Statistical Modelling of Extreme Rainfall

T.W.Anderson

April 30, 2014

Abstract

The aim of this report is to show the flaws in methods such as simplest counting when working with extreme values and therefore the need to use the GEV distribution and the GPD. The data we have worked with consists of hourly recordings of rainfall in tenths of millimetres over 63 from over 1000 sites across England and Wales. We have shown that hourly values have a decrease in trend while the aggregated daily and 5 day totals have less trend or no trend at all. These are shown, along with 100 year return levels, in colour density plots.

Acknowledgements

I like to acknowledge Chris Kilsby for supplying the original project proposal along with Hayley Fowler and Stephen Blenkinsop for setting up the links to the data. I would also like to thank the UK Meteorological office for supplying the data and the Environmental Agency for the CONVEX project funded by the NERC.

Contents

Chapter 1

Introduction

1.1 Motivation

Over the past few years there has been a growing need for an accurate way to predict extreme weather. With Hurricane Katrina (2005) and Hurricane Sandy (2012) devastating the east coast of America, together killing over 2,000 people and causing an estimated damage of over $140 billion; it is clear that the current defences are inadequate. Extreme weather is not only an international problem: in the UK, the 2007 summer floods damaged over 48,000 homes and nearly 7,000 businesses, while the winter period of 2013/14 was the wettest since records began. Although we can consider ourselves lucky compared to our American neighbours, we are no better equipped to deal with extreme weather. This is somewhat surprising since existing data shows that the current flood defences are not enough to deal with storms which would not even be considered extreme.

1.2 Example

Rainfall was recorded in Jesmond Dene over a 28-year period, where the highest amount of rain in a single day per year was noted in Table 1.2.

Suppose Newcastle city council are interested in the probability that rainfall in a single day will exceed 35mm. The simplistic counting approach would involve counting the number of exceedances and dividing that by the number of observations:

$$Pr(rainfall\ exceeds\ 35mm) = \frac{15}{28} = 0.54$$

415	280	565	520	455	280	325
265	530	480	182	432	372	216
384	672	472	268	250	276	432
420	234	332	550	460	300	292

Table 1.1: Annual maximum rainfall in tenths of mm in Jesmond Dean, 1984-2011.

But what if they require the probability that the rainfall exceeds 70mm?

$$Pr(rainfall\ exceeds\ 70mm) = \frac{0}{28} = 0$$

According to this method, there is no chance of more than 70mm of rain in a single day. This is an unreasonable conclusion. Additionally, what if they wanted to know how high to build flood defences to protect against a flood they would expect to see once every 14 years? Since the data is annual, an intuitive argument would be to find the value for which $1/14 \approx 7\%$ of the data exceeds it. Here, we have 28 observations and 7% of these exceed 56.5mm. But what if they required the same information for a flood they would expect to see once every 50 years? We would need the value for which 2% of the data is above, but with only 28 observations we need about half on observation. So we do not have enough data to perform this calculation using these simplistic counting methods. It is imperative that we be able to estimate extremes that exceed those already observed; but we would require a statistical model that can extrapolate beyond our range. For this, we need to use extreme value theory.

1.3 The Extremal Types Theorem

The previous example shows daily rainfall totals with the annual maximum recorded, so the number of observations (for non-leap years) is n=365. We could use the notation $M_{365,i}, i = 1, ..., 28$ to denote the 28 observations. A more general approach to this is to suppose that $X_1, X_2, ..., X_n$ are a sequence of independent and identically distributed random variables with a common distribution function F, then $M_n = max\{X_1, X_2, ..., X_n\}$. But what is the distribution of M_n?

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

But in practice, the distribution function F is unknown, so we have to look asymptotically as $n \to \infty$ and determine if there is a distribution function G such that we can estimate M_n without referring to F. [Coles, 2001, p.45-46]

Taking the limit of the distribution M_n leads to a degeneracy as it converges to a single point on the real line with probability 1. This is comparable to the sample mean $\bar{X}$ converging to the population mean μ by the Central Limit Theorem. In that case the degeneracy is prevented by allowing a linear rescaling, so that

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

where $b_n = \mu$ and $a_n = \sigma/\sqrt{n}$, where σ is the population standard deviation and n is the sample size.

