Bayesian Inference for Sea-surge Extremes in the Gulf of Mexico

Alexandra Lee
110406389

Supervisor: Dr. Lee Fawcett

Academic Year: 2014/15

MMathStat: Master of Mathematics and Statistics

Abstract

Working primarily within the Bayesian framework, this report will consider various methods for accounting for both temporal and spatial dependence in environmental extremes, to build upon standard methods from the extreme value theory toolbox. Severe weather conditions affect all forms of life, resulting in different problems with each type of extreme event. This report will focus on sea-surges in the Gulf of Mexico as the severe weather condition, and will discuss how we can use statistical models to extrapolate into the future. Sea-surges lead to mass flooding so it is important that we put in preventative measures to keep damage, destruction and loss of life to a minimum. We will look at several techniques used to model extremes, first assuming independence, and then progressing through the report by accounting for dependence, showing the problems associated with incorrectly assuming extremes are independent. We will look at estimates that will help with the design of the sea walls along the Gulf coast of Mexico and comment on further work that could lead on from this project.

Contents

Chapter 1

Background and Motivation

1.1 Motivation

This report will consider various methods for accounting for both temporal and spatial dependence in environmental extremes, working primarily within the Bayesian framework to build upon standard methods from the extreme value theory toolbox. There is a vast range of practical applications of extreme value theory within this field. For example, the aiding of the design of a new flood defence system to protect against the once-in-a-hundred year flood event; informing design codes for new buildings and other structures, particularly bridges, to protect against wind speed extremes; and providing estimates of the severity of cold spells to help plan fuel stockpiles. Climate change is here and upon us, our environment is changing, and not necessarily for the better. Recent years have seen an increase in extreme weather conditions so it is vital that we can estimate how severe the next serious weather condition will be, to reduce undesirable consequences. As statisticians, we provide information which leads to the prediction of the r-year return level - which can be though of as the extreme (e.g. sea-surge, wind speed, temperature, or rainfall event) which we can expect to see, on average, once every r years[1]. By their very definition, extremes are scarce, thus to provide estimates for periods beyond what we have available data we must extrapolate past what has been observed, which can lead to some difficulties. However the intended benefits of these applications outweigh the problems as we can save lives and reduce the financial burden of extreme climatic events.

In the Netherlands, a low-lying country, the state commissions use extrapolation and have determined an acceptable wave height return period of $1,250$ years[2]. The Delta programme by the Dutch Delta Commissioner, a government scheme, with an average annual budget of €1 billion, aims to ensure that water safety and freshwater supply are robust by 2050, and to equip the country better to withstand weather extremes[3]. Since the great North Sea flood in 1953 (which affected the Netherlands, Belgium, Scotland and England), it was decided that the country needed more protection against flooding. Sixty years on, with more negative factors influencing potential flooding, for example, a higher population density; rising temperatures; more severe rainfall; and the subsidence of land, it was vital that plans were made to further protect the country. Hence the establishment of the Delta programme, which uses estimates of the $1,250$-year return levels for wave heights without having this many years of data.

Another practical example of extrapolation is the work that the British Standards Institution (BSI) complete. They use estimates of the 50-year wind gust speed to inform their design codes for new buildings and other structures to protect against wind damage we might expect to see, on average, once every 50 years. They incorporate altitude, season and direction to inform their design codes, which work specifically for sites in the UK[4].

In the Gulf of Mexico, BP's use of estimates of the 500-year return level for sea-surges, to protect their oil platforms against damage, is a further example of extrapolation in practice. There is a tendency to concentrate on the resulting disasters which are caused by oil spills, and the work that oil companies do to protect their oil rigs from damage is often overlooked. Extrapolating 500 years into the future is a difficult task and shows that to the best of their abilities, BP are planning for the next severe sea-surge.

We will now look at how inadequate planning for environmental extremes can result in extensive loss of life and gross monetary problems. The North Sea flood, 1953, was especially damaging due to an unusually high tide; strong winds which pushed water over the sea defences in place; and severe wave action[5]. The consequences were shocking, around $1,800$ deaths in the Netherlands, about 300 deaths in England, and around $72,000$ people were evacuated in the Netherlands alone[6], along with damage to buildings; many cattle dying; and the failure of crops. Figure 1.1 shows the height, above normal tide to which the sea rose, 2.5m above normal tide level in some areas, and figure 1.2 shows Canvey Island, Essex, after the floods. As the flood defences in place at the time were inadequate, if such a storm hit again in the near future it could have caused identical destruction, thus protective measures were established: the Thames barrier was built in England, and the Delta programme developed in the Netherlands. The Thames barrier, see figure 1.3, is a huge steel structure, with six swivelling steel gates acting as an enormous drawbridge. It cost approximately £1.6 billion[8] to construct, and was built to protect against floods which could occur once in every thousand years. With 174 closures since it became operational in 1982 (correct as of March 2014), it has protected many lives, and properties over 50 square miles[8]. A comparable storm to the one in 1953 was seen in 2007, but London did not flood because of the barrier. Figure 1.4 shows what could have been, had the Thames barrier not been in place. Recent years have seen an increase in the closure of the barrier, with 48 closures in 2014 alone[9], hence promoting the continual study of environmental extremes to establish whether our return level estimates change.

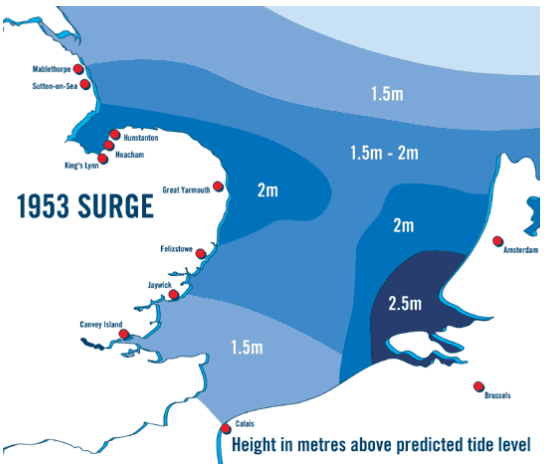

Figure 1.1: Estimated height of sea above normal tide level (metres)[6]

Figure 1.2: Residents of Canvey Island, Essex, are rescued by boat[7]

The European heatwave of 2003 saw temperature records broken in a number of countries, France saw some of the hottest on record, reaching 40 °C in Paris. Approximately

Figure 1.3: Thames barrier in use[10]

Figure 1.4: Estimated flooding which could occur without the Thames barrier[11]

70,000[13] died with around 15,000 of these being in France alone; in addition to crops failing, rivers drying up, and fires being fuelled[12]. Another European heatwave on a similar scale occurred in 2006, which was less intense and covered less geographical area than in 2003, however it did last longer. Maximum temperatures in France again reached 40 °C, yet an excess of only 2000 deaths occurred to the month's "norm", when 6500 were expected. Reasons for this include: denial of the seriousness of the event by authorities in 2003 thus emergency-level responses were not tested; preventative measures put in place after 2003, in which several countries initiated plans to prepare for the future with France developing a 'national heatwave plan'; surveillance activities; clinical treatment for heat-related illness; identification of the vulnerable; and improved infrastructure[13]. These actions reduced the death toll when the next heatwave occurred, thus promoting the study of extremes.

Preventative measures in place were not substantial when Hurricane Katrina, see figure 1.5, hit the Gulf Coast on 29th August 2005. There were around 2,000 deaths, and approximately 90,000 square miles of the United States were affected, resulting in over $100 billion worth of damage[14]. In parts of New Orleans, water levels reached 9 metres, thus many sea defences were breached resulting in numerous low-lying areas of land being completely submerged under water, see figure 1.6. Various changes were established in New Orleans, including but not limited to: increased height and sturdier structures for sea defences; storm-proofed pump stations, prepared for increased water volume; diversion of the Mississippi river freshwater containing nutrients to the wetlands around the city, which buffer hurricanes; and improved hurricane modelling, which used storm size and intensity[17].

The remainder of this report will consider sea-surges as the extreme weather condition. A sea-surge is one component of the overall height of the sea at any point in time (with another component being wave height, and another tide). During a storm, the combined effects of low air pressure, strong wind speeds and heavy rainfall induce extreme sea-surges, whereby the high pressure forces sea water to the coast and low pressure at the eye (centre) of the storm pulls the water level up, which is analogous to using a straw. High rainfall and strong winds then either push the water over the sea defences, or damage

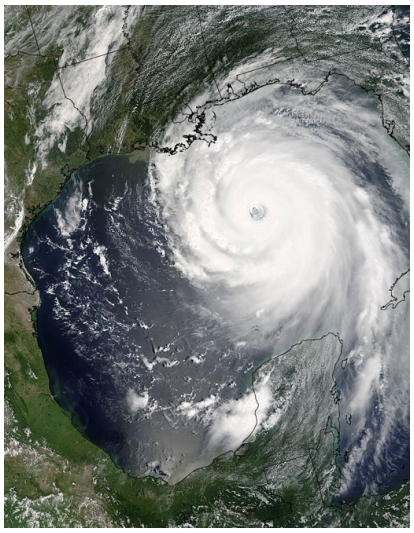

Figure 1.5: Satellite image of Hurricane Katrina[15]

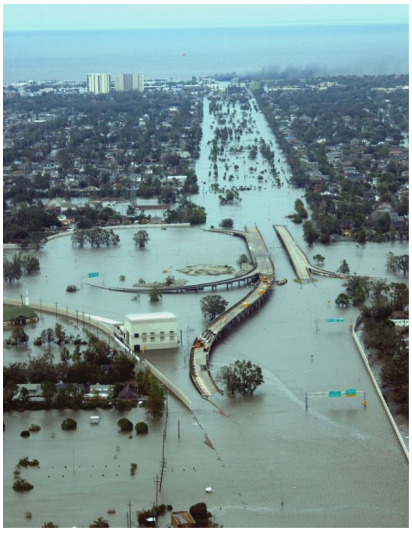

Figure 1.6: Flooded I-10/I-610 interchange and surrounding area of northwest New Orleans and Metairie, Louisiana[16]

them and so flooding can occur, hence it is coastal areas which are most damaged. The sea defences installed at coastal areas, defined as levees, are a barrier preventing water from getting to, and damaging land. Statisticians assist engineers with the design height of the levee, by relaying back the estimated return levels. When building levees, there is a trade-off between safety and financial burden, as building high levees is very expensive, and they must be built within the constraints of local authority budgets. A barrier never breached and a barrier often breached are wasteful, we need value for money, but without compromising safety. In essence we must find the optimal height.

1.2 The Data

1.2.1 Data Locations

This report will use data downloaded from the 'National Oceanic and Atmospheric Administration', (NOAA) webpage, for locations just off the coastline of Texas and Louisiana[18]. There are five years of hourly observations of water levels from five sites, and after removing elements where data was missing; there is approximately $31,000$ pieces of data for each site. Figure 1.7 shows the five locations studied. Sabine Pass and Galveston are geographically close, as are Port Fourchon and Grand Isle, thus similar return levels are expected at each of these two clusters, without first performing analyses. Berwick, a wetland area, with an inland position, could have reduced return level estimates. This is seen with coastal areas being worse hit by sea-surges and work by Masters, showing wetland and marsh areas can actually reduce the height of a sea-surge[20], though there is very large variation in this. Wetlands reducing the height of sea-surges were seen during Hurricane Rita in 2005, where a 15ft sea-surge hit the western coast of Louisiana, but was reduced by the wetlands by approximately 1ft per $2.1 - 3.6$ miles of inland penetration (simulation by Resio and Westerink[20]).

Figure 1.7: Map of the Gulf of Mexico showing the 5 locations studied (L-R): Galveston (G), Sabine Pass (SP), Berwick (B), Port Fourchon (PF) and Grand Isle (GI)[19]

1.2.2 Exploratory Analysis: Sabine Pass

We will now do a short exploratory analysis at one site - Sabine Pass. Figure 1.8 shows a time series plot, and a histogram of the surges seen at this site. From the time series plot

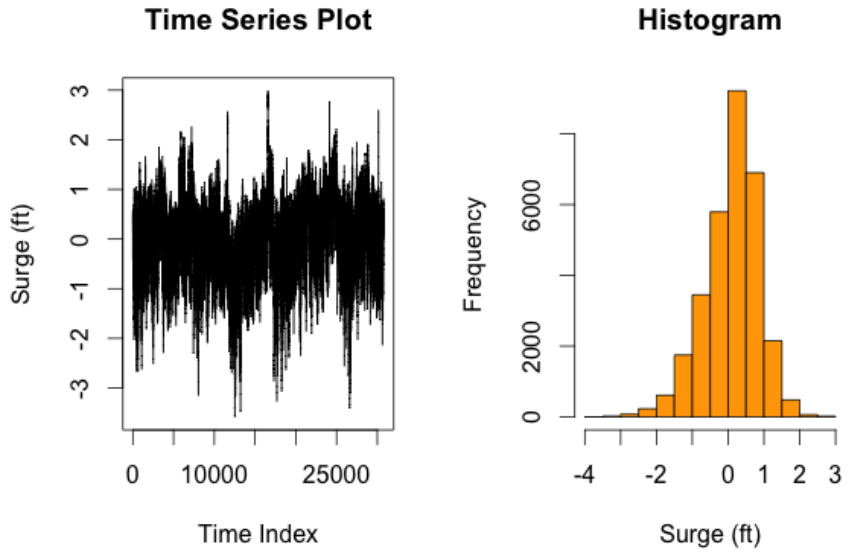

Figure 1.8: Exploratory Analysis at Sabine Pass

in figure 1.8 there are no increases/decreases, thus no trend in the data. If we speculate, one reason for the lack of trend could be that the data collected doesn't extend back far enough to show one, one may only be visible if we had 100 years of observations. There appears to be some seasonal variability as there are some peaks and troughs. However the seasonal variability may be caused by incorrectly assuming independence between observations, which we will account for this in chapter 2. The histogram in figure 1.8 shows all the surges recorded in the five years. The optimal surge height is from $0-0.5$ft, with a frequency of above 8000 times. The mean surge height is 0.102ft, but there are some surges reaching 3ft and some as low as -4ft.

1.3 Statistical Modelling of Extremes

There are various ways in which we can use the data collected to model extremes, and some of these will be discussed now.

1.3.1 Block Maxima Approach

The block maxima approach[21] was the first method to be used when modelling extremes and provides the foundation of extreme value theory. Suppose we have a sequence X_i, $i = 1, 2, \ldots, n$, of independent and identically distributed (IID) random variables with common distribution F, and we aim to focus on the statistical behaviour of

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

In practical applications, the X_i can be thought of as processes measured on a regular time scale, i.e. hourly sea-surges, thus M_n represents the maximum of the process over n time units. If we let n be the number of observations in a year, then M_n is the annual maximum. The distribution of M_n in equation 1.1, can be derived for all values of n, and can be seen in equation 1.2:

$$\begin{aligned} \Pr(M_n \leq x) &= \Pr(X_1 \leq x, X_2 \leq x, ..., X_n \leq x), \\ &= \Pr(X_1 \leq x)\Pr(X_2 \leq x)...\Pr(X_n \leq x), \\ &= \{F(x)\}^n, \end{aligned} \tag{1.2}$$

where we can multiply marginal components together because we have independent observations. The distribution of F is unknown, however if we model extremes in the way defined by equations 1.1 and 1.2, we are interested in limiting models for F^n regardless of what F is itself. This is not unlike the central limit theorem (CLT), where, for a large sample size, the distribution of the sample mean is approximately Normal - regardless of the distribution of the parent population. In essence we have an extreme value analog of the CLT.

