.4.

Parameter Stability Plot

A parameter stability plot is used to choose a suitable threshold value of u by plotting each parameter along with 95% confidence intervals for a range of u. When the estimates begin to converge to a common value we choose this to be the lower bound for u.

1.3.2 Return Levels

In exactly the same way we demonstrated return level estimation via the GEV, see section 1.2.1, we can estimate our return levels using the GPD based on threshold exceedances.

Since we are now working with threshold exceedances we need to incorporate the probability that the threshold has been exceeded. From equation 1.7, we have

$$Pr\left(Z > z \mid Z > u\right) = \left[1 + \xi\left(\frac{z-u}{\sigma}\right)\right]^{-1/\xi}. \tag{1.9}$$

From conditional probability we know that

$$\begin{aligned} Pr(Z > z \mid Z > u) &= \frac{Pr(Z > z \ \cap \ Z > u)}{Pr(Z > u)} \\ &= \frac{Pr(Z > z)}{Pr(Z > u)}. \end{aligned}$$

Rearranging gives

$$\begin{aligned} Pr(Z > z) &= Pr(Z > u)Pr(Z > z \mid Z > u) \\ &= Pr(Z > u)\left[1 + \xi\left(\frac{z-u}{\sigma}\right)\right]^{-1/\xi}. \end{aligned}$$

Setting $\lambda_u = Pr(Z > u)$ as the threshold exceedance rate, gives

$$Pr(Z > z) = \lambda_u\left[1 + \xi\left(\frac{z-u}{\sigma}\right)\right]^{-1/\xi}. \tag{1.10}$$

Now the formula for the return levels for GPD data can be derived in a similar way to the return level formula for the GEV using annual maxima, see equation (1.4). However, this will no longer give return level estimates on a convenient annual scale since we now have more than one observation per year. Thus, replacing r with $n_y \times N$ where n_y = number of observations per year and N = number of years, gives

$$\lambda_u\left[1 + \xi\left(\frac{z-u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{n_y \times N}.$$

Replacing the parameters by the estimated values and solving for $z = \hat{z}_r$ gives

$$\hat{z}_r = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(\hat{\lambda}_u(n_y \times N)\right)^{\hat{\xi}} - 1\right], \tag{1.11}$$

which, because of the inclusion of n_y, is now on the annual scale as before. The threshold exceedance rate λ_u is estimated as the observed proportion of threshold exceedances.

As discussed in the block maxima case, in a likelihood analysis standard errors can be obtained via the delta method, where

$$\text{Var}\left(\hat{z}_r\right) \simeq \nabla z_r^T V \nabla z_r.$$

However, V is now the variance covariance matrix of $\hat{\lambda}_u$, $\hat{\sigma}$ and $\hat{\xi}$, defined by:

$$V = \begin{pmatrix} \hat{\lambda}_u(1-\hat{\lambda}_u)/n & 0 & 0 \\ 0 & \text{var}(\hat{\sigma}) & \text{cov}(\hat{\sigma},\hat{\xi}) \\ 0 & \text{cov}(\hat{\sigma},\hat{\xi}) & \text{var}(\hat{\xi}) \end{pmatrix}. \tag{1.12}$$

Also,

$$\nabla z_r^T = \left[\frac{\partial z_r}{\partial \lambda_u}, \frac{\partial z_r}{\partial \sigma}, \frac{\partial z_r}{\partial \xi}\right],$$

evaluated at $\hat{\lambda}_u$, $\hat{\sigma}$ and $\hat{\xi}$.

1.3.3 Dependence

When working with block maxima data it is reasonable to assume that the annual maxima are independent. However when working with threshold exceedances we will likely have dependent data as we have taken data over a threshold so we may take consecutive data points. For example, if the sea surge is high at 1pm it will still likely be high at 2pm so we probably cannot make the assumption of independence. In order to overcome this issue there are several techniques in place. These include;

- Filtering out a set of independent threshold exceedances;
- Initially ignore dependence and fit the GPD to all threshold exceedances, but then use the extremal index to adjust estimates of return levels.

The first point is the most commonly used method and is known as declustering. The second point relies on the theory of dependent extremes.

Declustering

To decluster we use the following method:

1. Choose a declustering parameter κ.
2. A cluster of threshold exceedances is then terminated when at least κ consecutive observations fall below the threshold u.
3. Identify clusters in this way for the entire series.
4. Extract the maximum of each cluster.
5. Fit the GPD to the set of cluster peak excesses.

Whilst the method is easy to follow there is one main issue - how do we choose a suitable value of κ when declustering? If we choose a value of κ that is too high we risk losing suitable data as the cluster peaks will be too far apart. If κ is too low we will have clusters too close together which may lead to dependent data. If we do not have expert knowledge on the

subject in which we are analysing, for example sea surges, then we may not be able to pick a suitable value. In Fawcett and Walshaw [2012] it was shown that the value of κ can affect the parameter estimates and since it is often the case that κ is chosen randomly or at best by visual inspection. We need a method or expert knowledge to help us choose a sensible value of κ.

Extremal Index

In order to take dependence into account we can use the *extremal index* θ, where $\theta \in (0, 1]$. The extremal index is a measure of extremal dependence (see, for example, Coles [2001]). If

- $\theta \to 0$ the process is increasingly dependent.
- $\theta = 1$ the process is independent.

Theorem 1.3 describes the role of θ in extreme value theory.

Theorem 1.3 *(Extremes of Dependent Sequences)*
Let $X_1, X_2, \ldots, X_n$ be a stationary process and let $X_1^*, X_2^*, \ldots, X_n^*$ be an independent series with X and X^* having the same distributions. As before we define $M_n = max\{X_1, X_2, \ldots, X_n\}$ and $M_n^* = max\{X_1^*, X_2^*, \ldots, X_n^*\}$. Then, under certain regularity conditions (see, for example, Coles [2001]):

$$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G_1(z)$$

as $n \to 0$ where $G_1(z)$ is non-degenerate. It follows that

$$Pr\left\{\frac{M_n^* - b_n}{a_n} \leq z\right\} \to G_1^\theta(z) \tag{1.13}$$

where $0 < \theta \leq 1$.

The above theorem states that if the independent block maxima of a process converge then the maxima of the dependent series will converge to a related distribution. The extremal index θ shows how the two distributions are linked according to the dependence of the series.

Theorem 1.3 is stated in terms of block maxima. Thus, G (if it exists), will necessarily be a generalised extreme value distribution (and, therefore, so will G_1^θ). As discussed earlier, if our set of block maxima can be assumed GEV then, according to Theorem 1.2, the GPD also holds for excesses over some threshold u. Thus, the LHS of equation (1.13) can be replaced with $Pr(X - u|X > u)$, giving a GPD for G_1.

1.3.4 Seasonality

When looking at yearly data we can encounter issues of seasonality, which violates the assumption of independent and identically distributed observations. To avoid these issues we can extract certain months of data - for example, we can look at just one month or look at the hurricane season for each year. This can help to remove any seasonality in the data. A more complex way to overcome issues of seasonality is to model the seasonal effects for each month at each site of interest, see chapter 3 for more details.

1.4 Example: Sea Surges at Shell Beach, LA

We can now use the sea surge data to estimate return levels for locations in the Gulf of Mexico. As a demonstration we will examine sea surges at Shell Beach, looking at just one month of extremes in order to avoid issues of seasonality. Figure 1.2 suggests there is seasonality in the data for Shell Beach and based on this plot we will examine the month of September across a 5 year period. September is in the middle of the Atlantic hurricane season and experience suggests that the most devastating storms often strike in this month (eg Katrina). In order to take dependence into account in this analysis we will use the method of declustering, choosing $\kappa = 10$ as our value to identify clusters of extremes. In this section, we will use maximum likelihood to demonstrate the estimation of parameters. All plots and parameter estimates were obtained via the ISMEV package in R [Coles and Stephenson, 2011].

1.4.1 Threshold Selection

To identify cluster peaks we need to obtain the threshold value u. To do this we can use both methods discussed earlier, an MRL plot and parameter stability plots. (See section 1.3.1).

To start we can use a mean residual life plot in order to decide on a value for the threshold u. As the threshold increases the number of points above the threshold decreases causing the right hand side of the MRL plot to become unstable. We can also parameter estimate plots in order to view the stability of our estimates of σ and ξ as we increase the threshold.

In order to choose a suitable threshold value we can also look at parameter stability plots. As we can see from figure 1.6 the values begin to settle around $u = 1.3$ so we choose this a suitable value for the threshold.

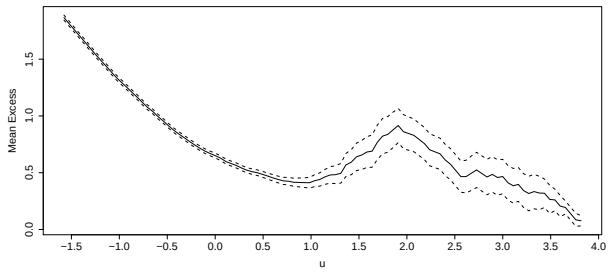

Figure 1.5: Mean Residual Life Plot

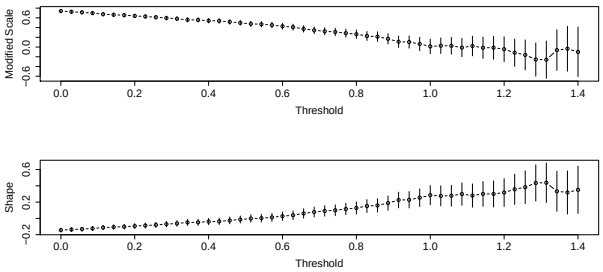

Figure 1.6: Parameter Estimate Plots

1.4.2 Parameter Estimation

Now we can fit the GPD to the cluster peaks above a threshold of 1.3 feet. To obtain the maximum likelihood estimates for the GPD shape and scale parameter σ and ξ as well as the

threshold exceedance rate λ_u we can use R to numerically maximise the GPD log-likelihood function. This gives the following maximum likelihood estimates:

$$\hat{\sigma} = 0.185 \ (0.0656), \qquad \hat{\xi} = 0.559 \ (0.313), \qquad \hat{\lambda}_u = 0.007 \ (0.001), \tag{1.14}$$

with standard errors obtained using the delta method. Now we have standard errors we can also work out 95% confidence interval for the estimates. The maximum likelihood estimators are normally distributed so due to the normality of the estimates we can obtain the confidence intervals using the formula MLE $\pm$ 1.96 $\times$ SE(MLE), giving

$$\hat{\sigma} : (0.056, 0.314), \quad \hat{\xi} : (-0.054, 1.172), \quad \hat{\lambda}_u : (0.004, 0.010).$$

The MLE for ξ suggests a very heavy tail. However the 95% confidence interval passes through zero suggesting a Gumbel type tail, which is not so heavy, might be appropriate.

1.4.3 Model Adequacy

Now we have fitted the GPD we can check that the model provides a good fit to the threshold exceedance data.

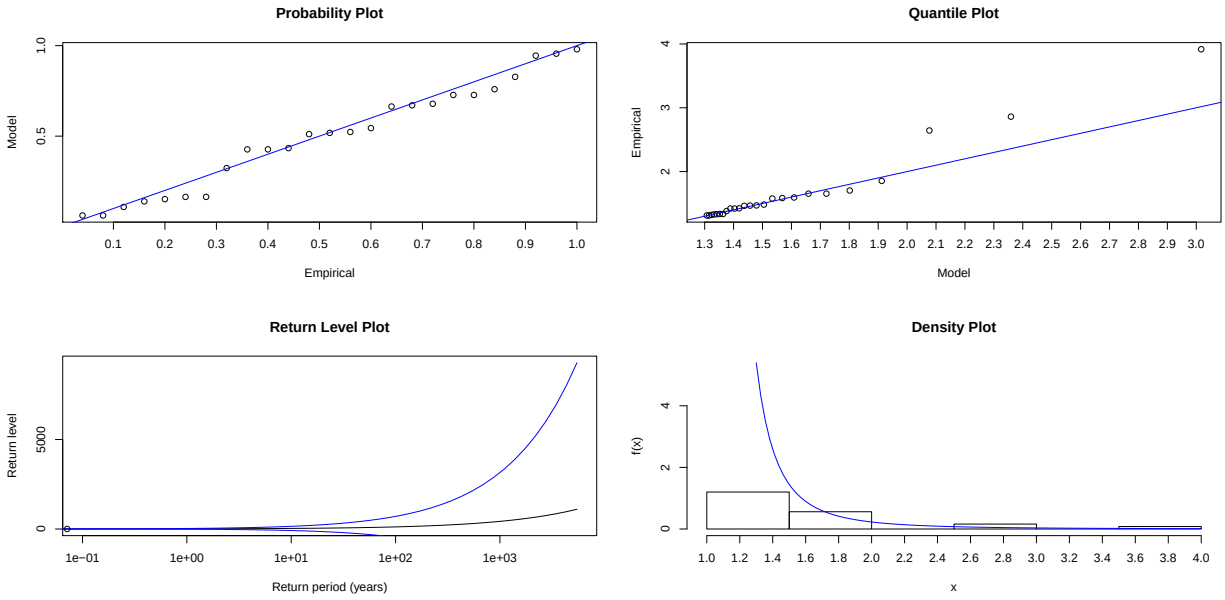

Figure 1.7: Model Adequacy Plots

As we can see from the model adequacy plots shown in figure 1.7, the model is a good fit. Since we have used cluster peaks we do not have many values, hence it is more likely the model will not fit well to a small number of points. The points that are included in this analysis follow the unit line of both the probability and quantile plots with some slight deviation in the tails. The return level plots show the estimated return level falls within the confidence intervals. The plot only contains one point due to the lack of data from the cluster peaks analysis.

1.4.4 Return Levels

Now to estimate our return levels, we can substitute our estimates for σ, ξ and λ_u into equation (1.11). Therefore the return level estimates for September, in feet, are:

$$\hat{z}_{10} = 3.921(1.820),\ \hat{z}_{50} = 8.227(7.589),\ \hat{z}_{100} = 11.662(13.315),\ \hat{z}_{1000} = 39.701(74.998).$$

These values show how high above the MHW level the sea level is estimated to be once in every 10, 50, 100 and 1000 years. The values in brackets are the standard errors associated with each estimate. As we can see as we increase the return period the standard errors increase substantially. For the 1000 year return level we can see that the standard error is much larger than the estimate itself. Such a large standard error offers little precision to our estimates, giving the return level estimate little significance when applied to estimating how high the structure of a sea wall should be built. Also in practice, the likelihood surface encountered for return levels is often severely asymmetric, leading to confidence intervals obtained in the usual way being extremely unreliable. (See, for example, Coles [2001]).

1.4.5 Profile Likelihood

In order to obtain more accurate confidence intervals we can use the profile likelihood. To obtain the profile likelihood we first rearrange the return level equation shown in equation (1.11) in order to make σ the subject of the equation. This gives

$$\sigma = \frac{\xi}{(\lambda_u r)^{\xi} - 1}(z_r - u). \tag{1.15}$$

We now have a formula for σ which is a function of the return levels, z_r and ξ. We can now re-write the log-likelihood formula for the GPD in order to make it a function of z_r and ξ by substituting in σ, equation (1.15). This then gives the formula:

$$\begin{aligned} \ell &= -n\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left(1+\frac{\xi y_i}{\sigma}\right) \\ \implies \ell_p &= -n\log\left(\frac{\xi(z_r-u)}{(\lambda_u r)^{\xi}-1}\right) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left\{1+\frac{y_i[(\lambda_u r)^{\xi}-1]}{(z_r-u)}\right\}. \end{aligned} \tag{1.16}$$

This is the equation for the profile log-likelihood. We now maximise ℓ_p for a range of values for the return level. We can plot these values and based on these plots we can calculate the 95% confidence intervals. These confidence intervals are shown in table 1.2. See Coles [2001] for a more detailed discussion on profile likelihood.

Return Period	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{100}$	$\hat{z}_{1000}$
Confidence Interval	(2.39, 28.35)	(3.2, 272)	(3.6, 734)	(5.1, 20150)

Table 1.2: 95% confidence intervals for return levels (feet)

As we can see from the table and the estimates given above these confidence intervals are positively skewed and hence imply non-normality. Figure 1.8 illustrates the 10-year return level profile likelihood surface, with the confidence intervals shown in red.

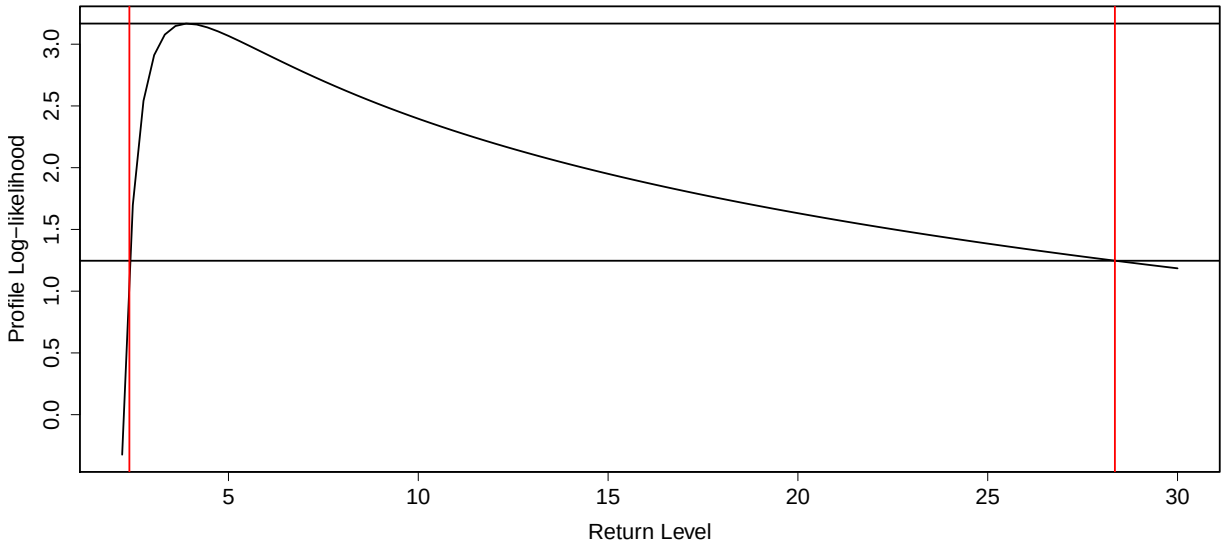

Figure 1.8: 10-year return level profile likelihood

1.5 Downfalls of a Likelihood Approach

Using the maximum likelihood method for estimating parameters and subsequently return levels is usually the most common method however we do encounter some drawbacks.