Recall the Central Limit Theorem, which says that

$$\frac{\bar{X} - b_n}{a_n} \xrightarrow{D} N(0, 1)$$

where $a_n = \mu$ and $b_n = \sigma/\sqrt{n}$, with $\bar{X}$, σ and n being the sample mean, population standard deviation and sample size, respectively. We can apply a similar rescaling to M_n to avoid convergence of the distribution to a singular point. This can be seen in the Extremal Types Theorem.

Theorem 1: *The Extremal Types Theorem*

The extremal types theorem[1] states that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\{(M_n - b_n)/a_n \leq x\} \to G(x) \qquad \text{as } n \to \infty, \tag{1.3}$$

where $G(x)$ is a non-degenerate distribution function, then G belongs to one of the following three families:

$$\begin{aligned} I : G(x) &= \exp\{-\exp(-x)\}, \quad -\infty < x < \infty \\ II : G(x) &= \begin{cases} 0, & x \leq 0 \\ \exp\{-x^{-\alpha}\}, & x > 0, \alpha > 0 \end{cases} \\ III : G(x) &= \begin{cases} \exp\{-(-x)^{-\alpha}\}, & x < 0, \alpha > 0 \\ 0, & x \geq 0. \end{cases} \end{aligned}$$

The three distributions in the Extremal Types Theorem are known the Gumbel, Fréchet and Weibull distributions respectively - known more generally as the *extreme value distributions*. Although the Extremal Types Theorem does not state which of the three distributions is applicable, nor does it ensure the existence of a non-degenerate limit for M_n, it does say that if a limiting distribution exists then no matter what the parent distribution F is, the limiting distribution of the sample maxima follows one of I, II or III.

However there is a problem with this in that although we have three distributions for the maximum values, we do not know which one of these would be the best to choose. This problem was resolved in 1954 and 1955 by Von Mises and Jenkinson[21] respectively, who worked separately to combine the three distributions into a single family of models known as the Generalized Extreme Value (GEV) Distribution. The cumulative distribution function (CDF) of the GEV is:

$$G(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \right\}, \tag{1.4}$$

where $a_+ = \max(0, a)$ and the parameters μ $(-\infty < \mu < \infty), \sigma$ (> 0) and ξ $(-\infty < \xi < \infty)$ are known as the location, scale and shape parameters respectively. It is the shape parameter ξ that differentiates between the three types of extreme value distribution and uncertainty between each one is accounted for in our uncertainty about ξ. The Fréchet (type II) and Weibull (type III) classes of extreme value distribution correspond to the cases $\xi > 0$ and $\xi < 0$ respectively. However, equation 1.4 does not hold if $\xi = 0$ and so we take the limit as $\xi \to \infty$, giving,

$$G(x; \mu, \sigma) = \exp\left\{ -\exp\left(\frac{x-\mu}{\sigma}\right) \right\}, \tag{1.5}$$

which gives the Gumbel (type I) class of extreme value distribution.

So we now have a distribution function for our maxima given by equations 1.4 and 1.5, however if we recall equation 1.3 we see that we also had constants a_n and b_n. In fact for large n, we have

$$\begin{aligned} \Pr\{M_n \leq x\} &\approx G\{(x - b_n)/a_n\} \\ &= G^*(x) \\ &= G(\mu^*, \sigma^*, \xi), \end{aligned} \tag{1.6}$$

where we have absorbed the constants a_n and b_n into μ^* and σ^*. In practice, we can simply fit the GEV to our maxima and ignore the normalization constants, because the GEV parameters must be estimated anyway.

Practical Use of the Block Maxima Approach

In practice there are 4 steps to follow when using the block maxima approach. We will discuss them here, however it is important to be aware that although the block maxima approach used to be the most widely used method to model extremes, there have now been new methods introduced, and they will be considered in more detail as this report advances. As I have previously mentioned, we need these methods as they provide the foundations for other methods. The four steps are detailed below.

1. As we often don't have filtered block maxima, the first step is to obtain the M_n. To do this you must choose a block length n, often we choose a calendar year. We then discard all but the maxima observation in each block. There can be some problems with choosing n, for example if n is too small then the limiting arguments we made will not hold. However if n is too large, then we won't have enough maxima to work with.

2. Next we estimate the GEV parameters. In a frequentist setting we maximize the log-likelihood to obtain these estimates, and appeal to standard asymptotic likelihood theory to obtain their standard errors. Within the Bayesian framework, the GEV likelihood is used as an "ingredient", along with prior beliefs about the model parameters, to formulate our posterior beliefs about these parameters - we then summarize the marginal posteriors for the parameters using posterior means/medians/modes, standard deviations and quantiles. The likelihood is formed by calculating

$$\prod_{i=1}^{n} g(x_i, \mu, \sigma, \xi),$$

where g is the probability density function (PDF) of the GEV:

$$\mathrm{g}(x_i, \mu, \sigma, \xi) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]_+^{-(1/\xi+1)} \exp\left\{-\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}.$$

3. We must then check various goodness of fit properties to check the overall adequacy of the fitted GEV. For example we could use probability plots or QQ-plots. With a probability plot the general idea is that the data are plotted against a theoretical distribution such that if F is a reasonable model for the population distribution, then the points of the probability plot should lie along the unit diagonal, and departures from linearity provide evidence of a poor fit of the model. A QQ-plot has the basic idea that we compute the theoretical expected value for each data point based on the distribution, and as with the probability plot, if F is a reasonable model then the points of the plot should lie close to the unit diagonal. See chapter 2.6.7 in Coles (2001)[1] for full descriptions.

4. The last but perhaps the most important step is to estimate the return levels, in essence this is the whole reason we are completing our work, and in context with my data it is so we can provide information to use in the estimation of the height the levees. To estimate the r-year return level, z_r, We must set the CDF for the GEV, equation 1.4 equal to $1 - 1/r$ and solve for $x = z_r$.

So we have:

$$\Pr(\text{annual maximum} > z_r) = \frac{1}{r}$$

Which can be rearranged to give:

$$1 - \Pr(\text{annual maximum} \leq z_r) = \frac{1}{r} \tag{1.7}$$

Now if we look at the left-hand-side of equation 1.7, we can see that in terms of our fitted GEV, it is:

$$1 - G(z_r; \mu, \sigma, \xi)$$

and so we can write:

$$1 - \exp\left\{-\left[1 + \xi\left(\frac{z_r - \mu}{\sigma}\right)\right]^{-1/\xi}\right\} = r^{-1}. \tag{1.8}$$

We then rearrange equation 1.8 and solve for z_r to get the estimate of the r-year return level

$$z_r = \mu + \frac{\sigma}{\xi}\left\{\left[-\log\left(1 - r^{-1}\right)\right]^{-\xi} - 1\right\}.$$

Recall that when $\xi = 0$ we must work with the limiting form of the GEV, equation 1.5. The parameters μ, σ and ξ can be replaced with their maximum likelihood estimates or - within a Bayesian context - draws from the marginal posteriors (see section 1.4) to obtain the posterior distribution for z_r. Estimation uncertainty can be accounted for via the delta method[1] within a frequentist setting, or by direct reference to the posterior standard deviation for z_r within the Bayesian framework.

The block maxima approach is not the best method to use, as to reduce problems of dependence and non-stationarity we need to use blocks as large as possible, for example yearly data as oppose to monthly data, which is very wasteful of data. If we used monthly data then there are often problems with seasonal variability with winter months observing lower surges than summer months (as the hurricane seasons is June - November). Taking annual blocks often avoids this issue, however in this dataset for this project, we would reduce the five years of data ($31,000$ observations) to just five pieces of data for each site, which throws away a tremendous amount of data, and leaves little data to make inferences on. This also has detrimental effect on the precision of return level estimates, with standard errors/posterior standard deviations being large, owing to the inclusion of so little data.

1.3.2 Method Of Threshold Excesses

A much more flexible, less wasteful method for classifying extremes is the method of threshold excesses[21], which considers all observations above some high threshold u.

Distribution of Threshold Excesses

For a large enough threshold u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{1.9}$$

for $y > 0$ and where:

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

Outline Proof[1]

Let X denote an arbitrary term in the X_i sequence and follow the distribution function F. If we assume the *Extremal Types theorem* holds, for large enough n,

$$F^n(x) \approx G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

for some parameters μ, $\sigma > 0$ and ξ. Hence:

$$n \log F(x) \approx -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}. \tag{1.10}$$

For large values of x, a Taylor expansion implies that

$$\log F(x) \approx -\{1 - F(x)\}, \tag{1.11}$$

and if we substitute equation 1.11 into equation 1.10, rearrange, and replace x by u, then we obtain:

$$1 - F(u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}, \tag{1.12}$$

for large u. Similarly, for $y > 0$,

$$1 - F(u+y) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}, \tag{1.13}$$

Hence, using equations 1.12 and 1.13,

$$\begin{aligned}
\Pr(X > u + y | X > u) &= \frac{1 - F(u+y)}{1 - F(u)} \\
&\approx \frac{n^{-1}[1 + \xi(u+y-\mu)/\sigma]^{-1/\xi}}{n^{-1}[1 + \xi(u-\mu)/\sigma]^{-1/\xi}} \\
&= \left[1 + \frac{\xi(u+y-\mu)/\sigma}{1 + \xi(u-\mu)/\sigma}\right]^{-1/\xi} = \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi},
\end{aligned}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ as required.

Equation 1.9 is the CDF of the Generalized Pareto Distribution (GPD) from the Generalized Pareto family, and this will be the distribution used for modelling threshold excesses. So the reason we needed the theory in section 1.3.1 is that if our block maxima have the GEV distribution, then the threshold excesses must have the distribution of the GPD. For notational convenience we drop the tilde on the σ and refer to the scale in both the GEV and GPD as σ, so we have parameters σ (> 0) and ξ ($-\infty < \xi < \infty$). It is the parameter ξ which determines the tail behaviour of the GPD, if $\xi < 0$ the distribution of excesses has an upper bound, if $\xi > 0$ the distribution has no upper limit, and finally, if $\xi = 0$ the distribution is unbounded, and we must take the limit $\xi \to 0$, giving:

$$H(y) = 1 - \exp\left(-\frac{y}{\sigma}\right). \tag{1.14}$$

How Do We Do This In Practice?

Discussed at the beginning of this section, we need to estimate what the threshold, which we will denote u_0, should be. By the threshold stability property of the GPD, if the GPD is a suitable distribution for excesses over some threshold u_0, then it is also valid for excesses over all thresholds $u > u_0$. Looking at the expected value of our threshold excesses, again conditional on being already greater than the threshold, we have:

$$E[X - u|X > u] = \frac{\sigma_{u_0} + \xi u}{1 - \xi}, \tag{1.15}$$

where σ_{u_0} is the GDP scale above the threshold. It can be seen that in equation 1.15 this function is linear in u. Note $E[X - u|X > u]$ is the mean of the excesses of the threshold u. We use a mean residual life (MRL) plot to identify a threshold, where we plot the mean threshold excess against u, and then we look for a value u_0 above which we observe approximate linearity. To see this in practice for one of the sites, say Sabine Pass, see figure 1.10 in section 1.5. From this plot approximate linearity is observed at 1.4ft, and so we take $u_0 = 1.4$ to be our threshold.

Analogous to the GEV, we must calculate the GPD parameters, in the same ways discussed. The likelihood is formed by calculating

$$\prod_{i=1}^{n} h(y_i; \sigma, \xi),$$

where h is the PDF of the GPD:

$$h(y_i; \sigma, \xi) = \frac{1}{\sigma}\left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi - 1}$$

As with the GEV we must obtain an equation for the r-year return level, however there is another parameter we must incorporate, λ, the threshold exceedance rate. We know

$$\Pr(X > u + y|X > u) \approx \left[1 + \frac{\xi y}{\sigma}\right]_+^{-1/\xi}, \tag{1.16}$$

for $\xi \neq 0$. If we focus on the left-hand-side of equation 1.16, we see that

$$\Pr(X > u + y|X > u) = \frac{\Pr(X > u + y)}{\Pr(X > u)},$$

which can be rearranged to form:

$$\Pr(X > u + y) = \Pr(X > u)\Pr(X > u + y|X > u). \tag{1.17}$$

If we substitute equation 1.16 into equation 1.17, we obtain:

$$\Pr(X > u + y) \approx \lambda_u \left[1 + \frac{\xi y}{\sigma}\right]_+^{-1/\xi}, \tag{1.18}$$

where $\hat{\lambda}_u = \Pr(X > u)$ - the threshold exceedance rate. Now if we substitute $y_i = x_i - u$ into equation 1.18 we obtain:

$$\Pr(X > x) \approx \lambda_u \left[1 + \xi\left(\frac{x - u}{\sigma}\right)\right]_+^{-1/\xi}. \tag{1.19}$$

We can now obtain an estimate of the return level z_t, which is exceeded, on average once every t observations, by setting equation 1.19 equal to $1/t$, and rearranging to make z_t the subject of to give:

$$z_t = u + \frac{\sigma}{\xi}\left[(t\lambda_u)^{\xi} - 1\right], \quad \text{if } \xi \neq 0,$$

and

$$z_t = u + \sigma \log(t\lambda_u), \quad \text{if } \xi = 0.$$

By construction, we have z_t as the t-observation return level, however it is much more convenient to have our return levels on an annual scale, as in we want the r-year return level and so we must replace t with $r \times n_y$, where n_y is the number of observations per year and r is the return level to be estimated. We can then define the equation for the r-year return level, z_r, to be

$$z_r = u + \frac{\sigma}{\xi}\left[(rn_y\lambda_u)^{\xi} - 1\right]. \tag{1.20}$$

As with the GEV, the parameters of the GPD, σ and ξ can be replaced with their maximum likelihood estimates or draws from the marginal posteriors in a Bayesian framework (see section 1.4) to obtain the posterior distribution for z_r. Again, estimation uncertainty can be accounted for via the delta method[1] within a frequentist setting, or by direct reference to the posterior standard deviation for z_r within the Bayesian framework.

1.4 Bayesian Inference

1.4.1 General Theory

Section 1.3 saw use look at the general theory of how to model extremes. The analyses in this report will focus on using a Bayesian framework, which is often preferable due to the inclusion of extra information, the potential of using predictive distributions, and more intuitive interpretation of the credible intervals.

We assume we have the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ which are realizations of a random variable from family $\mathcal{F} = \{f(\boldsymbol{x}; \theta) : \theta \in \Theta\}$. We can form the likelihood function: $f(\mathbf{x}|\theta)$, which is a function of θ for fixed $\mathbf{x}$, and if the x_i are independent then

$$f(\mathbf{x}|\theta) = \prod_{i-1}^{n} f(x_i; \theta).$$

We now believe that we can formulate beliefs about how likely θ is, without reference to the data, in a probability distribution. Such a distribution is called a prior distribution, which we denote $\pi(\theta)$, which is quite different to a frequentist view in which the parameter θ was thought of as an unknown constant. The way we choose our prior distribution depends entirely on how much information we already have about the parameter θ. For example, if we believe θ is a probability, and thus should take any value between $0 \leq \theta \leq 1$ but that these values are equally likely, then we could express our beliefs in the form $U \sim (0, 1)$. Whereas if θ is expected to be small in magnitude and real-valued, then a $\theta \sim N(0, 100)$ may be more appropriate[1].