- When $\xi < -1$ the MLE does not exist.
- When $-1 < \xi < -0.5$ the MLE exists but the usual properties do not hold.
- Return levels are not normally distributed.
- When the sample size is small it is difficult to obtain standard errors.
- When the sample size is small the MLEs for the GPD parameters are biased.

We don't have to worry too much about the first point as it is rare to obtain a value where $\xi < -1$ in extreme value analyses for environmental data.

The third point is one of the more important concerns with maximum likelihood estimation. Since the main aim in extreme value analysis is to obtain return levels we also need to calculate standard errors and from this estimate a confidence interval. If we were to estimate a value for the maximum sea surge in 100 years we would need to present this as a confidence interval as there is uncertainty in our estimation. This will give an engineer an interval to work to in order to estimate the height of sea defences needed to protect a city from such a high sea surge. Since the return levels are not normally distributed a normal 95% confidence interval would be skewed and would not give the desired coverage.

Chapter 2

Bayesian Inference

We will compare Bayesian inference to the MLE method using the GPD. There are numerous advantages to using a Bayesian approach instead of maximum likelihood estimation:

- Informative prior distributions, which can add much-needed precision to an extreme analysis.
- Predictive distribution which accounts for uncertainty in the model.
- Markov Chain Monte Carlo (MCMC) algorithms which will simulate realisations from the posterior distribution.

When using a Bayesian approach we have to specify prior distributions for the parameters in the GPD. In certain situations we could use vague prior distributions if we have no prior knowledge about our data. An extreme analysis also does not include a lot of data as extreme data is rare. Using Bayesian inference we can incorporate expert knowledge through prior distributions so we have more information to use. This gives a better starting point to finding posterior distributions and estimating return levels.

Intrinsic to the Bayesian framework is the predictive distribution. The predictive density has the property that it accounts for uncertainty in the model and uncertainty due to variation in future events. The predictive density is defined by:

$$f(z|\boldsymbol{x}) = \int_{\Theta} f(z|\theta) f(\theta|\boldsymbol{x}) d\theta. \tag{2.1}$$

This allows us to predict return levels using the distribution which is easier than the standard approach using maximum likelihood estimation.

Finally MCMC algorithms have popularised the Bayesian approach to statistical analysis. These algorithms allow us to simulate realisations from the posterior distribution in order to make inferences on the data. If this method is successful we can use the simulated sample to obtain estimates of the posterior distribution, i.e. the sample mean would be a good estimate to the posterior mean and a histogram of the simulated data would be a reasonable estimate of the posterior density. We will never generate *exact* estimates but the larger the

sample size, the more accurate they will become.

Now we can use Bayesian techniques to fit the GPD to our data. To estimate the parameters and estimate return levels we need to estimate the posterior distribution. Bayes theorem states:

$$\begin{aligned} \pi(\theta|x) &= \frac{\pi(\theta)L(x|\theta)}{f(x)} \\ \implies \pi(\theta|x) &\propto \pi(\theta) \times L(x|\theta) \\ \text{i.e. posterior} &\propto \text{prior} \times \text{likelihood} \end{aligned} \tag{2.2}$$

So in order to find the posterior we need to define a suitable prior. Since the GPD does not have a conjugate distribution we cannot obtain the posterior algebraically to obtain the posterior distribution. So we can use numerical techniques to find the posterior, ie MCMC algorithms. There are many different MCMC techniques; most-commonly used are the Gibbs sampler and the Metropolis-Hastings sampler.

2.1 Gibbs Sampler

The Gibbs sampler is an MCMC algorithm that allows us to simulate from a multivariate distribution based on the fact that we can simulate easily from the the full conditional distributions (FCDs).

Suppose we want to generate realisations from the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and that we can simulate from the full conditional distributions.

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \boldsymbol{x}) = \pi(\theta_i|\cdot), \qquad i = 1, \ldots, p.$$

The Gibbs sampler uses the following algorithm:

1. Initialise the iteration counter to $j = 1$. Initialise the state of the chain to $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})^T$.
2. Obtain a new value $\theta^{(j)}$ from $\theta^{(j-1)}$ by successive generation of values
$$\begin{aligned} \theta_1^j &\sim \pi(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\ \theta_2^j &\sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}) \\ &\vdots \qquad \vdots \\ \theta_p^j &\sim \pi(\theta_k|\theta_1^{(j)}, \ldots, \theta_{p-1}^{(j)}) \end{aligned}$$
3. Change counter j to $j + 1$ and return to step 2.

In this algorithm each simulated value depends upon the previously simulated value. This algorithm can be used when we have full conditional distributions that can easily be simulated from. However in practice we find this is not always the case. For example the GPD does not have full conditional distributions that we can use to easily simulate from. Therefore we need to use another MCMC algorithm, namely Metropolis-Hastings. We can also combine the Gibbs sampler with Metropolis-Hastings to make simulations easier.

2.2 Metropolis-Hastings

The Metropolis-Hastings MCMC algorithm, (see Hastings [1970]), works to simulate a sequence of values $\theta_1, \theta_2, \ldots$ from a proposal distribution $q(\boldsymbol{\theta^*}|\boldsymbol{\theta})$ which is easy to simulate from, unlike the FCDs. The algorithm follows the steps:

1. Initialise the iteration counter to $j = 1$. Initialise the chain to $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$
2. Generate a proposed value $\boldsymbol{\theta}^*$ using a proposal distribution with density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})$.
3. Evaluate the acceptance probability of the proposed move $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$ where
$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x})\ q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}|\boldsymbol{x})\ q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\}$$
4. Set $\boldsymbol{\theta}^j = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$. Otherwise set $\boldsymbol{\theta}^j = \boldsymbol{\theta}^{j-1}$
5. Change counter from j to $j + 1$ and return to step 2.

At each step a new value is generated from the proposal distribution. This value is either accepted or rejected so we either move on or stay where we are in the chain until a value is generated that is accepted. Whether or not we move on in the chain is dependent on the acceptance probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. If the acceptance probability is too high we will accept too many generated values, meaning it will take longer to reach the optimal stationary distribution. If the acceptance probability is too low we won't accept enough values. We can plot the simulations and we should hopefully see the simulations converge to the stationary distribution. The period it takes for the iterations converge is known as the *burn in* period. Once we have convergence we can disregard the burn in period as once the sampler has reached its stationary distribution we will then be simulating realisations of the posterior, which is what we are interested in. From here we can then plot density graphs to show the marginal posterior densities for each parameter.

We can also check the partial autocorrelations. For a good model we would want low autocorrelations however sometimes we may generate a high level of dependence so we may have high autocorrelations. In order to lower the autocorrelations we can carry out a process known as *thinning*. To do this we find the last point that has a high autocorrelation, say point m, and from our original sample we take every mth iteration. This process thins the sample and what we have left will have low autocorrelations.

2.3 Metropolis within Gibbs

This algorithm combines both the Gibbs sampler and Metropolis-Hastings algorithm in order to create a hybrid method for simulating from the posterior distribution. This hybrid goes through each full conditional distribution and simulates directly from the full conditionals wherever possible, and carrying out Metropolis-Hastings updates whenever we cannot simulate directly.

2.4 Generalised Pareto Distribution

Using the data for Shell Beach in September, we will use Metropolis-Hastings MCMC algorithms, as the GPD does not have a conjugate distribution, in order to simulate realisations from the posterior. To start our analysis we use the vague priors since we do not have any prior knowledge. It is easiest to use:

$$\log \sigma \sim N(0, 10000), \quad \xi \sim N(0, 1000), \tag{2.3}$$

as these distributions are almost flat distributions as they have such a large variance. To simulate from these distributions we need to set start values. We set our starting values to be:

$$\log \sigma^{(0)} = 2 \quad \xi^{(0)} = -2. \tag{2.4}$$

We can also vary the start value to see how the burn in period is affected. So as well as using (2.4) we can also use the following start values:

$$\log \sigma^{(0)} = 0.1 \quad \xi^{(0)} = 0.1 \tag{2.5}$$

$$\log \sigma^{(0)} = 1 \quad \xi^{(0)} = 3 \tag{2.6}$$

Using these start values we produced the trace plot shown in figure 2.1. We also set the variances of the proposal densities to be approximately 0.7.

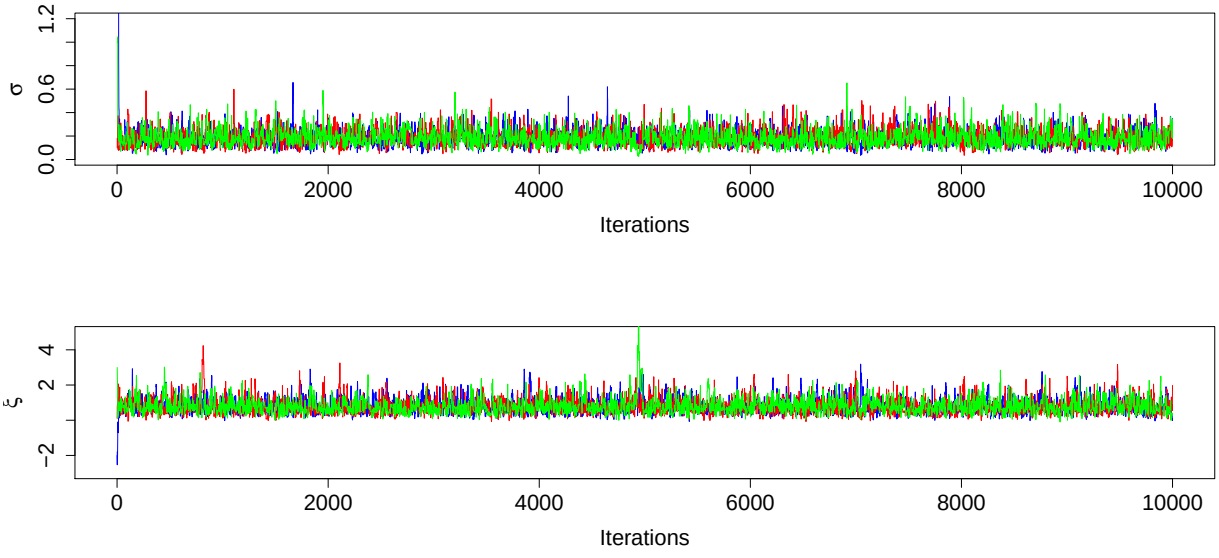

Figure 2.1: Trace Plots

As we can see from figure 2.1 the red line showing start values shown in equation (2.5) has the shortest burn in period, almost starting at the stationary distribution. The start values shown in equation (2.4) are shown by the blue line and the start values in equation (2.6) are

given by the green line. The acceptance probabilities for each parameter were approximately $39 - 44\%$ for each start value for each parameter. We aim to have an acceptance probability between 20-60% in order to move the chain along.

We choose to proceed using our original starting values. We therefore require a burn in period of approximately 300 iterations. We can remove this burn in period and use only observations taken after this period. We now obtain the posterior densities shown in figure 2.2.

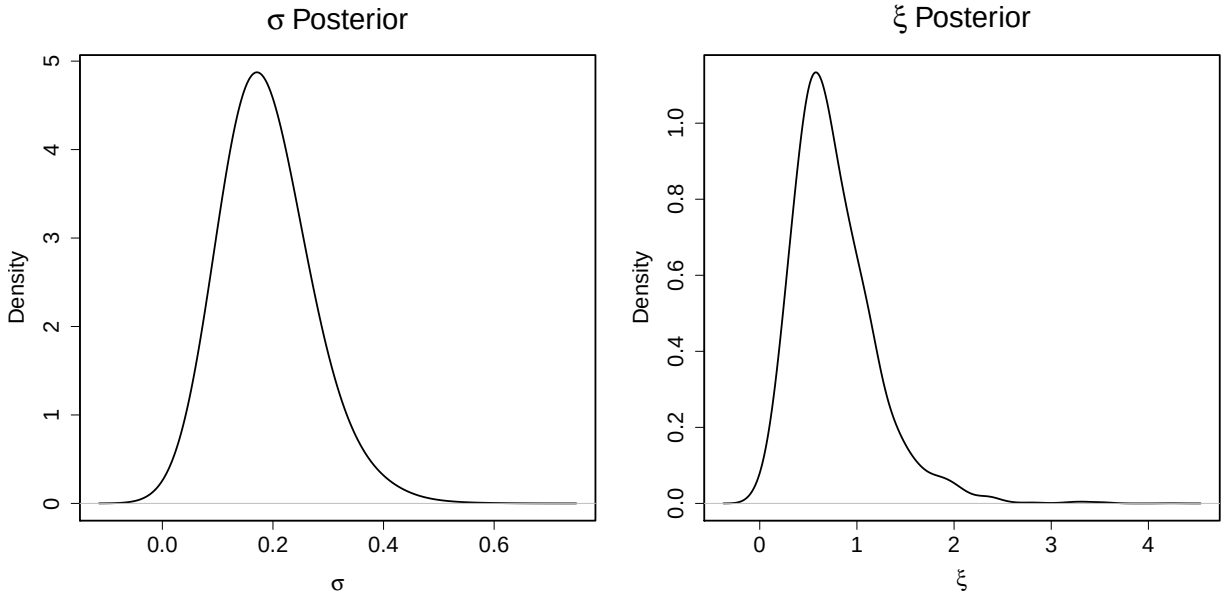

Figure 2.2: Marginal Posterior Densities for σ and ξ

We can see that the posterior plots show a small variance compared to the large vague prior variances we set earlier. The plots are centered around the posterior mean estimates for each parameter which we have worked out to be:

$$\bar{\sigma} = 0.188 \ (0.069) \quad \bar{\xi} = 0.769 \ (0.428), \tag{2.7}$$

where $\bar{\sigma}$ and $\bar{\xi}$ are the posterior mean estimates for σ and ξ. These estimates for σ and ξ are similar to those obtained through maximum likelihood estimation. (See equation (1.14)). The posterior standard deviations are shown in the brackets. These posterior standard deviations are also similar to the standard errors obtained through maximum likelihood. In order to reduce the posterior standard deviations we can include more information through specifying more informative priors.

2.4.1 Return Levels

Finally we can estimate our return levels. These are estimated by substituting each posterior draw for (σ, ξ) into the GPD return level equation shown in equation (1.11) Table 2.1, shows

the posterior median and posterior mode estimates for the distributions of the return levels. As we can see from table 2.1, the posterior median values are different from the return levels

Return Period	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{100}$	$\hat{z}_{1000}$
Median	4.992	13.168	20.763	100.368
Mode	3.3	6.4	7.7	12.5

Table 2.1: Posterior median and mode return levels (feet)

calculated using maximum likelihood estimation. The posterior modes are similar to the return level estimates for maximum likelihood. As we have used a cluster peaks analysis, we have only a small number of values extracted from the original data set. Due to the small number of observations used the return levels are postively skewed, explaining why the posterior median is much higher than the posterior mode. The non-normality can also be seen by looking at the 95% credible intervals, which, as we can see, have a very high upper bound. The credible intervals we obtained by taking the values at 2.5% and 97.5% in the density plots. Figure 2.3 also shows the non-normality in the return levels through the return level density plots.

Return Period	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{100}$	$\hat{z}_{1000}$
Credible Interval	(2.571, 56.682)	(3.636, 824.861)	(4.167, 2678.830)	(6.598, 136417)

Table 2.2: 95% Credible intervals for return levels (feet)

Table 2.2 shows wide credible intervals for all return periods, which is due to the fact that we have used cluster peaks to account for dependence. In figure 2.3 we can see how the non-normality displayed in the return levels. The plots show the points with the highest density, and as we can see the mode of the return levels corresponds to the peak of the density plots and the median return level is far into the tail of the distribution which shows that it would not be practical to use a normal 95% confidence interval.

The use of cluster peaks in our analysis has provided us with wide credible intervals, which are not useful in practical situations. It would be preferable to use more data in our analysis, and one way to do this would be to include all threshold excesses for September at Shell Beach. This would provide us with more data; however, we would somehow need to account for dependence with this method. Another method of improvement would be to use all threshold exceedances from all months of the year. This again would provide more data but we would have to account for the seasonality in our data as well as dependence between observations.

One way to overcome the issue of dependence and seasonality is to use a hierarchical model. The complex structure of a hierarchical model allows us to model the seasonal and site effects separately in order to obtain estimates for the GPD parameters. The value of being able to add more information to an extreme analysis is very high due to the rarity of extreme data. In chapter 3 we will see how this model can be used in accordance with our data.

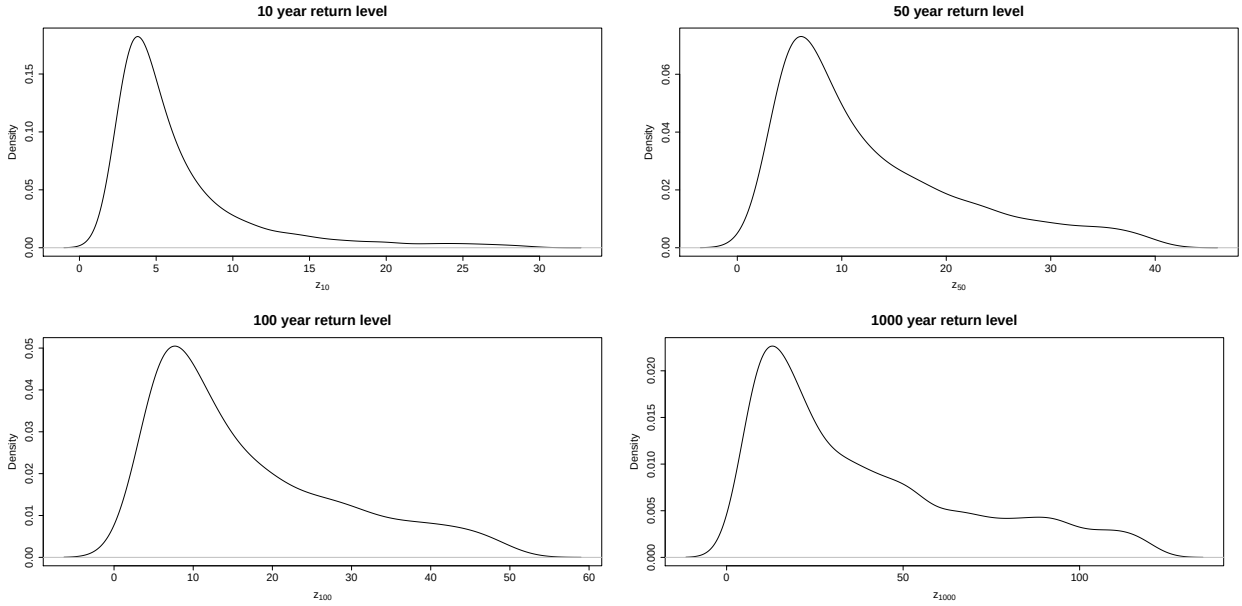

Figure 2.3: Return level density plots (feet)