Theorem 2: *Bayes' Theorem*

Bayes' Theorem[1] states:

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta)f(\mathbf{x}|\theta)}{\int_\Theta \pi(\theta)f(\mathbf{x}|\theta)d\theta}. \tag{1.21}$$

Bayes' Theorem provides us with a way of converting some initial beliefs and data we have observed into a posterior distribution. Thus we now have a complete distribution and our accuracy of the inference can be summarized by the variance of this posterior distribution, without using asymptotic likelihood theory. Those who support working in the Bayesian framework believe this supplementary information provided by the prior distribution is valuable and helps when little information is available. However those against the Bayesian view believe that it is very subjective as priors would be specified differently by different individuals. The main problem with this Bayesian framework is that the computation of Bayes' theorem, requires that of a difficult integral, see the denominator of equation 1.21. For some choices of prior distributions this problem is overlooked, and we do not have to compute the normalising integral that is we can simply compute posterior $\propto$ prior $\times$ likelihood. This happens when we choose a prior conjugate to the likelihood - that is that these prior distributions lead to posterior distributions from the same family. However conjugate priors do not necessarily adequately represent the prior beliefs that you have. Up until recently, for higher dimensions of θ, computing the integral in equation 1.21 was very difficult, however with the use of the Markov Chain Monte Carlo (MCMC) technique, has simplified the idea, and thus Bayesian techniques are now popular.

1.4.2 Markov Chain Monte Carlo (MCMC)

The idea of MCMC is simple, we want to produce simulated values from the posterior distribution. If this was possible to do exactly, we would expect the simulated mean to be the posterior mean, and the histogram of the simulated data to be the posterior density. The MCMC technique enables us to simulate values $\theta_1, \theta_2, \ldots$ from a distribution resembling the posterior distribution using the Metropolis Hastings algorithm. The Metropolis-Hastings scheme requires us to have $\pi(\theta)$ being the density of interest, and we also need a proposal distribution, which is easy to simulate from, with density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$. Examples include $(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \sim N(\boldsymbol{\theta}, 1)$ or $(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \sim \text{Gamma}(1, 1)$. Basically this distribution gives us a way of proposing new values $\boldsymbol{\theta}^*$ from the current value $\boldsymbol{\theta}$. Note it is not required that $\pi(\theta)$ is the stationary distribution of $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$. The Metropolis Hastings algorithm[23] used to implement this scheme can be seen below.

The Metropolis Hastings Algorithm

1. initialize the iteration counter to $j = 1$, and initialize the chain to $\boldsymbol{\theta}^{(0)}$.
2. Generate a proposed value $\boldsymbol{\theta}^*$ using the proposed distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})$.
   This procedure generates a first-order Markov chain, but the evolution of the $\boldsymbol{\theta}^{(0)}$ depends on q rather than the target density in equation 1.21.
3. Evaluate the acceptance probability $\alpha(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^*)$ of the proposed move, where

$$\alpha(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|x)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}|x)q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\}$$

4. Set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\theta}^*)$, and set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ otherwise.
   In other words, we either accept the proposed value, depending on the acceptance probability (which depends on the relationship between the density of interest and the proposal distribution), and then the chain moves, or reject it, again depending on the acceptance probability - resulting in the chain staying where it is.

5. Change the counter from j to $j+1$ and return to step 2.

So under simple regularity conditions, the generated sequence is a Markov chain, with its stationary distribution as the target distribution as in equation 1.21. For large i, $\theta_{i+1}, \theta_{i+2}, \ldots$ is approximately stationary, and with marginal distribution given by equation 1.21, with $\theta_1, \theta_2, \ldots, \theta_i$ defined as the burn in period, which we remove. Choosing the proposal distribution q can be difficult and two commonly used proposals are symmetric chain proposals and random walk proposal. This study will use the random walk proposal.

Random Walk Proposal

To use the random walk proposal[23] we must first consider $\boldsymbol{\theta}^*$ - the proposed value, at stage j to be:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(j-1)} + \mathbf{w}_j \tag{1.22}$$

where we define $\mathbf{w}_j$ to be independent and identically distributed random $p \times 1$ vectors, i.e. completely independent at the start of the chain. We say that the $\mathbf{w}_j$ have a distribution we can simulate from, with a mean $\mathbf{0}$, and we require the distribution to be symmetric about its mean. We can then simulate an innovation $\mathbf{w}_j$, from $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(j-1)} + \mathbf{w}_j$, taking this to be the proposed value. We let $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = f(|\boldsymbol{\theta}^* - \boldsymbol{\theta}|)$, note this is always symmetric, and $f(.)$ is an arbitrary density which can be used to calculate the acceptance probability. The problem now is to find the distribution for the innovation distribution $f(.)$ and find it's variance. We often use the Uniform or the Normal distribution for $f(.)$, noting that the Normal distribution is often better, but has a higher computational cost. We choose the variance to give us an optimal acceptance rate of about 0.234, however between $20 - 30\%$[24] is usually regarded as okay, thus the proportion of moves accepted relies on the variance of the distribution. If the variance is too low, there is a high acceptance rate and many small steps will be made. If the variance is too high, there is a low acceptance rate, and few large steps will be made, see figure 1.9 to show examples of different ranges of variance.

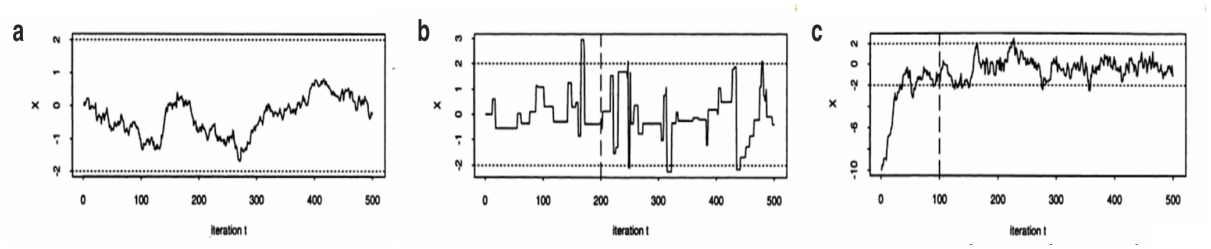

Figure 1.9: a) shows MCMC from a proposal with too low variance, b) MCMC from proposal with too large variance c) Correctly converging proposal, with burn-in 100 iterations[22]

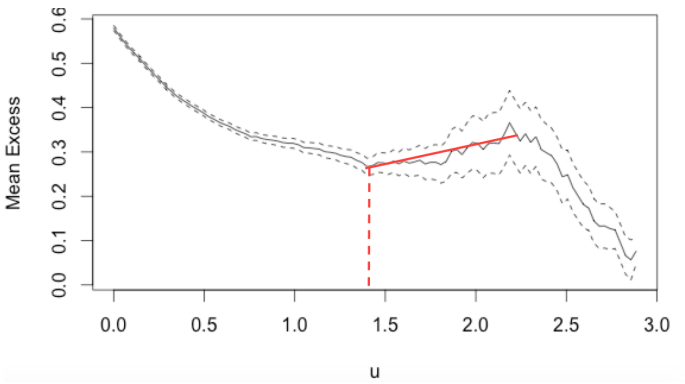

Figure 1.10: MRL plot for Sabine Pass, with associated 95% Confidence Intervals

The sequence we obtain using this random walk proposal can then be used to estimate the posterior mean and uncertainty about the parameter of interest. MCMC gives use the opportunity to overcome the problems of the integration in equation 1.21 and use Bayesian techniques to model extremes

1.5 Illustrative Application: Sabine Pass

We will now look at a more in depth analysis, which will lead on to the estimation of the return levels. Due to the downfalls of the block maxima approach, we will use the method of threshold excesses, as detailed in section 1.3.2. We must first choose an appropriate threshold, so we must look at a mean residual life (MRL) plot. The plot can be seen in figure 1.10, along with the associated 95% confidence intervals too. You can clearly see with help from the overlaid line, that this the MRL plot becomes linear at approximately 1.4ft thus after this value the mean excess (the expectation of the GPD) will be a linear function of the threshold, (see section 1.3.2: How do we do this in practice? for more details), and hence we choose 1.4ft to be our threshold. From this method, the percentage of observations kept is 2.73%, just over 800 values, which is far greater than just the five observations that we would have conserved had we been using the block maxima approach with annual blocks. We are now at a stage to estimate the values of σ and ξ at a threshold of 1.4ft using an MCMC algorithm with $10,000$ iterations. To run the algorithm in a Bayesian framework and in accordance with section 1.4 we need appropriate priors for σ and ξ. Since σ is a scale parameter, we re-parameterize as $\log(\sigma)$ to retain the positivity of the scale parameter in our MCMC. We choose our priors to be:

$$\log(\sigma) \sim N(0, 100) \tag{1.23}$$

$$\xi \sim N(0, 10) \tag{1.24}$$

Little is known about the two parameters hence we let them be Normally distributed, each with a large variance. The variance for ξ is smaller that that of σ because we rarely see large deviations from 0 for the ξ parameter. As detailed in section 1.4.2 we use a Metropolis Hastings scheme, with random walk updates, since conjugate priors do not need to be known to progress with this approach. The algorithms were tuned, to get acceptance probabilities within the range $20-30\%$, to obtain the optimal convergence rate as discussed in section 1.4.2. Multiple starting points were used to check for convergence, however for ease of clarity, figure 1.11 shows only two. Clearly, from the two starting

points for both σ and ξ we get convergence to the same value. The plots show means and confidence intervals for the two parameters, so we can say that the approximate values for σ and ξ are 0.25 and 0.05 respectively.

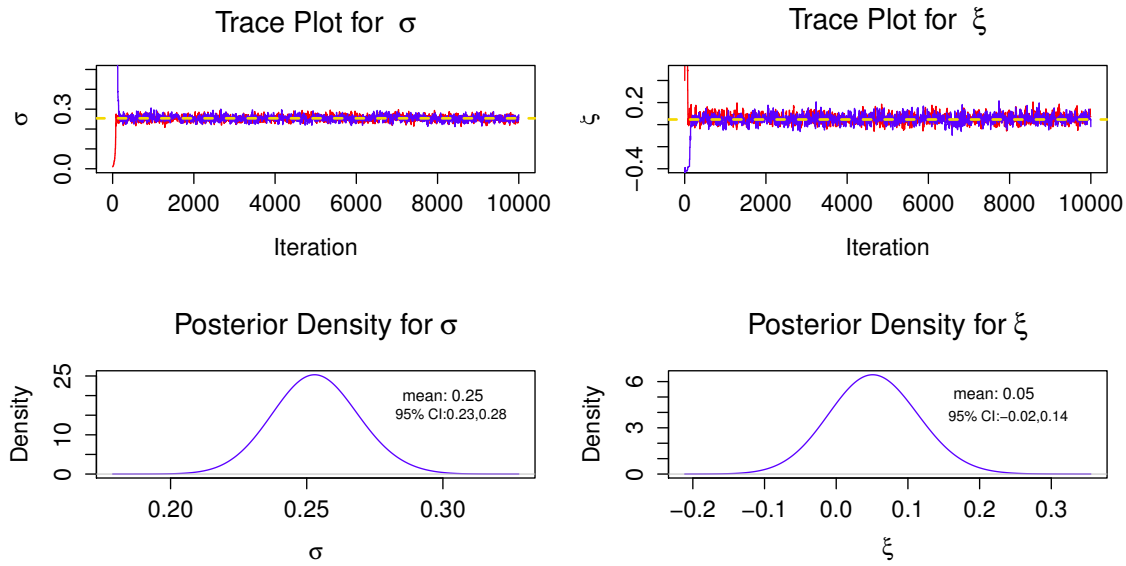

Figure 1.11: Plots showing MCMC convergence and posterior densities for σ and ξ

As discussed in section 1.3.1 it is important to perform model adequacy checks on your data to check whether the model that has been fitted has correctly encompassed your beliefs. Figure 1.12 shows a probability plot and a quantile plot (see section 1.3.1 for details), which have been constructed with reference to the posterior means. We can see that both plots fit well to the unit diagonal and so the GPD can be considered an adequate model.

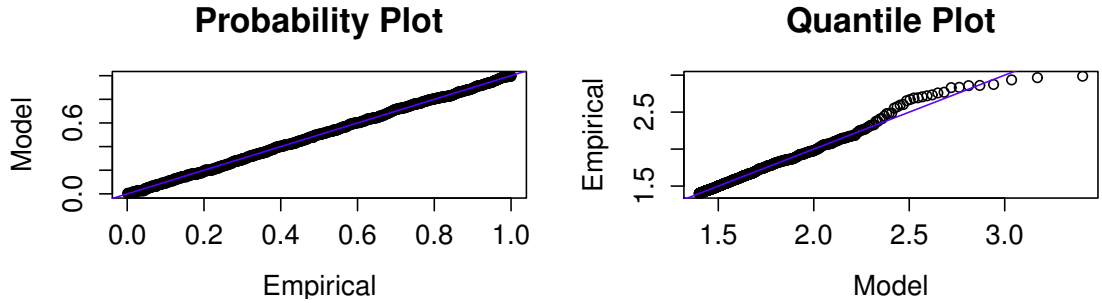

Figure 1.12: Probability & Quantile Plots for Sabine Pass

After fitting a model we will now obtain trace plots and densities for the 100 and 1,000 year return levels, given in figure 1.13. These return levels were estimated as z_{100} : 4.82 $(3.90, 6.46)$ and z_{1000} : 5.95 $(4.43, 8.93)$, with confidence intervals in parenthesis. Basically for z_{100} this means that in 100 years the counterpart of the sea-surge that we have been calculating would be expected to be around 4.82ft, but could take values from 3.90ft to 6.46ft approximately 95% of the time. Since the x-axis scales for the two density plots are the same you can also visually see that confidence interval for the z_{1000} return level is wider. This is to be expected, as there would be a more uncertainty surrounding our estimates if we were calculating for 1,000 years into the future. Interestingly, the highest

sea surge observed in the five years of data collected was 2.986ft, and so when we extrapolate past our time period we expect the highest sea-surge counterpart to increase almost 2ft in 100 years.