Chapter 3

Random Effects Model

3.1 Introduction

Another way to model our extreme sea level data is to use a hierarchical model which will account for seasonal and site variation within the data. Random effects models are a useful way of unravelling complex structure in environmental data, such as seasonal and site effects. Using this method it is also much easier to make inferences within the Bayesian framework than within a standard likelihood framework. MCMC algorithms can be used to estimate the parameters more easily than through maximum likelihood estimation due to the complex nature of the likelihood equations.

A random effects model can allow us to use data from all months at each site which enables us to include more information in our inferences. Such a method can give a more accurate account of the seasonal variability and more precise estimates of parameters as we can pool information across sites and seasons.

Another advantage of this analysis is the fact that we can easily incorporate data from surrounding sites rather than just exclusively analyse sea surges from one site at a time, as in chapters 1 and 2. By including data from nearby sites we should be able to see similarities in return levels as we would expect to obtain similar readings from nearby sites. In the analysis of sea surges at Shell Beach in Chapter 1 we only used data from one month at one site. The return levels we estimated will have greater precision if we include data from different sites, and different seasons.

In this chapter, we also account for dependence without wastefully declustering our data by taking the extremal index θ into account. We can estimate the extremal dependence between readings for each site and use this in our return level estimation procedure in order to account for dependence rather than filter out observations to remove dependence as we had previously done in chapters 1 and 2.

To start we will first look at the variation when using seven different sites, shown in figure 1.1. We will then work to model the temporal dependence between extremes at each

site. To take seasonal variation into account we need to partition the data into seasons; for convenience, we take our seasonal unit to be the calendar month. In order to see if the random effects analysis provides better estimates for parameters and return levels we will then compare the outcomes with a typical maximum likelihood analysis, which also accounts for dependence through estimation of the extremal index, but does not account for site and seasonal variation. The method displayed in this chapter follows the method laid out by Fawcett and Walshaw [2006] for a hierarchical model for extreme wind speed data.

3.2 Theory

We now have the following parameters $\tilde{\sigma}_{m,j}$ and $\xi_{m,j}$ for the GPD which we assume to be valid for threshold exceedances in season (month) m and site j, where $m = 1, ..., n_m$ and $j = 1, ..., n_s$, where n_m = number of months and n_j = number of sites:

$$\log \tilde{\sigma}_{m,j} = \gamma_{\tilde{\sigma}}^{(m)} + \epsilon_{\tilde{\sigma}}^{(j)} \tag{3.1}$$

$$\xi_{m,j} = \gamma_{\xi}^{(m)} + \epsilon_{\xi}^{(j)}, \tag{3.2}$$

where γ and ϵ represent seasonal and site effects respectively. Here we use $\tilde{\sigma}_{m,j} = \sigma_{m,j} - \xi_{m,j}u_{m,j}$ in place of $\sigma_{m,j}$. This gives, for all values $u_{m,j} > u^*_{m,j}$, a scale parameter which is threshold independent. We use $\log \tilde{\sigma}$ in order for the scale parameter $\tilde{\sigma}$ to remain positive in our MCMC simulations. We assume a threshold $u_{m,j}$ for each month and site for identifying extremes, and obtain these thresholds using a combination of MRL plots and parameter stability plots, as demonstrated in section 1.4.1.

All of the random effects for $\log \tilde{\sigma}_{m,j}$ and $\xi_{m,j}$ are taken to be normally distributed with mean 0 and precision τ or ζ. The seasonal effects are:

$$\gamma_{\tilde{\sigma}}^{(m)} \sim N(0, \tau_{\tilde{\sigma}}^{-1}), \quad \gamma_{\xi}^{(m)} \sim N(0, \tau_{\xi}^{-1}), \quad m = 1, \ldots, n_m, \tag{3.3}$$

and the site effects are:

$$\epsilon_{\tilde{\sigma}}^{(j)} \sim N(0, \zeta_{\tilde{\sigma}}^{-1}), \quad \epsilon_{\xi}^{(j)} \sim N(0, \zeta_{\xi}^{-1}), \quad j = 1, \ldots, n_s. \tag{3.4}$$

We fix the means to be zero to avoid over parameterisation, although, in principal, this could be generalised to use non-zero means.

The final layer of the model is defined by the random effects distribution parameters. Here we use conjugate distributions to simplify calculations. These distributions are defined by:

$$\tau_{\tilde{\sigma}} \sim Ga(a_{\tilde{\sigma}}, b_{\tilde{\sigma}}), \qquad \tau_{\xi} \sim Ga(a_{\xi}, b_{\xi}), \tag{3.5}$$

$$\zeta_{\tilde{\sigma}} \sim Ga(c_{\tilde{\sigma}}, d_{\tilde{\sigma}}), \qquad \zeta_{\xi} \sim Ga(c_{\xi}, d_{\xi}). \tag{3.6}$$

To start we obtain our full conditional distributions (FCD). For example, for our site effects for $\log \tilde{\sigma}_{m,j}$, we have the Normal pdf from equation (3.4):

$$f(\epsilon_{\tilde{\sigma}}^{(j)}) = \sqrt{\frac{\zeta_{\tilde{\sigma}}}{2\pi}} \exp\left\{-\frac{1}{2}\zeta_{\tilde{\sigma}}(\epsilon_{\tilde{\sigma}}^{(j)})^2\right\}. \tag{3.7}$$

Now we can see that the likelihood is:

$$L(\epsilon_{\tilde{\sigma}}^{(j)}) = \left(\frac{\zeta_{\tilde{\sigma}}}{2\pi}\right)^{n_s/2} \exp\left\{-\frac{1}{2}\zeta_{\tilde{\sigma}} \sum_{j=1}^{n_s} (\epsilon_{\tilde{\sigma}}^{(j)})^2\right\}. \tag{3.8}$$

As previously stated the priors are taken from equations 3.6. Here, the distribution function for the prior of $\zeta_{\tilde{\sigma}}$ is:

$$\pi(\zeta_{\tilde{\sigma}}) \propto \zeta_{\tilde{\sigma}}^{c_{\tilde{\sigma}}-1} \exp\left\{-\zeta_{\tilde{\sigma}} d_{\tilde{\sigma}}\right\}. \tag{3.9}$$

Using Bayes theorem, equation (2.3), we can obtain the posterior distribution;

$$\begin{aligned} \pi(\zeta_{\tilde{\sigma}}|\epsilon_{\tilde{\sigma}}^{(j)}) &\propto \pi(\zeta_{\tilde{\sigma}}) L(\epsilon_{\tilde{\sigma}}^{(j)}) \\ &\propto \zeta_{\tilde{\sigma}}^{c_{\tilde{\sigma}} + {}^{n_s}/_2 - 1} \exp\left\{-\zeta_{\tilde{\sigma}}\left[d_{\tilde{\sigma}} + \frac{1}{2}\sum_{j=1}^{n_s}(\epsilon_{\tilde{\sigma}}^{(j)})^2\right]\right\} \end{aligned}$$

$$\implies \zeta_{\tilde{\sigma}}|\epsilon_{\tilde{\sigma}}^{(j)} \sim Ga\left(c_{\tilde{\sigma}} + \frac{n_s}{2}, d_{\tilde{\sigma}} + \frac{1}{2}\sum_{j=1}^{n_s}(\epsilon_{\tilde{\sigma}}^{(j)})^2\right). \tag{3.10}$$

Similar calculations can be carried out for the site effects for ξ as well as the seasonal effects for $\log\tilde{\sigma}$ and ξ. Therefore our full conditional distributions are gammas. We can work out the remaining three FCDs which gives:

$$\zeta_{\xi}|\epsilon_{\xi}^{(j)} \sim Ga\left(c_{\xi} + \frac{n_s}{2}, d_{\xi} + \frac{1}{2}\sum_{j=1}^{n_s}(\epsilon_{\xi}^{(j)})^2\right), \tag{3.11}$$

$$\tau_{\tilde{\sigma}}|\gamma_{\tilde{\sigma}}^{(m)} \sim Ga\left(a_{\tilde{\sigma}} + \frac{n_m}{2}, b_{\tilde{\sigma}} + \frac{1}{2}\sum_{j=1}^{n_m}(\gamma_{\tilde{\sigma}}^{(m)})^2\right), \quad \textit{and} \tag{3.12}$$

$$\tau_{\xi}|\gamma_{\xi}^{(m)} \sim Ga\left(a_{\xi} + \frac{n_m}{2}, b_{\xi} + \frac{1}{2}\sum_{j=1}^{n_m}(\gamma_{\xi}^{(m)})^2\right). \tag{3.13}$$

For these parameters the MCMC algorithms are simple as we can use Gibbs sampling to simulate directly from the full conditional distributions. In order to simulate from the GPD we have to use Metropolis-Hastings sampling due to the lack of conjugacy of the GPD at this level of the model. In other words we can use a hybrid sampler which uses Metropolis-within-Gibbs. (See section 2.3).