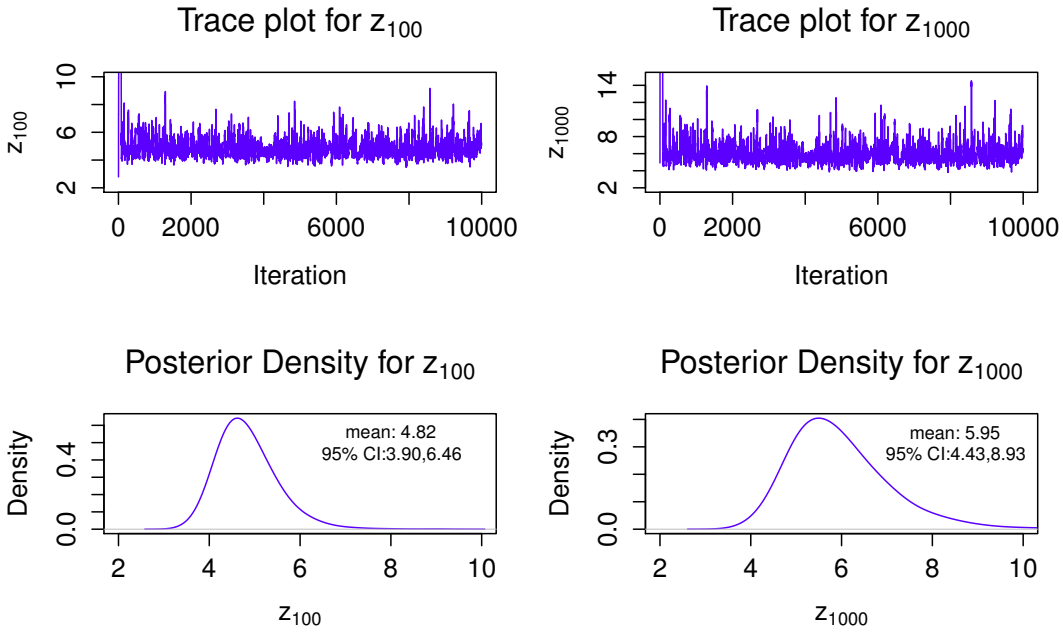

Figure 1.13: Return Level Plots for z_{100} and z_{1000}

Table 1.1 gives the estimates in both Frequentist and Bayesian frameworks, of the GPD scale and shape parameters, various return levels and their associated 95% confidence intervals, for the five different sites. The 95% Frequentist confidence intervals were constructed using a technique called profile likelihood, as calculating confidence in the standard way of estimate $\pm 1.96 \times$ s.e. often fails due to large standard error and potential non-normality for z_r, as a result of severe asymmetry in the return level likelihood. As this report will focus on using a Bayesian framework to make analyses, and this method is frequentist, we will not explain it here, but for details see Chapter 2.6.6 Coles, (2001)[1].

Table 1.1 has some interesting features. Out of all the sites Port Fourchon appears to have the highest return level estimates, in both the Bayesian and Frequentist framework, with the $1,000$ year return level reaching 62.80 and 37.73 in the Bayesian and Frequentist frameworks respectively. However it must be noted that the measure of variability in each framework is largest at Port Fourchon, implying that we are less confident about our estimates for this site. Another interesting site to look at is Berwick, where in both frameworks, the return level estimate is relatively unchanged as we increase r. It has previously been discussed in section 1.2 that Berwick was in an inland, marsh area, and that these types of areas can reduce the height of sea-surges. This appears to be what is happening at this site.

We can also look at comparing the Bayesian and Frequentist views on calculating parameter and return level estimates. For each parameter estimated there is barely any difference in the two views, and although the Frequentist estimates do appear to be lower, we note that there is an overlap in the 95% confidence interval, which implies there is no significant difference in the two approaches for this data. However saying this,

this report will still focus on using a Bayesian framework, as there are numerous reasons why this framework is better, which we now discuss.

Parameters	Bayesian Posterior			Frequentist		
	Mean	Standard Deviation	95% Credible Interval	MLE	Standard Error	95% Confidence Interval
σ_{SP}	0.25	0.01	(0.23,0.28)	0.25	0.01	(0.23,0.28)
σ_{G}	0.37	0.03	(0.31,0.44)	0.37	0.03	(0.31,0.44)
σ_{GI}	0.22	0.01	(0.20,0.25)	0.24	0.01	(0.21,0.26)
σ_{PF}	0.22	0.02	(0.18,0.27)	0.22	0.02	(0.18,0.27)
σ_{B}	1.48	0.04	(1.40,1.58)	1.48	0.05	(1.39,1.57)
ξ_{SP}	0.05	0.04	(–0.03,0.13)	0.05	0.04	(–0.03,0.12)
ξ_{G}	–0.08	0.06	(–0.19,0.06)	–0.10	0.06	(–0.21,0.01)
ξ_{GI}	0.22	0.05	(0.14,0.31)	0.22	0.05	(0.13,0.30)
ξ_{PF}	0.39	0.13	(0.23,0.59)	0.37	0.09	(0.20,0.54)
ξ_{B}	–0.50	0.02	(–0.55,–0.45)	–0.50	0.03	(–0.55,–0.45)
$z_{10,SP}$ (ft)	3.87	0.32	(3.34,4.64)	3.78	0.30	(3.33,4.58)
$z_{10,G}$ (ft)	3.78	0.28	(3.43,4.38)	3.68	0.21	(3.41,4.35)
$z_{10,GI}$ (ft)	5.82	0.86	(4.46,7.75)	5.54	0.90	(4.34,7.80)
$z_{10,PF}$ (ft)	8.60	2.81	(5.05,16.26)	7.51	2.07	(4.93,15.01)
$z_{10,B}$ (ft)	4.23	0.07	(4.12,4.37)	4.23	0.07	(4.12,4.38)
$z_{100,SP}$ (ft)	4.86	0.63	(3.90,6.43)	4.67	0.56	(3.83,6.27)
$z_{100,G}$ (ft)	4.35	0.54	(3.70,5.70)	4.10	0.36	(3.65,5.30)
$z_{100,GI}$ (ft)	9.75	2.31	(6.37,15.20)	9.09	2.29	(6.22,15.5)
$z_{100,PF}$ (ft)	22.18	13.26	(8.59,61.00)	16.54	7.75	(8.04,51.58)
$z_{100,B}$ (ft)	4.25	0.07	(4.13,4.41)	4.25	0.07	(4.13,4.42)
$z_{200,SP}$ (ft)	5.18	0.76	(4.04,7.09)	4.95	0.67	(3.99,6.88)
$z_{200,G}$ (ft)	4.41	0.64	(3.74,5.77)	4.21	0.40	(3.70,5.60)
$z_{200,GI}$ (ft)	11.50	3.04	(7.22,18.60)	10.55	2.96	(6.85,19.29)
$z_{200,PF}$ (ft)	29.53	20.87	(9.84,87.13)	21.15	11.18	(9.27,75.18)
$z_{200,B}$ (ft)	4.25	0.07	(4.13,4.40)	4.25	0.07	(4.13,4.42)
$z_{1000,SP}$ (ft)	6.01	1.12	(4.42,8.87)	5.65	0.95	(4.32,8.51)
$z_{1000,G}$ (ft)	4.84	0.90	(3.87,7.11)	4.43	0.52	(3.81,6.35)
$z_{1000,GI}$ (ft)	16.61	5.62	(9.10,30.40)	14.91	5.21	(8.72,31.5)
$z_{1000,PF}$ (ft)	62.80	59.52	(14.50,229.88)	37.73	25.31	(13.50,185.00)
$z_{1000,B}$ (ft)	4.26	0.07	(4.13,4.42)	4.25	0.07	(4.14,4.43)

Table 1.1: Frequentist and Bayesian estimates of the GPD scale and shape parameters, various return levels and their associated 95% confidence intervals, for the five different sites.

Why use Bayesian Inference?

Bayesian inference is considered the preferred analysis for several reasons, being a natural progression from the Frequentist view. One of these is that due to the very nature of extremes, we have few data points, and so any way in which we can incorporate another source of information - in this case through the use of a prior distribution, is desirable.

We can often use this prior to incorporate an experts beliefs, who may have studied this subject for many years, and by incorporating their opinion we can obtain a much more precise estimate for the return levels. Also the 95% confidence intervals have a more practical interpretation within the Bayesian scheme, in the way that they contain the true parameter with 95% probability, as the parameter is a random variable. However in the Frequentist view, the probability of lying within the 95% confidence interval is either 0 or 1 as the parameter is a fixed, but unknown constant, which on 95% of occasions will lie in the interval. Clearly it is much more intuitive to think of the probability of lying in the interval is 95%, thus giving credit to Bayesian statistics. Also, Bayesian inference is not dependent on regularity assumptions, for example the asymptotic theory of the maximum likelihood. In fact when $\xi < -0.5$ maximum likelihood breaks down[1], and we cannot use this, however within the Bayesian framework this does not happen, thus it gives us a way to provide estimates where maximum likelihood would fail. A final reason discussed here is that we can perform a more complete inference, the result of our work is to find out the probability of future events, and in a Bayesian framework we can do this via predictive distributions, see section 2.4 for an explanation of the theory behind this. It is because of these reasons that this report will not carry on the analysis using the Frequentist view, and will use a Bayesian framework from now on.

Chapter 2

Serial Correlation

2.1 Exploratory Analysis

This report will account for two types of dependence, the first being dependence between successive observations, (serial correlation), and the second being between-site correlation. This chapter will focus on the first of these types of dependence, and chapter 3 will focus on the latter. This is a natural place to start as often data is only available for one site, and in this case serial correlation can have a large effect on return level estimates, as we will see as this chapter progresses.

Recall the method of threshold excesses, see section 1.3.2, where we include all values above some pre-determined threshold u_0. The inclusion of more data compared with the block maxima approach should lead to reduced standard errors/posterior standard deviations for return levels, however this method brings about its own problems in the area of temporal dependence. We have hourly observations that are often dependent on each other from one hour to the next. The values above a threshold likely include all values from the same storm and thus they will depend on each other, breaking one of our modelling assumptions that our series of extremes is independent. Hence, we must find methods of accounting for this dependence, as the analyses conducted in chapter 1 were over-optimistic and thus led to too small standard errors/posterior standard deviations attached to parameter/return level estimates, as they assumed we had more independent observations than we actually do have in practice.

We start by looking at whether there actually is serial correlation at each of our sites, but we will focus on extreme sea-surges at Sabine Pass for illustrative purposes, although similar findings were obtained at the other sites. We first look at data from the partial autocorrelation function of the series, as this can give us an idea of the serial correlation within the data. There is a significant autocorrelation between observations one (lag 1) hour apart, with a value of 0.961; however this doesn't focus on the extremes. We can look at a plot of each observation against the preceding observation, see figure 2.1, from which we can see a very strong dependence between successive observations. This figure also allows us to look at the correlation in the extremes: by superimposing the threshold in red, which we calculate using an MRL plot, as we did in section 1.5. We can see there is still a very strong dependence above this threshold. Figure 2.2 shows a small section of the sea-surge data (50 observations). We can clearly see that the threshold is breached by successive observations; in fact, considering the series as a whole, we can see that the extreme sea-surges occur in clusters, which follows what we believe, as a storm is likely to last for several hours. We need a method to overcome this dependence within the clusters.

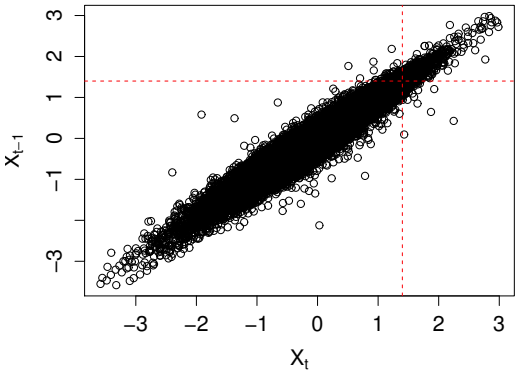

Figure 2.1: Plot of the hourly sea-surges observations against preceding observations, with superimposed threshold in red, at Sabine Pass,TX

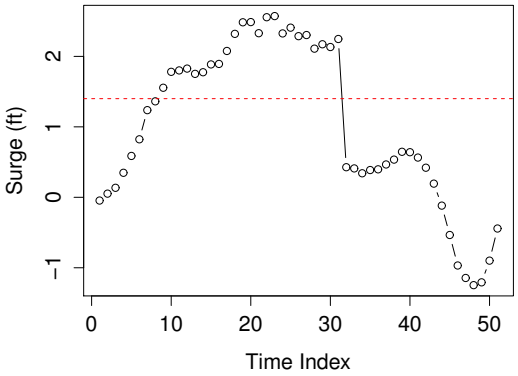

Figure 2.2: Time series plot of a small section of the sea-surge data for Sabine Pass, TX with threshold

2.2 Temporal Filtering: Runs Declustering

One technique we use to account for extremal dependence, seen in figures 2.1 and 2.2, is to filter out an independent set of threshold exceedances - a process commonly referred to as declustering (see, for example, Coles (2001)[1]). The analysis presented here uses MCMC sampling as in section 1.4 and the same priors as in equations 1.23 and 1.24. Declustering is the most widely used approach in practice, but current research[25] shows other approaches can be superior, for example explicitly modelling the temporal dependence in the process, which we will see in section 2.3. The technique of declustering, particularly runs declustering (the most commonly used method), requires us to choose a cluster termination interval, κ, often arbitrarily. Occasionally, some physical knowledge of the process being studied (e.g. knowledge of tides) might indicate an "optimal" choice of κ; see, for example, Coles and Tawn (1991)[26]. We then say that a cluster of threshold excesses has terminated as soon as at least κ consecutive observations fall below the threshold. Finally, to account for the dependence, we extract the peak of each of the clusters obtained, and fit the GPD to the set of cluster peak excesses. We must note that the GPD parameter estimations (σ, ξ) are sensitive to the choice of κ. A κ value which is too large will leave too few cluster exceedances on which to complete an analysis and if κ is too small we may not be able to assume independence, as our cluster peaks may be too close. As well as this, return level estimates can also be sensitive to the choice of κ, which is shown in Fawcett and Walshaw (2012)[27]. We will now look at the sensitivity in return level estimates using declustering.

Figure 2.3 shows 10, 100, 200 and 1000-year return level estimates for a range of values for κ, for Sabine Pass. Although the 95% credible intervals overlap, and thus we don't have significantly different estimates for κ, (also seen in Fawcett and Walshaw (2015) [28]), in practice, practitioners often work to the upper end-point of a 95% confidence interval as this can be interpreted as the most plausible "worst case scenario". Consider the declustering for the 10-year return level using the data at Sabine Pass (top left of figure 2.3). If we choose $\kappa = 5$, then the engineers would build this component of the levee to just above 4ft, however if we choose $\kappa = 18$, the engineers could be building the

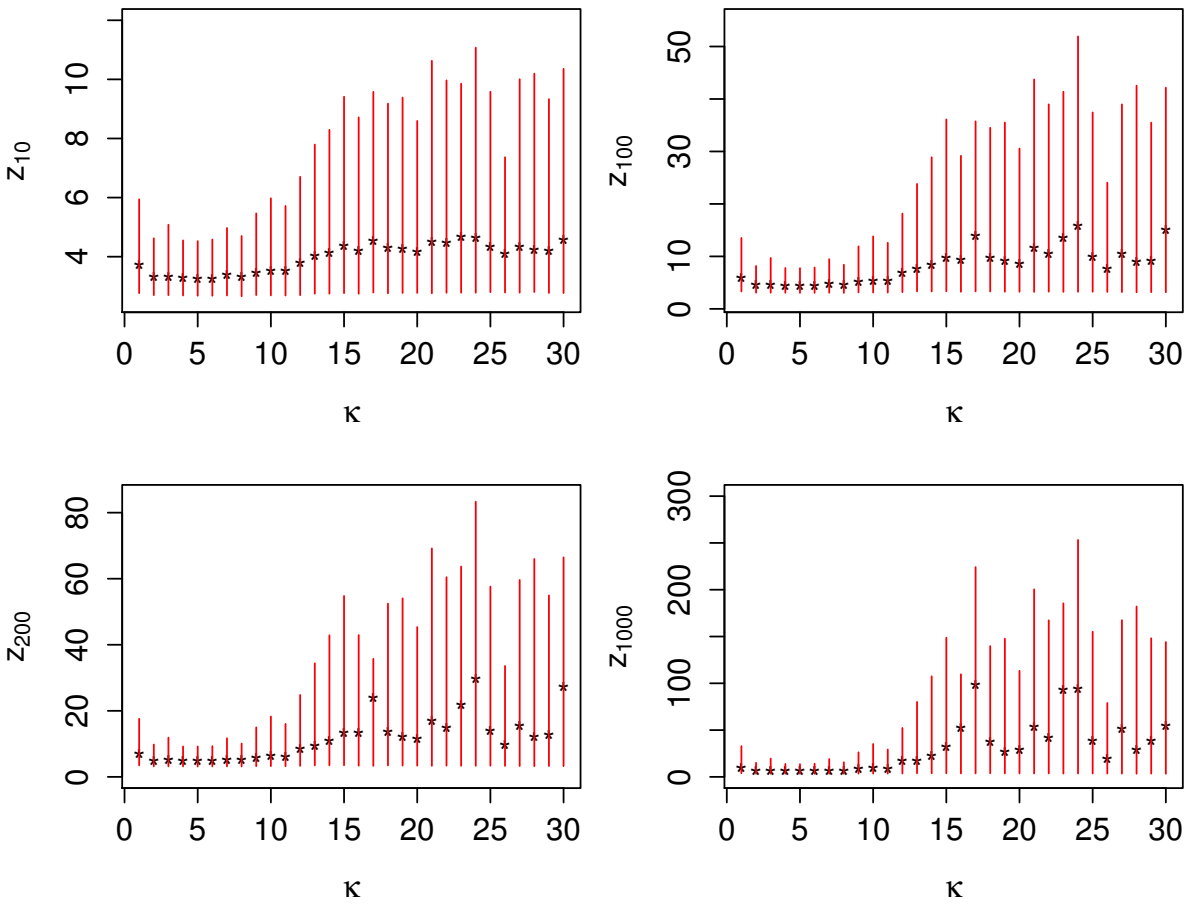

Figure 2.3: Plots for $10, 100, 200, 1000-$year return levels where * indicates the return level posterior mean and the 95% credible intervals are superimposed in red, for a range of κ values

component to above 10ft. Clearly, this is evidence that return level estimates and their 95% credible intervals are sensitive to the choice of κ. In fact, the estimates generally increase in value, and credible intervals increase in width as κ increases, which results in less precision about the return level estimate, thus discrediting the declustering technique. Parameter estimate sensitivity to the choice of κ is not the only downfall of this method, another is that the peak from a certain cluster may not be as extreme as some of the other observations we throw away from another cluster. Hence because of these two downfalls we must look into better, more reliable analyses.

2.3 Accounting for Dependence: the Extremal Index

2.3.1 Extremal Index

We now look at a second, more useful method for accounting for dependence in the extremes of our process; one which allows the inclusion of all extremes, not just a filtered set of independent extremes. First, we need Leadbetter's $D(u_n)$ condition, which ensures that long-range dependence is sufficiently weak and hence it does not affect the asymptotics of an extreme value analysis[21]. Now in more formal terms, Leadbetter's $D(u_n)$

condition[1] states that:
A stationary series $X_1, X_2, ...$ is said to satisfy the $D(u_n)$ condition if, for all $i_1 < ... < i_p < j_1 < ...j_q$ with $j_1 - i_p > l$,

$$\begin{aligned} &|\Pr\{X_{i1} \leq u_n, ..., X_{ip} \leq u_n, X_{j1} \leq u_n, ..., X_{jq} \leq u_n\} \\ &- \Pr\{X_{i1} \leq u_n, ..., X_{ip} \leq u_n\}\Pr\{X_{j1} \leq u_n, ..., X_{jq} \leq u_n\}| \ \leq \ \alpha(n, l) \end{aligned} \tag{2.1}$$

Basically this means that if you have a sequence of independent variables then the difference in probabilities in equation 2.1 will be zero for any sequence u_n. For a sequence with threshold u_0 we require the $D(u_n)$ condition to hold, and thus the condition ensures that, for sets of variables that are far enough apart (i.e. a large sea surge in January does no affect a large sea surge in June), the difference in probabilities in equation 2.1 has no effect on limits laws for extremes, as it is sufficiently close to zero.

We now look at a theorem - *Extremes of Dependent sequences*[21]. This theorem enables us to account for dependence using the extremal index, often denoted by θ.

Theorem 3: *Extremes of Dependent Sequences*

Let $\tilde{X}_1, \tilde{X}_2, ...$ be a stationary series which satisfies Leadbetter's $D(u_n)$ condition from equation 2.1, and we also let $\tilde{M}_n = \max\{\tilde{X}_1, \ldots, \tilde{X}_n\}$. We also have $X_1, X_2, \ldots$ which is an independent series with X having the same distribution as $\tilde{X}$ and $M_n = \max\{X_1, \ldots, X_n\}$. Then since

$$\Pr\{(M_n - b_n)/a_n \leq x\} \to G(x),$$

which is a non-degenerate limit for M_n, then under certain regularity conditions,

$$\Pr\{(\tilde{M}_n - b_n)/a_n \leq x\} \to G^\theta(x). \tag{2.2}$$

As the extremal index, θ, is a measure of dependence in the extremes, we say that if $\theta = 1$ then the extremes of a process are completely independent, and if $\theta \to 0$ the extremes of the process become increasingly dependent. This leads to the conclusion that if the maxima of a stationary series converge in distribution (which we know they do, to the GEV, from section 1.3), and Leadbetter's $D(u_n)$ condition holds (i.e. long-range dependence is negligible), then the limit distribution is related to that of the independent series, in fact it is $G^\theta(x)$. This is in fact another GEV distribution function, with a different location and scale to the independent series, as we absorb θ into the new location and scale.

Unfortunately, when considering the tail of our process according to threshold excesses, the extremal index cannot be absorbed into the parameters of the GPD after powering by θ; thus, as used by Fawcett and Walshaw (2012)[27], we have

$$H^\theta(y) = \left\{1 - \lambda_u \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}\right\}^\theta. \tag{2.3}$$

Inversion of this expression then gives an expression for the r-year return level, which accounts for extremal dependence, see equation 2.5.

2.3.2 Intervals Estimator

Section 2.1 showed strong serial correlation at Sabine Sabine Pass; similar dependencies were observed at the other sites. We now discuss various ways in which we can calculate θ, as detailed in Fawcett and Walshaw (2012)[27], but will settle on the method they proposed, due to it's ease-of-use and the fact that it is unbiased. The first method they propose fits an extreme value Markov chain model to successive pairs of extremes in the series, however this is quite a subjective approach, and requires the Markov model used to be suitable for this method to work. The second method suggests that the extremal index can be found through methods which identify the cluster of extremes, with the estimate being the reciprocal of the mean cluster size. Again this is subjective, this time to the choice of κ, and as parameter estimates are sensitive to this choice, this might also not be a suitable method. A third method estimates θ as the reciprocal of the mean cluster, where clusters are identified by splitting the data into l blocks of length τ and the threshold exceedances in each block are treated as a single cluster. Again this is subjective, this time to the choice of τ, and there are problems with choosing block length. The method that this report will use to estimate θ is the intervals estimator, shown in Fawcett and Walshaw (2012)[27] to be one of the best estimators when comparing it's estimates of true values from simulated data, and also it relies on no assumptions regarding the form of the extremal dependence structure. This method looks at estimating θ by looking at the inter-arrival times of the threshold exceedances, as proposed by Ferro & Segers (2003)[29]. We have $T_i = S_{i+1} - S_i$ for $i = 1, ..., K-1$, where T_i are the inter-arrival times and K the exceedance times observed: $S_1 < S_2 < ... < S_K$. The distribution of T_i was derived by Ferro and Segers, giving a bias-corrected moments-based estimator for θ as:

$$\hat{\theta} = \min\left(1, \frac{2\left\{\sum_{i=1}^{K-1}(T_i - a)\right\}^2}{(K-1)\sum_{i=1}^{K-1}(T_i - b)(T_i - c)}\right) \tag{2.4}$$

where $a = b = c = 0$ if the largest inter-arrival time is no greater than 2, and $a = b = 1$ and $c = 2$ if the largest inter-arrival time is greater than 2. The values of θ that were obtained from this method can be seen in table 2.1. You can clearly see that for all sites there is a strong dependence between successive observations, with all θ estimates being close to 0.

Parameter	Estimate
θ_{SP}	0.198
θ_G	0.162
θ_{GI}	0.166
θ_{PF}	0.181
θ_B	0.014

Table 2.1: Point estimates of θ for all 5 sites

It is important that we can incorporate the parameter θ into our return level estimate, so that we can effectively account for the dependence we have seen. We now have estimates of θ for each site, and so we must form an equation for the return levels, incorporating the dependence. We first let $y = z_r - u$ in equation 2.3, then set this equal to $1 - rn_y$ and rearrange to find z_r to obtain an equation for the r-year return level, in a similar fashion

to the independent case. We start with

$$\left(1-\lambda_u\left[1+\xi\left(\frac{z_r-u}{\sigma}\right)\right]^{-1/\xi}\right)^{\theta}=1-(rn_y)^{-1},$$

which we rearrange to obtain the r-year return level which correctly accounts for the dependence between successive observations:

$$z_r = u+\frac{\sigma}{\xi}\left[\left({\lambda_u}^{-1}\left\{1-[1-(rn_y)^{-1}]^{\theta^{-1}}\right\}\right)^{-\xi}-1\right]. \tag{2.5}$$

We now look at the return level estimates for all sites, using MCMC output for the GPD scale and shape parameters, to give posterior draws for the return levels, using equation 2.5. These can be seen in table 2.2. Comparing table 2.2 with table 1.1 we can clearly see that every posterior mean return level estimate is lower once we account for serial correlation. This implies that had we incorrectly assumed independence, we would have consistently been relaying overestimates to the engineers. This would most likely result in levees being built at a height that would unlikely be breached, which could be financially wasteful. We can also compare the posterior mean return level estimates and their 95% credible intervals for Sabine Pass from table 2.2 to the declustered return level estimates and confidence interval in figure 2.3. It was stated in Coles and Tawn (1991)[26] that a value of $\kappa = 30$ is sufficient to account for wave propagation time, and thus after

Parameter	Bayesian Posterior	
	Mean	95% Credible Interval
$z_{10,SP}$ (ft)	3.25	(2.97, 3.67)
$z_{10,G}$ (ft)	3.34	(3.14, 3.67)
$z_{10,GI}$ (ft)	3.86	(3.31, 4.70)
$z_{10,PF}$ (ft)	4.59	(3.48, 6.71)
$z_{10,B}$ (ft)	4.01	(3.95, 4.09)
$z_{100,SP}$ (ft)	4.14	(3.53, 5.11)
$z_{100,G}$ (ft)	3.89	(3.50, 4.65)
$z_{100,GI}$ (ft)	6.41	(4.78, 9.07)
$z_{100,PF}$ (ft)	10.51	(5.62, 22.71)
$z_{100,B}$ (ft)	4.18	(4.08, 4.32)
$z_{200,SP}$ (ft)	4.44	(3.69, 5.65)
$z_{200,G}$ (ft)	4.04	(3.58, 4.96)
$z_{200,GI}$ (ft)	7.48	(5.32, 11.12)
$z_{200,PF}$ (ft)	13.82	(6.52, 33.65)
$z_{200,B}$ (ft)	4.21	(4.10, 4.35)
$z_{1000,SP}$ (ft)	5.17	(4.05, 7.10)
$z_{1000,G}$ (ft)	4.38	(3.73, 5.73)
$z_{1000,GI}$ (ft)	10.75	(6.78, 17.99)
$z_{1000,PF}$ (ft)	27.00	(9.12, 84.02)
$z_{1000,B}$ (ft)	4.24	(4.12, 4.39)

Table 2.2: Posterior mean return levels estimates and 95% credible intervals for all five sites accounting for Dependence, using Ferro and Segers' estimator for θ

this value we can safely say we have independence. If we compare the return level estimate z_{10} for this value of κ, 4.55ft, to that using the extremal index, 3.25ft, we can see a significantly reduced estimate when using the extremal index. Perhaps more important, is the difference in the 95% credible intervals, which are $(2.77, 10.35)$ft and $(2.97, 3.67)$ft for the declustering and the extremal index method respectively. We have much smaller credible intervals when using the extremal index method, due to the fact we have incorporated more data, and thus we have reduced uncertainty; we are more precise with our estimates. This reduced uncertainty, along with the added benefit of not having to choose a value for κ gives weight to the extremal index method. However, this analysis is not completely Bayesian, because Ferro and Segers' intervals estimator is not likelihood-based but is moments-based. An analysis of the fully Bayesian approach, assuming a likelihood-based estimator for θ has been completed in section 2.3.3.

2.3.3 A Fully Bayesian Approach

The extremal index estimators assessed by Fawcett and Walshaw (2012)[27] failed to take into account more recent advances of Ferro and Segers' work. For example, Süveges and Davison (2010)[30] propose a modification to the distribution of threshold inter-arrival times, as discussed in Section 2.3.2, to give a bias-corrected maximum likelihood estimator for the extremal index. More recent work in Fawcett and Walshaw (2015)[28], including extensive simulation studies, suggests this as a real contender to the moments-based estimator of Ferro and Segers (2003)[29]; we exploit the fact that this new method is likelihood-based to propose a fully Bayesian analysis of our sea-surge extremes, using the same non-informative priors for the GPD scale and shape as used earlier in equations 1.23 and 1.24. In the absence of any expert prior knowledge regarding the extremal dependence present in our processes, we adopt Uniform priors over the range (0,1) for the extremal index at each site. In our MCMC scheme, the marginal parameters and the extremal index are updated independently; an assumption often made in such analyses, and verified on practical grounds in, for example, Fawcett and Walshaw (2015)[28] and Davison et al. (2012).[31]

Table 2.3 shows the estimations for the return level using the Bayesian Posterior means, along with their associated 95% confidence intervals. In this complete Bayesian approach, we have used a likelihood for θ and thus expect wider confidence intervals, due to their being more variability in our estimate. Comparing table 2.2 to 2.3 we can see that mostly all the confidence intervals do increase when we estimate using a fully Bayesian approach. However these increases are only marginal, which is to be expected when using only a vague prior. Figure 2.4 plots the posterior mean return level estimates for the moments-based estimator against those for the likelihood-based estimator, and can be used to determine if the estimates appear to differ from the two approaches. We observe that the there appears to be little difference in the two approaches, as they plot fairly well to a straight line, i.e. what we see with one estimator almost mirrors that for the other estimator. Even though the estimates are very similar, it is still important that we complete a fully Bayesian approach as we are now accounting for all the possible variability in the parameters.