3.3 Analysis

In the first stage of the analysis we will include data from the seven sites shown in figure 1.1. In order to proceed with our analysis we have to set threshold values $u_{m,j}$ in order to obtain our exceedances. We use MRL plots in order to identify appropriate threshold values

for each season m and each site j.

We run an MCMC algorithm to estimate all site and seasonal effects in our model. Each update of the algorithm will include simulating a value from the conditional distributions defined in equations (3.10) – (3.13), and a Metropolis step for the random effects themselves. We choose start values of

$$a_{\cdot} = b_{\cdot} = c_{\cdot} = d_{\cdot} = 1,$$

as we are using vague priors. The MCMC algorithm ran for 30,000 iterations and two different starting points were chosen for each effect parameter to check for convergence. However we have chosen to thin the MCMC output by taking every 10th iteration as this will reduce dependence within the chain. The output produced 38 trace plots showing the convergence for each site and seasonal effect for $\log \tilde{\sigma}$ and ξ. Figure 3.1 shows the trace plots for the seasonal effect for each month of the year for $\log \tilde{\sigma}$.

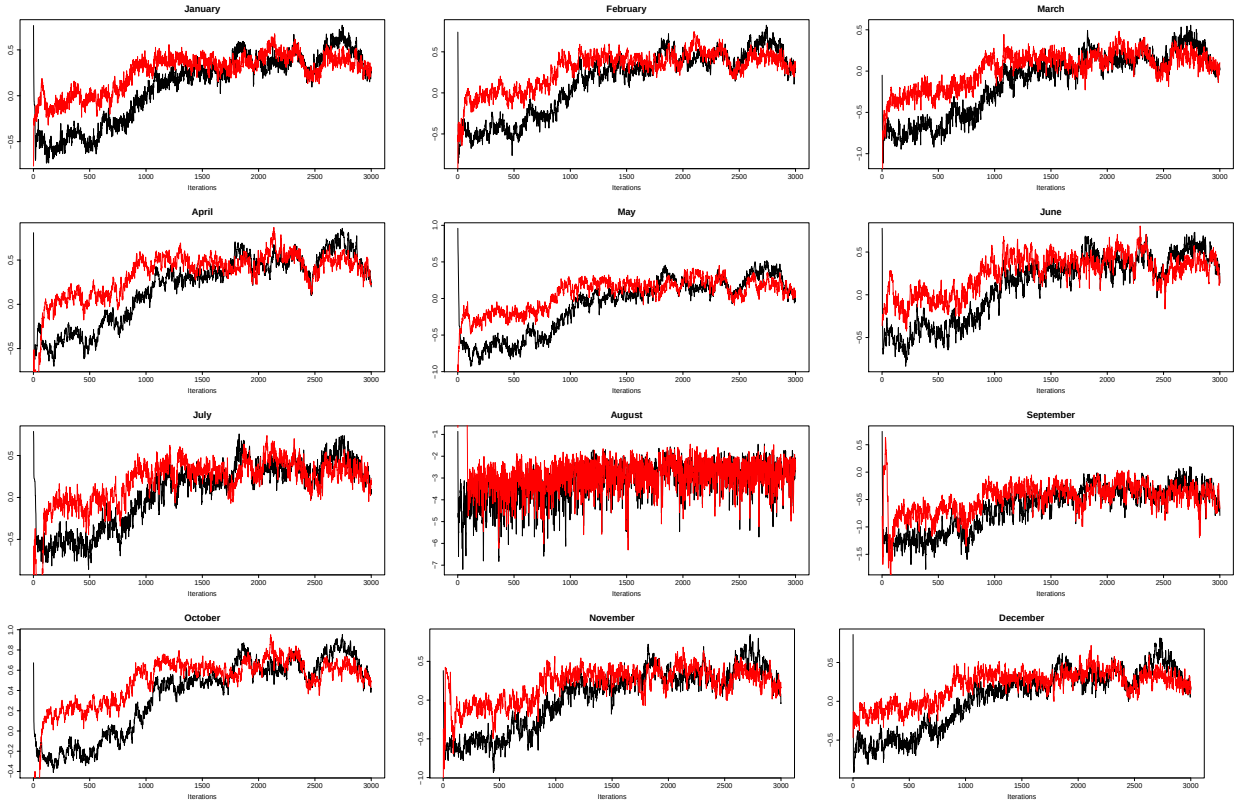

Figure 3.1: Trace plots of the seasonal effects for $\log \tilde{\sigma}$

As we can see from the plots in figure 3.1, the two different start values give chains which converge to the same stationary distribution. The thinned series is 3000 iterations long with convergence seen after approximately 1000 iterations, hence we disregard the first 1000 iterations as the "burn in" period and work with the remaining iterations. Ideally we would run the sampler for much longer - perhaps for 1 million iterations or more - but due to time restrictions it was not feasible to run for that many iterations. As an extension of this analysis, we could also experiment with the MCMC tuning parameters to increase the efficiency

of the sampler.

Figure 3.2 shows the MCMC output for Shell Beach in September after recombining the site and seasonal effects for Shell Beach and September respectively, for the GPD scale and shape parameters. Also shown are the plots of the posterior densities for these parameters. As table 3.1 shows, there is a substantial gain in precision over the separate site / separate season likelihood analysis due to the pooling of information across sites and seasons in the random effects model.

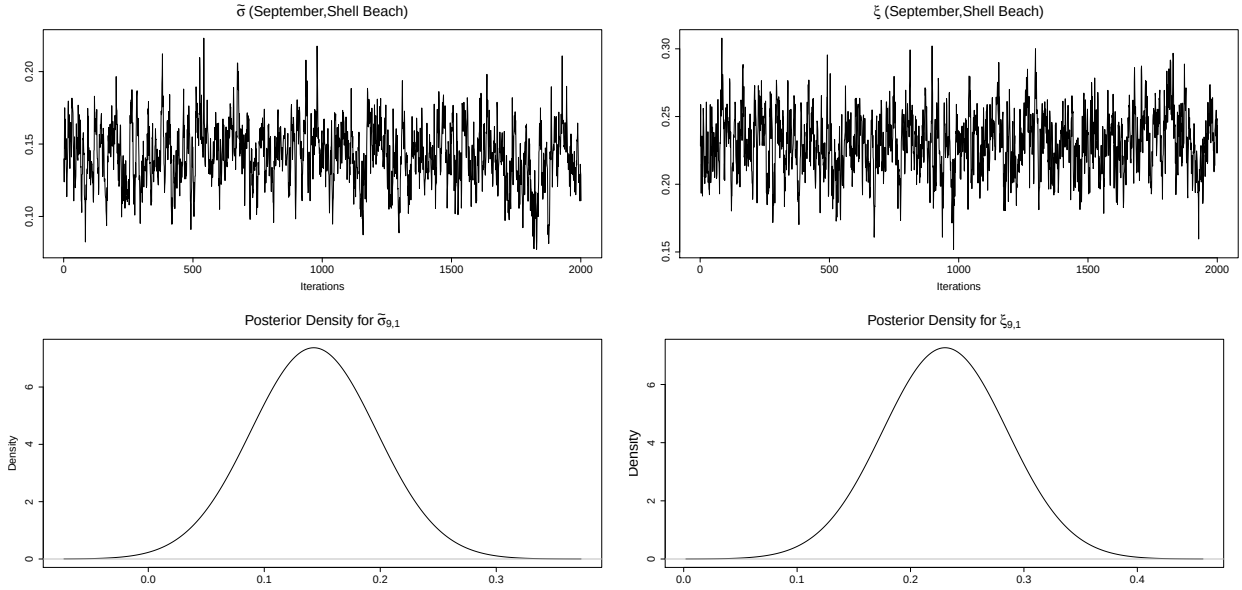

Figure 3.2: MCMC output for Shell Beach in September

Figure 3.3 shows hierarchical estimates of the GPD parameters $\tilde{\sigma}$ and ξ against the corresponding maximum likelihood estimates, (albeit from a separate site, separate seasons analysis). Both graphs show the range for the two parameters is much smaller when using the hierarchical model, in particular the range for ξ is much smaller for the hierarchical model. Again this shrinkage can be attributed to the way the random effects model pools information across sites and seasons.