Parameter	Bayesian Posterior	
	Mean	95% Credible Interval
$z_{10,SP}$ (ft)	3.26	(2.95,3.67)
$z_{10,G}$ (ft)	3.34	(3.12,3.70)
$z_{10,GI}$ (ft)	3.89	(3.29,4.77)
$z_{10,PF}$ (ft)	4.56	(3.44,6.62)
$z_{10,B}$ (ft)	3.51	(-4.74,4.21)
$z_{100,SP}$ (ft)	4.14	(3.51,5.10)
$z_{100,G}$ (ft)	3.92	(3.50,4.75)
$z_{100,GI}$ (ft)	6.47	(4.79,9.15)
$z_{100,PF}$ (ft)	10.46	(5.68, 22.17)
$z_{100,B}$ (ft)	3.64	(-4.73,4.34)
$z_{200,SP}$ (ft)	4.42	(3.67,5.60)
$z_{200,G}$ (ft)	4.09	(3.59,5.09)
$z_{200,GI}$ (ft)	7.56	(5.35,11.18)
$z_{200,PF}$ (ft)	13.76	(6.61, 30.96)
$z_{200,B}$ (ft)	3.67	(-4.74, 4.36)
$z_{1000,SP}$ (ft)	5.16	(4.05,6.96)
$z_{1000,G}$ (ft)	4.44	(3.75,5.96)
$z_{1000,GI}$ (ft)	10.88	(6.90,17.96)
$z_{1000,PF}$ (ft)	26.81	(9.48, 75.55)
$z_{1000,B}$ (ft)	3.69	(-4.74, 4.39)

Table 2.3: Posterior mean return levels estimates and 95% credible intervals for all five sites accounting for Dependence, using Süveges and Davison's estimator for θ

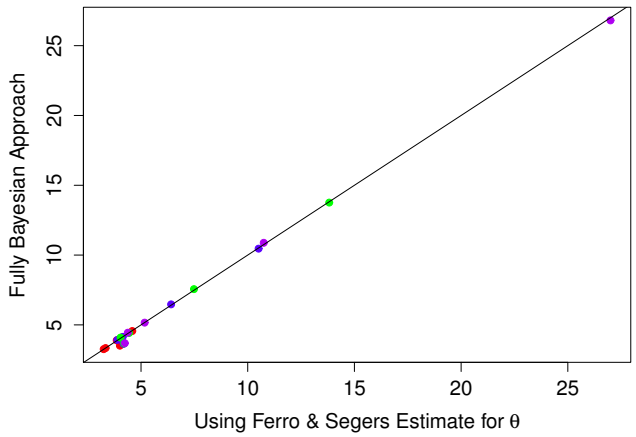

Figure 2.4: Plot using Ferro and Segers estimator for θ against the θ in the fully Bayesian approach. The 10, 100, 200 and 1000-year return levels are coloured in red, blue, green and purple respectively.

Figure 2.5 shows the return level estimates without accounting for serial correlation in blue and accounting for serial correlation, in a fully Bayesian approach through the extremal index, in red. The plots show up to the $10,000$-year return level, however it is conventional to put them on the scale $-\log(-\log(1-r^{-1}))$, as this emphasizes the tail of the distribution

for easy examination of model fit and extrapolation. We can see a marked reduction from the independence to the dependence case for all five sites, at all the year return levels, again giving credit to the method of accounting for dependence. This shows that even 10,000 years into the future we expect our return levels to be lower when we account for dependence, and as you can imagine, incorrectly assuming independence would lead to levees which would never be breached, which would be a financial waste.

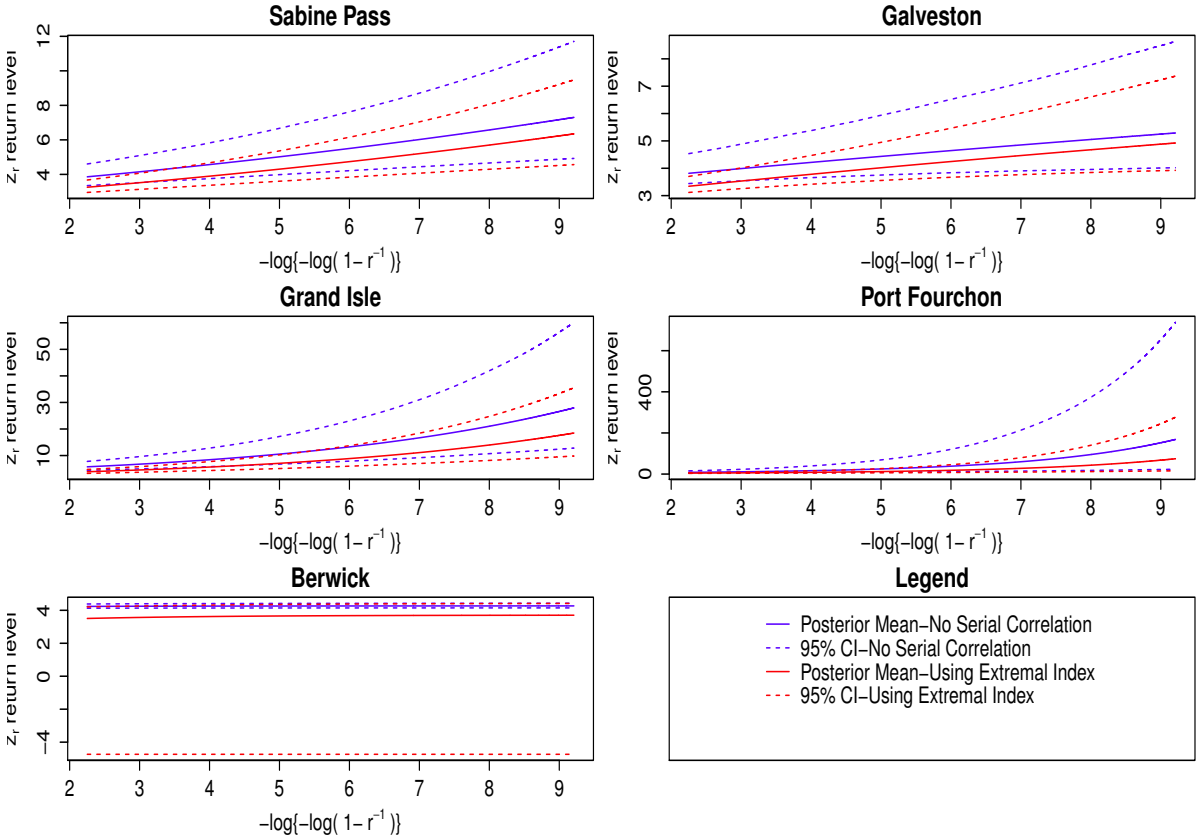

Figure 2.5: Return levels accounting for Independence and Dependence for 10,000 years, for all five sites

2.4 Predictive Return Level Inference

Up until now we have simply been making estimates about the parameters in order to make estimates for our return levels, however this section[1] will look at how we can predict the outcome if an event was to happen again. We are clearly uncertain about future outcomes, call them z, and so we model possible outcomes in a probability density function given by $f(z|\theta)$. We could assume that $\hat{\theta}$ is an estimate of θ, then make inferences on $f(z|\theta = \hat{\theta})$, however it is doubtful that $\theta = \hat{\theta}$. In a Bayesian framework we can use the predictive distribution, which has been touched upon previously. This distribution forecasts how likely values are in the future, using the predictive probability density function of z given $\boldsymbol{x}$:

$$f(z|\mathbf{x}) = \int_{\Theta} f(z|\theta)\pi(\theta|\mathbf{x})d\theta. \tag{2.6}$$

The predictive probability density function accounts for the variability in θ by weighting possible values of θ in the future experiment $f(z|\theta)$, by how likely we believe its values could occur, i.e. through the posterior distribution $\pi(\theta|\mathbf{x})$. Predictions from the predictive distribution are usually the best predictions we can achieve, as θ is generally unknown. If we knew θ to be some value, say $\theta = \theta_0$ then we could use $f(y|\theta = \theta_0)$ and this would obviously be best.

We use the predictive distribution in extreme value analyses as we often want to estimate the likelihood of reaching certain extreme levels. Say a suitable model for threshold excesses is $Z \sim GPD(\sigma, \xi)$, and we use the methods discussed in chapter 1 to obtain estimates for the parameter vector $\theta = (\sigma, \xi)$, based on the hourly observations seen $\mathbf{x} = x_1, \ldots, x_n$. Then if we use equation 2.6 we obtain:

$$f(Z \leq z|\mathbf{x}) = \int_{\Theta} f(Z \leq z|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}). \tag{2.7}$$

where this gives the distribution of future threshold excesses $f(Z \leq z|\mathbf{x})$, incorporating parameter uncertainty and variability in future observations. Now the key problem here is that we cannot implement equation 2.7 analytically, we need a distribution for the posterior, and we do not have this, we simply have our MCMC which should converge to the posterior distribution. However as seen by Coles (2001)[1] it is possible to approximate the left-hand-side of equation 2.7 via:

$$f(Z \leq z|\mathbf{x}) \approx \frac{1}{s}\sum_{i=1}^{s} Pr\{Z \leq z|\theta_i\}, \tag{2.8}$$

where s is the iterations left after removing the burn-in period. So if we set our approximation, as in equation 2.8 equal to $1 - 1/rn_y$, as we have done previously, we can obtain the analogue of the r-year return level. This method incorporates uncertainty due to model estimation, and is straightforward to implement using a numerical solver.

Although comparative results were found for all five sites, for illustrative purposes we will focus here on just Sabine Pass and Grand Isle. The predictive return levels for Sabine Pass were 3.27, 4.24, 4.57 and 5.49 and for Grand Isle were 3.91, 6.63, 7.82, and 11.60, for the 10, 100, 200 and 1,000-year return level respectively. If we compare these values to those obtained for the fully Bayesian approach in table 2.3 we can see that there is an increase in each prediction in comparison to the estimate. Some increases were larger than others, for example the 1,000 year return level for Grand Isle was 10.88ft in the fully Bayesian approach, but the predictive value was nearly a foot more. When using a predictive distribution we give a single number without confidence intervals to practitioners, which encompasses all of the variability, in the form of parameter uncertainty and randomness in future observations, thus leading to an increased value. Practitioners often like the use of the predictive return level estimate, as it is a single number the levee heights can be designed to, which is clearly much easier to work with. This credits the Bayesian way of thinking as this analysis is not possible to undertake in a Frequentist setting.

Chapter 3

Spatial Dependence

3.1 Motivation

Having accounted for serial correlation we now make a natural progression to considering extremal dependence between sites; in other words - *spatial* dependence. In this chapter we switch from looking at estimates of return levels to estimates of joint exceedance probabilities of marginal quantiles, focusing primarily on the bivariate case by examining pairs of sites. Work in Fawcett and Walshaw (2014)[32] shows the marginal serial correlation has no effect on the estimation of such joint exceedance probabilities, and so we focus solely on the effects of dependence between extremes occurring spatially. We must first look to see if there is any spatial dependence between our sites, however for illustrative purposes we will show a plot for Sabine Pass and Galveston only. There is strong extremal dependence between all pairs of sites, except for each site with Berwick. Although Berwick is geographically rather close to Grand Isle and Port Fourchon, it is an inland, swamp location and as such is subject to very different shifts in sea level, comparative to the other sites. As a result of this, there will be no continued analysis with Berwick in the chapter. Figure 3.1 shows a plot of sea-surges at Sabine Pass against those at Galveston; collected at the same times, numbered into four separate regions. Region 1 is defined as being above the thresholds for both sites, region 2 is above the threshold at Sabine Pass only, region 3 is above the threshold at Galveston only, and region 4 is below the threshold at both locations. From this plot it is obvious that there is very strong dependence between sea-surges at these sites, even in the extremes. This gives a clear motivation for considering bivariate extreme value theory as we will now see.

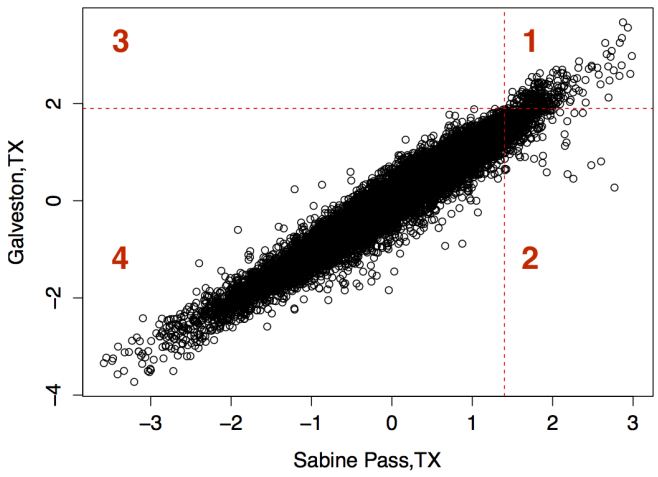

Figure 3.1: Plot of sea-surge extremes at Sabine Pass against those at Galveston showing numbered regions above and below the thresholds

3.2 Bivariate Extreme Value Theory

3.2.1 Componentwise Maxima

We now consider the theory of bivariate extremes[1]. Again we must first look at how we model this in a block maxima approach, as we did with the univariate case, because the theory in the bivariate case is based on limiting behaviour of the block maxima. Suppose we have observations at a pair of sites, $(X_1, Y_1), (X_2, Y_2), \ldots$, and in context with this report think of these as sea-surges at two different sites. However, we note that it is also possible to model extremes of two different processes at a single location, for example sea-surges against rainfall. For site X, we let

$$M_{x,n} = \max_{i=1,\ldots n} \{X_i\},$$

with an analogous definition for $M_{y,n}$ at site Y. We then let the componentwise maxima be defined as

$$\boldsymbol{M_n} = (M_{x,n}, M_{y,n}). \tag{3.1}$$

We aim to analyse $\boldsymbol{M_n}$ in equation 3.1, as $n \to \infty$ and thus when we consider them separately, we say that $\{X_i\}$ and $\{Y_i\}$ are independent univariate random variables, hence standard univariate extreme value theory applies to each marginal component. Assuming X_i and Y_i have standard Fréchet marginal distributions, with CDF:

$$F(z) = \exp(-1/z),$$

we can then obtain a simple normalization of the maxima giving:

$$\Pr(X_i < x) = \Pr(M_{x,n}/n \leq x) = \exp(-1/x), \tag{3.2}$$

again, with an analogous form for Y_i. If we now consider a rescaled vector, obtained using equation 3.2, we have

$$\begin{aligned} \boldsymbol{M_n^*} &= \left(\max_{i=1,\ldots n} \{X_i\}/n, \max_{i=1,\ldots n} \{Y_i\}/n \right) \\ &= (M_{x,n}^*, M_{y,n}^*). \end{aligned} \tag{3.3}$$

We now have unit Fréchet margins for all n, thus we can characterise the limiting behaviour of $\boldsymbol{M_n^*}$ without worrying about the marginals. Theorem 4[1] gives a characterization of the limiting distribution of $\boldsymbol{M_n^*}$, in essence is it a bivariate analog to theorem 1.

Theorem 4: *Limiting Distributions for Bivariate Extremes*

We have $\boldsymbol{M_n^*}$ as defined in equation 3.3, where (X_i, Y_i) are independent, with standard Fréchet marginal distributions. Then if

$$\Pr(M_{x,n}^*, M_{y,n}^*) \xrightarrow{d} G(x, y), \tag{3.4}$$

where G is a non-degenerate distribution function, then G has the form:

$$G(x, y) = \exp\{-V(x, y)\}, \;\; x > 0, \; y > 0, \tag{3.5}$$

where

$$V(x,y) = 2\int_0^1 \max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right) dH(\omega), \tag{3.6}$$

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

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

The family of distributions arising from equation 3.4 is known as the family of bivariate extreme value distributions.