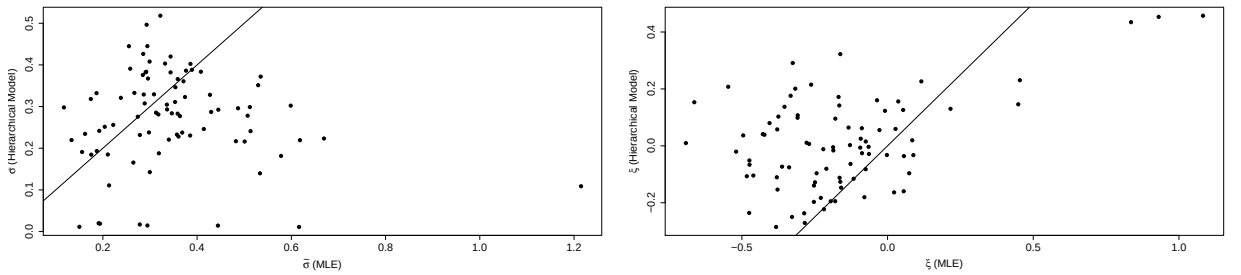

Figure 3.3: Posterior means against maximum likelihood estimates of GPD parameters

Table 3.1 shows the posterior means and standard deviations for Shell Beach in September

and Rockport in January. By choosing these two sites we can see the contrast of the site and seasonal characteristics by picking sites far apart and one month in the hurricane season with one month outside the hurricane season.

	Shell Beach, September		Rockport, January	
	Mean (st. dev.)	*MLE(st. err.)*	Mean (st. dev.)	*MLE(st. err.)*
$\gamma_{\tilde{\sigma}}^{(m)}$	-0.406 (0.187)		0.341 (0.152)	
$\gamma_{\xi}^{(m)}$	-0.0015 (0.0618)		-0.173 (0.0663)	
$\epsilon_{\tilde{\sigma}}^{(j)}$	-1.553 (0.155)		-1.805 (0.148)	
$\epsilon_{\xi}^{(j)}$	-0.232 (0.0625)		0.0657 (0.0617)	
$\tilde{\sigma}_{m,j}$	0.142 (0.0208)	*0.299 (0.040)*	0.232 (0.0144)	*0.278 (0.023)*
$\xi_{m,j}$	0.230 (0.0227)	*0.454 (0.119)*	-0.107 (0.0268)	*-0.483(0.048)*

Table 3.1: Posterior means and standard deviations with MLE estimates for GPD parameters

We can also see that table 3.1 shows the corresponding maximum likelihood estimates when applied to each site and season separately. The table shows one of the advantages of using the hierarchical model - a reduction in sampling variation. It is shown here that the standard deviations obtained from the posterior distributions are substantially smaller than the asymptotic standard errors from a likelihood analysis. This implies that combining the information into one complex model increases the precision of the analysis, hence favouring a random effects model here.

3.3.1 Seasonal Effects

Figure 3.4 shows the posterior means and 95% credible intervals for the seasonal effects for each GPD parameter. The first figure shows the sites effects for $\log \tilde{\sigma}$ and as we can see only August and September have a seasonal effect significantly different from zero as the credible intervals do not pass through zero. It is obvious from this plot that there is a strong seasonal effect in August, which we would expect as this is considered the height of the hurricane season. We would expect the hurricane season months to have a stronger effect than the other months. We can see this is true and August and September tend to be the busiest hurricane months and are the months here showing the most effect.

The second plot in figure 3.4 shows the ξ component of the seasonal effects. Here we can see the hurricane months, July, August and September, have a higher value for ξ.

3.3.2 Site Effects

Figure 3.5 shows the site effects and credible intervals for the GPD parameters. Sites 1,4 and 5 lie closely to each other and we can see this has an effect on the parameters. The three sites have similar parameter values for both $\log \tilde{\sigma}$ and ξ. The same can also be said for sites 2 and 3. Sites 6 and 7 were chosen to be further from all other sites, with site 6 in Florida and site 7 in Texas. We can see for $\log \tilde{\sigma}$ sites 2 and 4 have a higher value whilst for ξ sites 1,5 and 6 have a higher value.

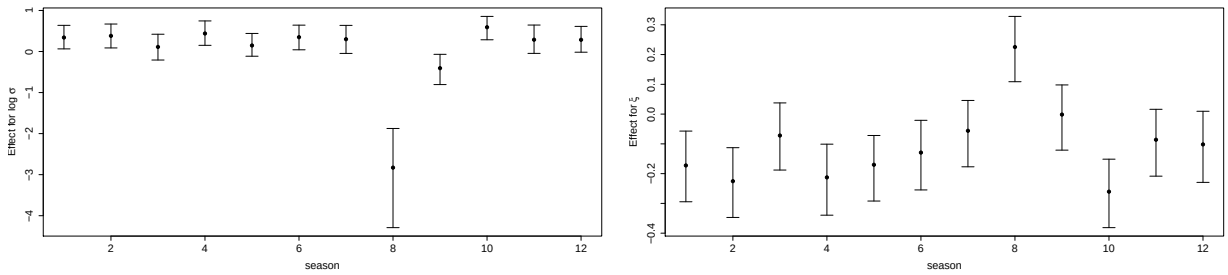

Figure 3.4: Posterior means and 95% credible intervals for the seasonal effects

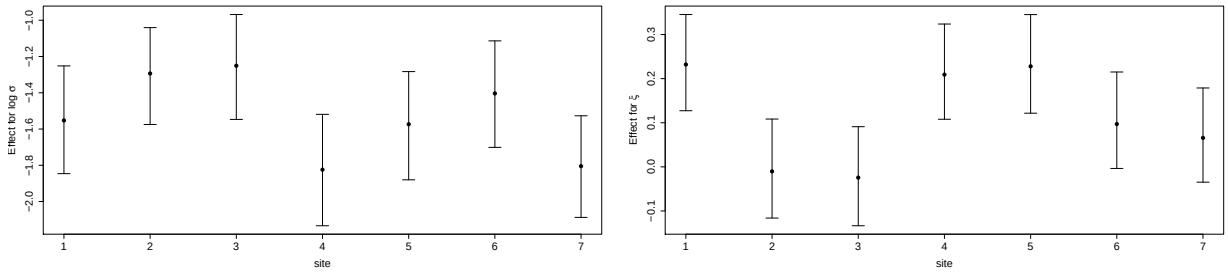

Figure 3.5: Posterior means and 95% credible intervals for the site effects

3.4 Dependence

The next stage of the analysis involves building in the dependence between observations for each site. We assume that since we have taken observations above a threshold we have dependence between one observation and the following observation. The autocorrelation plot in figure 1.3 shows a very large value at lag 1, hence implying an assumption of first order dependence might be suitable. Assuming first order dependence we can now model dependence between successive pairs of observations. Given a model $g(x_i, x_{i+1}; \psi)$ specified by parameter vector ψ, it follows that the likelihood equation for ψ is given by:

$$L(\psi) = f(x_1; \psi) \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \psi) / \prod_{i=1}^{n-1} f(x_i; \psi) \tag{3.14}$$

The denominator is the GPD for observations above a threshold. The numerator can be modelled by using a bivariate extreme value model, which will enable us to capture the first order Markov structure. Many models can be used here, however due to its flexibility and accessibility we choose the logistic model, with joint distribution function:

$$G(x_i, x_{i+1}) = \exp\left\{-\left(x_i^{-1/\alpha} + x_{i+1}^{-1/\alpha}\right)^{\alpha}\right\}. \tag{3.15}$$

The logistic distribution is symmetric which implies that the variables, x_i and x_{i+1}, are exchangeable. The flexibility and full coverage of all levels of dependence owe to the popularity of this choice of model. Other choices of distribution include the bilogistic model and the

Dirichlet model. (See Coles [2001], chapter 8, for more detail).

The extremal index, see section 1.3.3, can be estimated through the logistic dependence parameter α, where $\alpha \in (0, 1]$. It has been shown in Fawcett and Walshaw [2012] that there is a simple cubic equation which approximates θ through a cubic relationship with the logistic dependence parameter α. The link is given by the equation:

$$\theta = 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3. \tag{3.16}$$

A separate study of the dependence structure in our extremes at each site suggests the following estimates for the logistic dependence parameter α_j at each site $j = 1, \ldots, 7$:

$$\begin{aligned} \alpha_1 = 0.187, \quad \alpha_2 = 0.315, \quad \alpha_3 = 0.284, \quad \alpha_4 = 0.128, \\ \alpha_5 = 0.224, \quad \alpha_6 = 0.371, \quad \alpha_7 = 0.101. \end{aligned}$$

Most of the values are close to zero which implies there is strong dependence between readings at each site. To carry out further research using this model we could incorporate the logistic dependence parameter into the random effects model itself in order to more accurately model the logistic dependence. From these values of logistic dependence we can now find the corresponding values for the extremal index θ, via equation (3.16).

$$\begin{aligned} \theta_1 = 0.0547 \quad \theta_2 = 0.142 \quad \theta_3 = 0.118 \quad \theta_4 = 0.0295 \\ \theta_5 = 0.0756 \quad \theta_6 = 0.193 \quad \theta_7 = 0.0215 \end{aligned}$$

Now we can use these values to account for dependence in our return level estimation procedure.