If we now suppose that X and Y are GEV with parameters (μ_x, σ_x, ξ_x), with equivalent for Y then we can transform to obtain unit Fréchet margins via:

$$\tilde{x} = \left[1 + \xi_x\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^{1/\xi_x}, \tag{3.7}$$

again, with similar for Y, we have

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

which has the appropriate Fréchet margins to be valid for $V(.)$ and is a bivariate extreme value distribution. We will come back to valid functional forms for V shortly.

3.2.2 Bivariate Threshold Excesses

Now we have the theory for modelling bivariate extremes of componentwise maxima, we can extend this to look at how to model bivariate extremes under a threshold-based approach[1]. We define our bivariate excesses as those which exceed a threshold in either one of the margins. We have previously looked at how to model approximations to the tail of an arbitrary distribution, say F, and this had distribution function for the excesses (focusing on extremes for site X for now), to be:

$$H(x) = 1 - \lambda_{u_x}\left\{1 + \frac{\xi_x(x-u_x)}{\sigma_x}\right\}^{-1/\xi_x}, \tag{3.8}$$

using equations 1.9 and 1.19. We say F is the joint distribution function for independent realizations of a random variable (X,Y) and we aim to find a bivariate version of equation 3.8, that is a joint distribution $G(x,y)$ valid on $x > u_x$ and $y > u_y$. As we did with the GEV case in equation 3.7, we transform x and y to unit Fréchet. For x we obtain

$$\tilde{x} = -\left(\log\left\{1 - \lambda\left[1 + \frac{\xi(x-u_x)}{\tilde{\sigma}}\right]^{-1/\xi}\right\}\right)^{-1}, \tag{3.9}$$

with an analogous result for y. To get the equation for G we first need the homogeneity property of V, which states that $V(a^{-1}x, a^{-1}y) = aV(x,y)$. Substituting this into equation 3.5, we obtain $G^n(x,y) = G(n^{-1}x, n^{-1}y)$. Thus we can now find a distribution function for $\tilde{F}$ with variable $(\tilde{x},\tilde{y})$, we can say, for large n, by equation 3.5 we have:

$$\begin{aligned}\tilde{F}(\tilde{x},\tilde{y}) &= \{\tilde{F}^n(\tilde{x},\tilde{y})\}^{1/n} \\ &\approx [\exp\{-V(\tilde{x}/n, \tilde{y}/n)\}]^{1/n} \\ &= \exp\{-V(\tilde{x},\tilde{y})\}.\end{aligned}$$

Now since $F(x,y) = \tilde{F}(\tilde{x},\tilde{y})$ we can say that

$$\begin{aligned} F(x,y) &\approx G(x,y) \\ &= \exp\{-V(\tilde{x},\tilde{y})\} \;\; x > u_x, \; y > u_y, \end{aligned} \tag{3.10}$$

where $V(x,y)$ satisfies the mean constraint condition in equation 3.6, and $\tilde{x}$ and $\tilde{y}$ defined as in equation 3.9.

As with the univariate case, we need the likelihood function as an ingredient in Bayesian inference. To obtain the likelihood, we transform to unit Fréchet as just discussed, and then differentiate equation 3.10, to obtain $g(x,y)$. When working with Bivariate extremes, there is an added complication that a bivariate pair may exceed a threshold for just one of its components, i.e. regions 2 and 3 in figure 3.1, as well as exceeding both thresholds, (region 1), or neither thresholds (region 4). We obtain contributions to the likelihood from all four regions as below, noting that $\boldsymbol{\theta}$ denotes the dependence parameters in V:

$$g(x,y;\boldsymbol{\theta}) \;\; = \;\; \begin{cases} \left.\dfrac{\partial^2 G}{\partial x \partial y}\right|_{(\tilde{x},\tilde{y})} & \text{if } (x,y) \in \text{Region 1} \\ \left.\dfrac{\partial G}{\partial x}\right|_{(\tilde{x},\tilde{u}_y)} & \text{if } (x,y) \in \text{Region 2} \\ \left.\dfrac{\partial G}{\partial y}\right|_{(\tilde{u}_x,\tilde{y})} & \text{if } (x,y) \in \text{Region 3} \\ G(\tilde{u}_x,\tilde{u}_y) & \text{if } (x,y) \in \text{Region 4} \end{cases} \tag{3.11}$$

We then form the likelihood as

$$L(\theta;x,y) = \prod_{i=1}^{n} g(\tilde{x}_i,\tilde{y}_i). \tag{3.12}$$

3.2.3 Functional Forms for V: Logistic and Bilogistic Models

Before we look at some analysis of bivariate extremes we must first look at what forms V can take in equation 3.10. There are various forms for V, depending on the dependence structure in the extremes, for example whether it is symmetric or asymmetric. A symmetric dependence structure is where X depends on Y to exactly the same degree as Y depends on X, in other words X and Y are exchangeable. An asymmetric dependence structure is where either X has a greater influence on Y than Y has on X, or vice versa. This report will focus on using a logistic and a bilogistic model to model symmetric and asymmetric dependence structures respectively, as these are the most commonly used choices from the same family, thus we can easily compare them.

The logistic model gives the form of V[1] as:

$$V(x,y) = (x^{-1/\alpha} + y^{-1/\alpha})^{\alpha}, \tag{3.13}$$

where $x > 0, \; y > 0, \; \alpha \in (0,1)$. We say that as $\alpha \to 1$ our variables become increasingly independent in the extremes; when $\alpha \to 0$ the extremes of our variables become increasingly dependent.

The bilogistic model has the following for $V^{[1]}$:

$$V(x,y) = -x\gamma^{1-\alpha} - y(1-\gamma)^{1-\beta}, \tag{3.14}$$

where $\alpha \in (0,1),\ \beta \in (0,1)$ and $\gamma = \gamma(x,y;\alpha,\beta)$ solves

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

In this approach, when $\alpha = \beta \to 1$ and one of α, β approaches 1 and the other is fixed, $\alpha - \beta$ determines the extent of asymmetry we have in the dependence structure. Also, if $\alpha = \beta$ then the bilogistic model reduces to the logistic model.

3.3 Illustrative Example

We will now look at bivariate extreme value theory in practice. From section 3.2 we know that our first step in conducting a bivariate analysis is to transform the data above the thresholds from both sites to unit Fréchet margins; we can check our transformations via probability plots. Probability plots for the four sites which exhibit between-site dependence can be seen in figure 3.2. The transformations are clearly correct, as all points lie almost directly on the unit diagonal. Once the extremal data has been transformed,

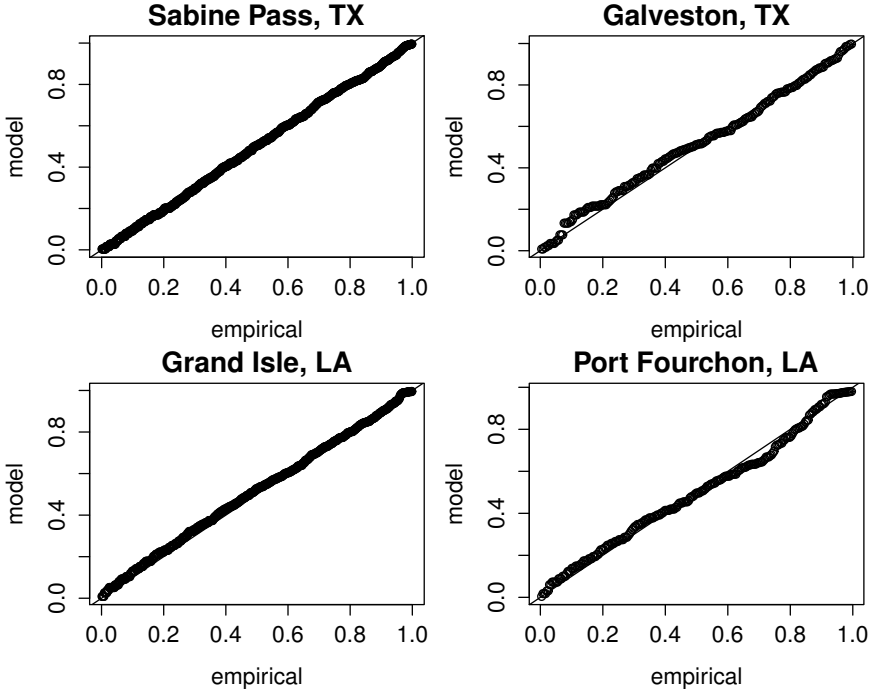

Figure 3.2: Probability Plots for the four sites exhibiting between-site dependence

we then move on to obtaining the likelihoods for each region as shown in figure 3.1, using equation 3.11. Figure 3.3 shows the data contributions to the likelihood for each region for Sabine Pass and Galveston. Note that there are no contributions from region 3.

At this stage we can either use the logistic or the bilogisttic model as a functional form for V. For the logistic model we run our MCMC scheme for $10,000$ iterations using a $U(0,1)$

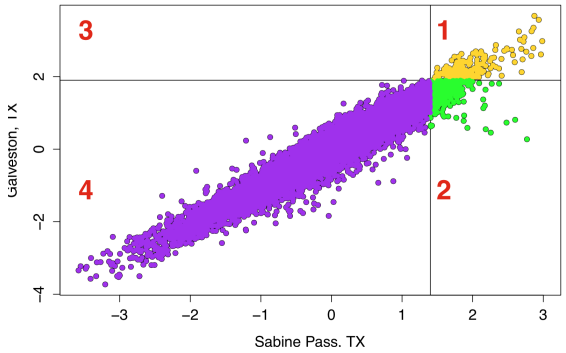

Figure 3.3: Plot showing which data points contribute towards the likelihood for each region

prior for the dependence parameter α, to find posterior draws for this parameter for each pairing, taking care to get suitable acceptance probabilities (see appendix 1 for code). For the bilogistic model we again run our MCMC scheme for $10,000$ iterations, taking $U(0,1)$ priors for both α and β, to find posterior draws for both dependence parameters for each pairing. Table 3.1 shows the posterior means for α in the logistic model, and α and β in the bilogistic model, after burn-in was discarded, for each pairing.

Site Pairings	Logistic		Bilogistic			
	α	Credible Interval	α	Credible Interval	β	Credible Interval
SP ∼G	0.42	(0.39, 0.46)	0.42	(0.36, 0.58)	0.31	(0.24, 0.40)
SP∼GI	0.77	(0.74, 0.79)	0.68	(0.62, 0.74)	0.62	(0.56, 0.70)
SP∼PF	0.83	(0.80, 0.86)	0.69	(0.63, 0.80)	0.92	(0.67, 0.79)
G∼GI	0.78	(0.74, 0.81)	0.63	(0.51, 0.75)	0.73	(0.62, 0.85)
G∼PF	0.82	(0.78, 0.85)	0.68	(0.55, 0.81)	0.64	(0.51, 0.79)
GI∼PF	0.41	(0.37, 0.44)	0.33	(0.26, 0.38)	0.35	(0.31, 0.40)

Table 3.1: Posterior means for α in the logistic model, and for α and β in the bilogistic model, with their 95% associated confidence intervals, for each site pairing

Recall section 3.2.3, where we discussed that if $\alpha = \beta$ then the biologistic model reduces to the logistic model. If we look at table 3.1 we can clearly see that the 95% credible intervals of all pairings overlap and thus we can cannot say α and β are significantly different. This means that in this case we need only do more analyses using the logistic model. In section 3.2.3 when considering the logistic model we also discussed how the value of α is a measure of the between-site dependence. From this we can say that sea-surge extremes at Sabine Pass and Galveston are quite dependent on each other, as are those observed at Grand Isle and Port Fourchon. There is not complete independence between extremes observed at the other pairings of sites, however they do not depend on each other quite as much as those just mentioned. If we think about this in terms of geographical positions (see figure 1.7 for site locations), we would expect there to be a higher between-site dependency in extremes (i.e. smaller α) for sites which are closer together. Using table 3.1 we expect Sabine Pass and Galveston, and Grand Isle and Port Fourchon to be close by and this is the case. Figure 3.4 plots the estimates of α against the distance the sites are from each other, to see if smaller α estimates correspond to

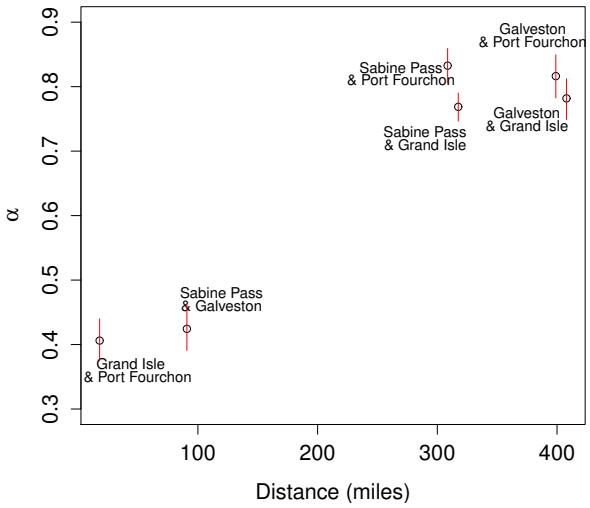

Figure 3.4: Posterior means (circles) for α with associated 95% confidence intervals plotted against geographical distance (miles) between sites.

geographical closeness and vice versa. Clearly we can see that sites further away from each other appear to have larger estimates of α and thus extremes here are less dependent on each other, and the sites which are closer together have smaller α estimates, and thus are more dependent on each other. It is clear that we must account for this between-site dependence in order to obtain the most accurate exceedance probabilities we can.

When examining bivariate extremes it is common practice to consider joint exceedance probabilities, which are basically a measure of how likely it is that certain extremes are breached at multiple sites. When assuming independence, we first need the probability of exceeding a value, say x for site X, which we can obtain using equation 3.8, by

$$\Pr(X > x) = 1 - H(x), \tag{3.15}$$

with similar for a value y at site Y. Then to obtain the exceedance probability we simply multiply the two marginal distributions together for $\Pr(X > x)$ and $\Pr(Y > y)$. Clearly the estimation is more involved when assuming dependence and we must follow the procedure as in section 3.2 to find α, as we have just done. Then to estimate the exceedance probability assuming dependence, say x is a value at site X and y a value at site Y, we use equation 3.10 where V has the form as in equation 3.13 to obtain

$$1 - \exp\{(x^{-1/\alpha} + y^{-1/\alpha})^{\alpha}\}. \tag{3.16}$$

Figure 3.5 shows posterior means, with 95% credible intervals, for the probability of simultaneously exceeding successively high marginal thresholds at two pairs of sites: Galveston and Sabine Pass, and Port Fourchon and Grand Isle, assuming independence in blue, and dependence in red. If we look at the scales given, it is evident for the pairings given that assuming independence results in lower exceedance probabilities than assuming dependence. Although this would be easier to see if we superimposed the dependence and independence exceedance probabilities on the same graph, results of this lead to distorted plots, as the independence probabilities are significantly lower than those for dependence. Since all 6 pairings had some degree of dependence if we incorrectly assume

independence, we consistently underestimate the chance of simultaneously breaching all the heights proposed, relative to an approach which assumes dependence. Taking Sabine Pass and Galveston as an example the chance of seeing a sea-surge at their thresholds (i.e. 1.4ft Sabine Pass, 1.9ft Galveston) has about a 0.0002 chance when assuming independence, but increases to about a 0.025 chance when accounting for dependence. Even though when assuming dependence the exceedance probability is still small, just for this one example it is about 100 times larger than when we assume independence. This would give us a false sense of security, we would believe a flood would be much less likely to happen than it actually was. We would underestimate the flood defences that needed to be in place, a consequence of ignoring extremal dependence.