3.5 Return Levels

In order to estimate return levels we also take dependence into account. To do this we set the return level equation for the GPD equal to $1 - r^{-1}$ in order to obtain the r-year return level. For each site j, $j = 1, \ldots, 7$ the return level z_r is given by

$$\sum_{m=1}^{12} \left\{1 - F_{m,j}(z_r)^{h_{m,j}\theta_j}\right\} = \frac{1}{r} \tag{3.17}$$

where $m = 1, \ldots, 12$. Here $h_{m,j}$ is the the number of hours in month m for each site j. This will put the return level on an annual scale. This equation adopts the standard approach to account for seasonal variability and dependence. The formula includes the extremal index, which we can see is included in the power. This is related to Theorem 1.3. The theorem showed that the dependent excesses will tend to a related distribution as the independent excesses, which is demonstrated here through the inclusion of θ. The extremal index θ_j has been estimated in the previous section, giving values for each site implicitly from the logistic dependence parameter α_j. Here $F_{m,j}$ is the GPD distribution function in month m for site j.

Substituting the formula for the GPD, see equation (1.10), into equation (3.17), we obtain the return level formula:

$$\sum_{m=1}^{12}\left\{1-\left(\lambda_u\left[1+\xi\left(\frac{z_r-u}{\tilde{\sigma}}\right)\right]^{-1/\xi}\right)^{h_{m,j}\theta_j}\right\}=\frac{1}{r}, \qquad j=1,\ldots,7. \tag{3.18}$$

This expression can now be solved numerically for z_r for each posterior observation of $\tilde{\sigma}_{m,j}$ and $\xi_{m,j}$ to obtain posterior observations for the return levels. The posterior means for two sites, Shell Beach and Rockport, are shown below in table 3.2. (Posterior standard deviations are shown in brackets). For comparision the MLEs with associated standard errors are also shown.

Model	Results for Shell Beach for the return periods (years)				Results for Rockport for the return periods (years)			
	10	50	100	1000	10	50	100	1000
Hierarchical	2.965	3.933	4.449	6.753	1.718	2.037	2.182	2.688
	(0.065)	(0.099)	(0.127)	(0.331)	(0.033)	(0.0453)	(0.0491)	(0.0549)
Maximum	2.023	4.214	6.545	45.77	1.148	1.554	1.719	2.000
Likelihood	(0.133)	(0.201)	(0.363)	(2.055)	(0.101)	(0.190)	(0.202)	(0.668)
Predictive	4.213	5.628	7.501	10.323	2.400	3.569	4.511	4.888

Table 3.2: Return Levels for Shell Beach and Rockport (feet)

As we can see from table 3.2 the return levels calculated for Shell Beach using the hierarchical model are much lower and reasonable compared to the values calculated using maximum likelihood estimation. It is reasonable to assume that during a hurricane we will see high sea levels however it is incredibly unlikely to see the height of the sea to be 45 feet above the Mean High Water level. The estimates for the return levels for Rockport are similar when comparing both methods.

As discussed in chapter 2, and one of the virtues of a Bayesian analysis is its natural to prediction. Using the MCMC samples, the RHS of equation (2.1) can be approximated with

$$Pr(X\leq z|x)=\frac{1}{B}\sum_{k=1}^{B}Pr(X\leq z|\theta^k) \tag{3.19}$$

where θ is the full parameter vector and B is the number of iterations after burn-in. Solving equation (3.19) for $z=\hat{z}_{r,pred}$ for various values of r gives estimates of the r-year predictive return level; a single value including all sources of variability in parameter estimation and future observations so greatly valued by practitioners. These values are higher than the values estimated through the hierarchical model as well as the likelihood analysis. These results however are more reliable as we have accounted for all variability in parameter estimation.

We can also compare the posterior standard deviations and the asymptotic standard errors. It is relatively simple to obtain the posterior standard deviations as we have generated

3000 return level estimates from the MCMC output. The asymptotic standard errors were obtained by using the delta method, as demonstrated in chapter 1. Table 3.2 shows a substantial reduction in the posterior standard deviations compared to the asymptotic standard errors. The pooling of information across seasons and sites has resulted in a reduction in variability, giving more precise return level estimates.

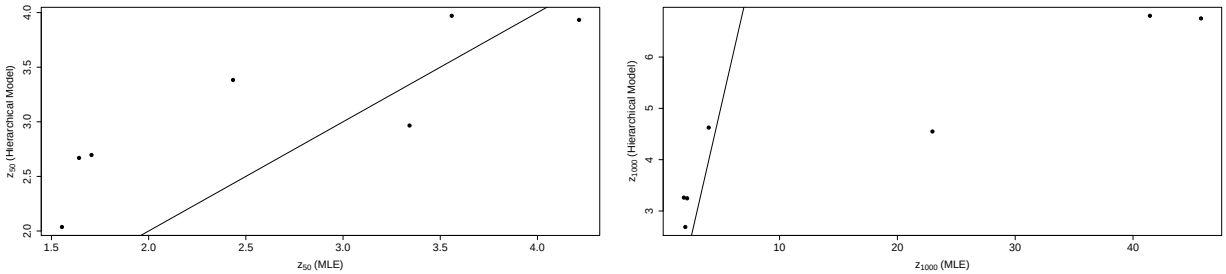

Figure 3.6: Hierarchical mean estimates against maximum likelihood estimates of i) 50 year return level and ii) 1000 year return level

Figure 3.6 plots the return levels for the hierarchical model against the maximum likelihood estimates. As we can see the range for return levels is much smaller when using the hierarchical model, mainly demonstrated through the 1000-year return level plot. The values obtained using maximum likelihood estimation appear to be much higher and in practical terms are unrealistic. The shrinkage in the range of the estimates for the return levels shows that there is a reduction in variability when using the hierarchical model. This implies that the complex model helps to account for seasonal and site variability and increases the accuracy when estimating the values for the parameters for the GPD as well as when estimating the return levels.

Chapter 4

Conclusion

Finally we can compare the models we have used to analyse the sea surge data in order to decide which is the most effective model.

When comparing the maximum likelihood estimates obtained in chapter 1 with the Bayesian estimates in chapter 2, we notice there is little difference between the estimates and between the posterior standard deviations and standard errors, as shown in table 4.1. In order to reduce the posterior variability in the Bayesian analysis we can include more information in our analysis, through the prior distributions. An extension of this project could include asking a civil engineer to provide expert information on how sea surges are taken into account when building a new section of a flood defence system such as a sea wall; or, perhaps, involving an oceanographer in the elicitation of informative priors for the GPD parameters. This could help increase the precision of the parameter estimates as well as the precision of the estimated return levels.

	Maximum Likelihood	Bayesian
$\hat{\sigma}$	0.185 (0.066)	0.188 (0.069)
$\hat{\xi}$	0.559 (0.313)	0.769 (0.428)

Table 4.1: MLE (std. err.) and posterior means (std. dev)

The hierarchical model constructed in chapter 3 gives a demonstration of how a complex model can be used in an extreme value analysis of environmental data. Due to the inherent variability within environmental data it is preferable to use a complex model such as a random effects model which can better account for the seasonality, and site effects, within the data. This is demonstrated in detail in chapter 3.

The hierarchical model has accounted for the seasonal and site variation, as we can see from table 3.2 showing the standard deviations for return levels. These values are considerably smaller than the asymptotic standard errors from the likelihood analysis. This shows that the complex model has reduced the variability giving more precise estimates for return levels as well as more realistic estimates. The return levels estimated from the likelihood analysis are unrealistic and, along with larger standard errors, are impractical to work with from a practitioners standpoint. Table 3.2 also shows the predictive return levels. These

estimates incorporate all sources of variability and are advantageous in practical situations as we no longer need a standard error.

Overall the hierarchical model used has shown to be the preferred approach to the analysis of this data set. The ability to account for the seasonal and site effects through a hierarchical structure has shown to be a more efficient approach, leading to an increase in precision of parameter estimates.

An improvement on the model would be to use informative priors instead of vague priors as we do in our analysis. Also, it would be preferable to run the MCMC algorithms used in chapter 3 for much longer. This would give more iterations to work with once the burn-in period has been removed.

In chapter 3 we only included the extremal dependence parameter θ when estimating return levels. It would be preferable to include the dependence parameter in the hierarchical model in order to obtain more precise estimates. Also, we only only looked at first order dependence by assuming dependence between consecutive pairs of observations. Further research could provide an insight into how estimates change when looking at 2nd order dependence or longer range dependence through multivariate extreme value models compared to the bivariate analysis carried out in this project.

Finally, further research could investigate the use of other models for our random effects parameters. Figure 3.3 shows the degree of shrinkage of our parameter estimates relative to a more simple likelihood analysis. It could be that the hierarchical model has over-smoothed, and t-distributions could be used to give more realistic estimates.

References

S. Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

S. Coles and E. A. Powell. Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review/Revue Internationale de Statistique*, 64:119–136, 1996.

S. Coles and A. Stephenson. *ismev: An Introduction to Statistical Modeling of Extreme Values*, 2011. URL `http://CRAN.R-project.org/package=ismev`. R package version 1.36.

L. Fawcett and D. Walshaw. A hierarchical model for extreme wind speeds. *Applied Statistics*, 55:631–646, 2006.

L. Fawcett and D. Walshaw. Estimating return levels from serially dependent extremes. *Environmetrics*, 23:272–283, 2012.

R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. 24(02):180–190, 1928.

B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, pages 423–453, 1943.

W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.