As a final note, although we cannot see this on the plots, the 95% credible intervals do not overlap in any of the 6 given pairings, giving a significant result. This gives credit to this approach and so we must account for this dependence in order to have the most accurate exceedance probabilities.

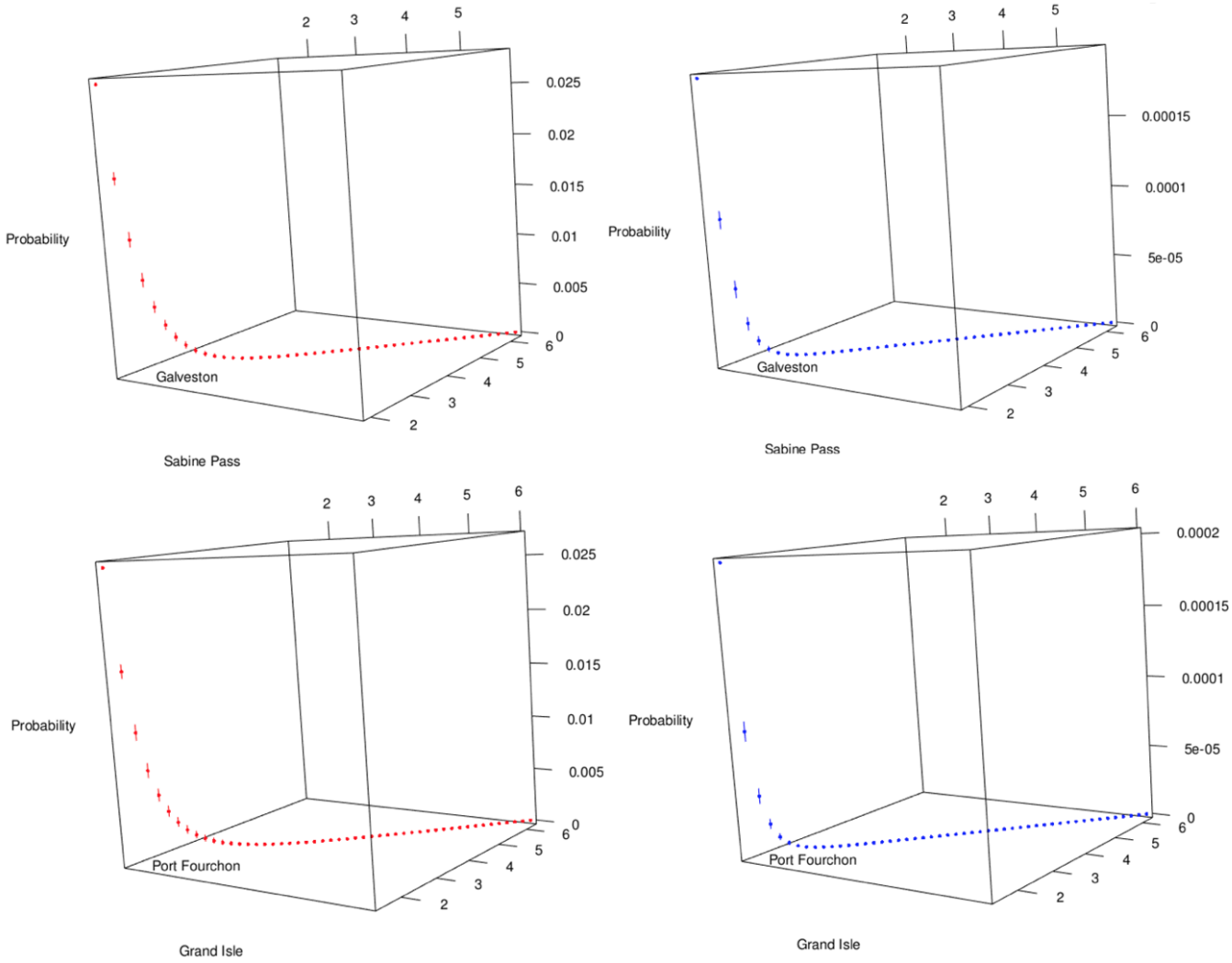

Figure 3.5: 3D plots showing exceedance probabilities (circles) with associated 95% confidence intervals of the pairings Sabine Pass with Galveston and Grand Isle with Port Fourchon. Plots in red assume between-site dependence and plots in blue assume between -site independence.

Chapter 4

Conclusion and Further Work

The main aim of this report was to understand the practical importance of studying environmental extremes and to find the best way to estimate return levels for sea-surges on the Gulf Coast of Mexico. We discussed the practical importance of studying environmental extremes in detail in chapter 1. If we can accurately estimate return levels, then we can put in preventative measures which can reduce loss of life and destruction and we considered various instances where this happened historically. We also looked at the theory behind the most common methods to model extremes, discounting the block maxima approach at it is wasteful of data, in favour of the method of threshold excesses. We discussed why we should adopt a Bayesian view when estimating return levels; there is a more intuitive interpretation for Bayesian confidence intervals, and because of the use of predictive return levels, which can not be used in a Frequentist setting.

As the report progressed, we demonstrated that if we fail to account for dependence when it exists we can get significantly over-estimated return level estimates, which could lead to levees built above the required safe height and would be financially wasteful. Various ways of accounting for serial correlation were discussed, however we settled on a complete Bayesian approach using a likelihood-based estimator for our inter-site dependence parameter θ. The runs declustering technique was discredited due to the sensitivity of the return level estimates to the choice of the declustering parameter κ, and that it has the potential of throwing away more extreme observations than the peak of some clusters. A moments-based estimator for the extremal index was also considered, however, being moments-based rather than likelihood-based, we could not take into account the uncertainty of this estimator in our MCMC scheme. We discussed how using a Bayesian framework enables us to make predictions into the future, incorporating all our uncertainty into one parameter, and we found these estimates were consistently higher than the return level estimates for the fully Bayesian approach, and as such are more convenient for practitioners to work with.

A final, important area of study was how we looked at modelling for bivariate extremes. We looked at Bayesian estimations of α, a dependence parameter in the commonly used logistic model for bivariate extremes. We found that generally, sites closer to each other had a smaller value of α, and thus sea-surge extremes at these sites were more dependent on each other. We finished our investigations by looking at the probability of simultaneously exceeding high marginal quantiles at pairs of sites and saw how if we incorrectly assume independence then we significantly underestimate the exceedance probabilities at all possible site pairings.

This work has shown us that when independence is present, we must account for it, as we can give return level estimates which are too high, and thus are a financial waste, or perhaps worse, we could significantly underestimate the probability of various extremes happening, where the results obtained coincided with those obtained by Fawcett

and Walshaw (2012)[27].

There are many areas of research that could stem from this project. Perhaps most importantly we could look at using informative priors for σ and ξ. Should we aim to include informative prior distributions, these would probably need to be constructed initially on some parameters an expert would feel comfortable with. For example, we might well be able to elicit expert prior distributions for return levels - these can then be transformed to obtain priors for the GPD parameters themselves. Although, in principle, this sounds straightforward, ensuring the expert specifies prior information regarding several return levels coherently could be a challenge. However if this was undertook, the inclusion of more information via this informative prior could lead to better return level estimations, which in turn could save money, and lives.

Another area of study would try looking for trend in the data, however for this we would need to obtain more than five years of data to accurately assess whether a trend was present. Also, throughout this report we have not accounted for seasonal variation, one approach here would be to allow the GPD parameters to vary by season, or to vary smoothly through time, see Fawcett and Walshaw (2012)[27].

We saw that there was strong between-site dependence for all pairings of sites in chapter 2.1. A further area of study could be to look into extending the bivariate model to account for dependence between all site triples, and to find out what may happen if we assume independence incorrectly. Leading on from this, a relatively new area of research has looked at using a spatial model to estimate how tall a barrier would need to be to account for a whole coastal front, by interpolating between the sites that we have. This would be much easier to relay back to practitioners and which could perhaps lead to safer coastlines. Much attention in recent literature has been given to the estimation of max-stable processes for estimating spatial extremes. Such processes are the infinite-dimensional extension of models for bivariate/multivariate extremes, and can allow for the estimation of spatial return level maps for whole coastlines. Such methods are comprehensively reviewed in Davison et al. (2012)[31].

There is clearly a vast area of study within extreme value theory and with its practical importance and potential areas of further work, it is a topic that must be considered and studied in more detail in the future.

Chapter 5

Appendix

1

```
bayesforalpha5=function(n,fx1,fy1,fx2,UY,UX,fy3,alphastart,erralpha){
  alpha=alphastart
  canalpha=vector("numeric")
  canalpha[1]=alpha
  x=vector("numeric")
  aprobalpha<-vector("numeric")
  x[1]=canalpha[1]
  loglik=function(fx1,fy1,fx2,UY,UX,fy3,ALPHA)  {
    if(ALPHA<0.00001)return(as.double(-1000000)) #if <0 return nothing
    if(ALPHA>0.9999)return(as.double(-1000000))     #if >0 return nothing
    loglik1= log(((fx1*fy1)^(-(1/ALPHA+1)))*((fx1^(-1/ALPHA)+
             fy1^(-1/ALPHA))^(ALPHA-2))*(((fx1^(-1/ALPHA)+
             fy1^(-1/ALPHA))^(ALPHA))-(1-1/ALPHA))*(exp(-(fx1^(-1/ALPHA)+
             fy1^(-1/ALPHA))^ALPHA)))
    loglik2= log((exp(-(fx2^(-1/ALPHA)+UY^(-1/ALPHA))^ALPHA))*((
             fx2^(-1/ALPHA)+UY^(-1/ALPHA))^(ALPHA-1))*(fx2^-(1/ALPHA+1)))
    loglik3= log((exp(-(UX^(-1/ALPHA)+fy3^(-1/ALPHA))^ALPHA))*((
             UX^(-1/ALPHA)+fy3^(-1/ALPHA))^(ALPHA-1))*(fy3^-(1/ALPHA+1)))
    loglik4= log(exp(-(UX^(-1/ALPHA)+UY^(-1/ALPHA))^ALPHA))
    loglik=sum(loglik1)+sum(loglik2)+sum(loglik3)+sum(loglik4)  }
  for(i in 2:n){
   canalpha[i]=x[i-1]+rnorm(1,0,erralpha)
   likely=exp((loglik(fx1,fy1,fx2,UY,UX,fy3,canalpha[i]))-
            (loglik(fx1,fy1,fx2,UY,UX,fy3,x[i-1])))
aprobalpha[i]=min(1,(likely*(dunif(canalpha[i],0,1))/(dunif(x[i-1],0,1))))
   u=runif(1)
   if(u<aprobalpha[i]){x[i]=canalpha[i]}
   if(u>=aprobalpha[i]){x[i]=x[i-1]}
   print(i)  }
  results=matrix(ncol=2,nrow=n)
  results[,1]=x
  results[,2]=aprobalpha
  return(results)  }
## Sabine Pass & Galveston
tryalpha_S.G=bayesforalpha5(10000,fS1_S.G,fG1_S.G,fS2_S.G,UG_S.G,US_S.G,
                            fG3_S.G,0.7,0.08)
alpha_S.G=tryalpha_S.G[,1]
acc.prob_S.G=mean(tryalpha_S.G[,2],na.rm=TRUE)
```

Bibliography

[1] Coles S. 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London

[2] 02/04/15, http://unitedkingdom.nlembassy.org/binaries/content/assets/postenweb/v/verenigd_koninkrijk_van_groot_brittannie_en_noord_ierland/nederlandse-ambassadeinlonden/import/keytopics/floodriskandwatermanagementinthe-netherlandsa2012update1.pdf

[3] 02/04/15, http://www.deltacommissaris.nl/english

[4] 15/03/15, http://www.academia.edu/5551438/BS_6399_Loading_for_buildings._Part_2_Wind_loads

[5] 02/04/15 http://rsta.royalsocietypublishing.org/content/363/1831/1293#sec5

[6] 03/04/15, http://www.floodsite.net/juniorfloodsite/html/en/student/thingstoknow/hydrology/1953flood.html

[7] 05/04/15, http://www.channel4.com/news/uk-weather-storms-1953-floods-norfolk-suffolk-essex

[8] 05/04/15, http://www.telegraph.co.uk/news/weather/10646439/TheThamesBarrierhassavedLondonbutisittimeforTB2.html

[9] 05/04/15, https://www.gov.uk/thethamesbarrier

[10] http://news.bbc.co.uk/1/hi/sci/tech/4075140.stm

[11] http://www.dailymail.co.uk/sciencetech/article2520867/NewmapshowsLondon-underwatercityThamesBarrierbuilt.html

[12] 04/04/15, http://www.earth-policy.org/plan_b_updates/2006/update56

[13] Field CB, Barros V, Stocker TF, Dahe Q. 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change

[14] 05/04/15, http://www.history.com/topics/hurricanekatrina

[15] 05/04/15, http://www.nasa.gov/vision/earth/lookingatearth/h2005_katrina.html

[16] 05/04/15, http://www.history.auxpa.org/collections/photographs/katrina/Aux KATRapuscg_05BV4.jpg

[17] http://www.livescience.com/11177hurricanekatrinahitorleanstoday.html

[18] 15/10/14, http://tidesandcurrents.noaa.gov/waterlevels.html

[19] 10/11/15,http://www.ezilon.com/maps/united-states/louisiana-geographical-maps.html

[20] 04/04/15, http://www.wunderground.com/hurricane/surge_wetlands.asp

[21] Fawcett L. 2014. MAS8304: Topics in Statistics, Environmental Extremes

[22] Gilks WR, Richardson S, Spiegelhalter DJ. 1997. Markov Chain Monte Carlo in Practice

[23] Boys RJ. 2014. MAS3321: Bayesian Inference.

[24] Gelman A, Gilks WR, Roberts GO. 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms

[25] Fawcett L, Walshaw D. 2007. Improved estimation for temporally clustered extremes

[26] Coles SG, Tawn JA. 1991. Modelling extreme multivariate events.

[27] Fawcett L, Walshaw D. 2012. Estimating return levels from serially dependent extremes

[28] Fawcett L, Walshaw D. 2015. Sea-surge and wind speed extremes: optimal estimation strategies for planners and engineers

[29] Ferro CAT, Segers J. 2003. Inference for clusters of extreme values. Journal of the Royal Statistical Society, Series B (Statistical Methodology)

[30] Süveges M, Davison AC. 2010. Model Misspecification in Peaks Over Threshold Analysis

[31] Davison AC, Padoan SA, Ribatet M. 2012. Statistical Modelling of Spatial Extremes

[32] Fawcett L, Walshaw D. 2014. Return level estimation: A review and best practice

[33] Coles SG, Tawn JA. 2005. Bayesian modelling of extreme surges on the UK east coast