So applying the same method to M_n, if there exists a sequence of constants $a_n > 0$ and b_n such that, as $n \to \infty$

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

for some non-degenerate distribution G, then it can be shown that G is of the same type as one of the following distributions:

$$I : G(z) = \exp\{-\exp(-z)\}, \ -\infty < z < \infty; \tag{1.1}$$

$$II : G(z) = \begin{cases} 0, & z \leq 0, \\ \exp\{-z^{-\alpha}\}, & z > 0, \alpha > 0; \end{cases} \tag{1.2}$$

$$III : G(z) = \begin{cases} \exp\{-(-z^{\alpha})\}, & z < 0, \alpha > 0, \\ 1, & z \geq 0. \end{cases} \tag{1.3}$$

The three distributions of the Extremal Types Theorem (I, II *and* III) are known as the Gumbel, Fréchet and Weibull types respectively. These can all model extreme data without the use of the parent distribution F, but how do we know when to use each of these distributions? This is when they can be combined to form the Generalised Extreme Value distribution. [Tippett, 1928]

1.4 The Generalised Extreme Value (GEV) distribution

Working with three distributions is inconvenient, but there does exist a distribution which combines all three. This is referred to as the generalised extreme distribution (GEV) and this was derived independently by Von Mises

[1954] and Jenkinson [1955] and has the cumulative distribution function

$$G(z;\mu,\sigma,\xi) = \exp\Big\{ -\Big[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}. \tag{1.4}$$

However, $\xi = 0$ is not defined in Equation (1.4), so it is taken as the limit as $\xi \to 0$, given by

$$G(z;\mu,\sigma) = \exp\Big\{ -exp\Big(\frac{z-\mu}{\sigma}\Big)\Big\}. \tag{1.5}$$

Here $\mu(-\infty < \mu < \infty)$ is the location parameter, $\sigma(> 0)$ is the scale parameter and $\xi(-\infty < \xi < \infty)$ is the shape parameter. Different values of ξ determine which of the three extreme value distributions will be used. $\xi = 0$ corresponds to Equation (1.1), $\xi > 0$ corresponds to Equation (1.2) and $\xi < 0$ corresponds to Equation (1.3). [Coles, 2001, p. 47-48] Now we can get the probability density function of the GEV if we differentiate Equation (1.4). This is given by

$$g(z;\mu,\sigma,\xi) = \frac{1}{\sigma}\Big[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi+1} \exp\Big\{ -\Big[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}. \tag{1.6}$$

Now if we wanted to calculate the r-year return level, we have to calculate $Pr(annual\ maximum > z_r) = 1/r$, i.e.

$$1 - Pr(annual\ maximum \le z_r) = 1 - G(\hat{z}_r;\hat{\mu},\hat{\sigma},\hat{\xi}) = 1/r \tag{1.7}$$

which gives

$$\exp\Big\{ -\Big[1+\hat{\xi}\Big(\frac{\hat{z}_r-\hat{\mu}}{\hat{\sigma}}\Big)\Big]^{-1/\hat{\xi}}\Big\} = 1 - 1/r. \tag{1.8}$$

Solving Equation (1.8) for $\hat{z}_r$ gives the r-year return level as

$$\hat{z}_r = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\Big\{[-log(1-r^{-1})]^{-\hat{\xi}} - 1\Big\}. \tag{1.9}$$

Now we need to find the values for $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$. To do this, we use the method of maximum likelihood estimation and this can be implemented using the R package *ismev*.

So if we revisit the example from Section 1.2, we can easily obtain these estimates from the following output:

```
> library(ismev)
> jesmond=c(415,280,565,520,455,280,325,246,530,480,182,
+           432,372,216,384,672,472,268,250,276,432,420,
+           234,332,550,460,300,292)
> gev.fit(jesmond)
$conv
```

```
[1] 0

$nllh
[1] 173.3856

$mle
[1] 328.0209356 106.7688494  -0.1141183

$se
[1] 23.8781901 18.0003544  0.1899403
```

Looking at this output, *$cov=0* tell us that convergence has occured while*$nllh=173.3856* gives us an estimate for the log—likelihood. *$mle* tells us that $\hat{\mu} = 328.02$, $\hat{\sigma} = 106.77$ and $\hat{\xi} = -0.1141183$ and we are also given their standard errors. Putting these values back into Equation (1.9) we find that the 14-year return level $z_{14} = 56.8mm$ and the 50-year return level $z_{50} = 66.4mm$. Also, using Equation (1.7) we can now calculate that $Pr(rainfall\ exceeds\ 35mm) = 0.556$ and $Pr(rainfall\ exceeds\ 70mm) = 0.012$ i.e. once every 1.7 years and once every 85 years respectively.

The r-year return levels can also be calculated in R using the package *extRemes* and can even return a plot with confidence regions. The years in question are arbitrary but we have selected 14, 50, 100 and 1000 for this example. The output is as follows:

```
> library(extRemes)
> return.level(x,rperiods=c(14,50,100,1000))
$conf.level
[1] 0.05

$return.level
[1] 568.4022 664.2344 710.1394 838.2564

$return.period
[1]   14   50  100 1000

$confidence.delta
        lower     upper
[1,] 476.6302  660.1743
[2,] 489.2706  839.1983
[3,] 475.1873  945.0916
[4,] 362.0964 1314.4165
```

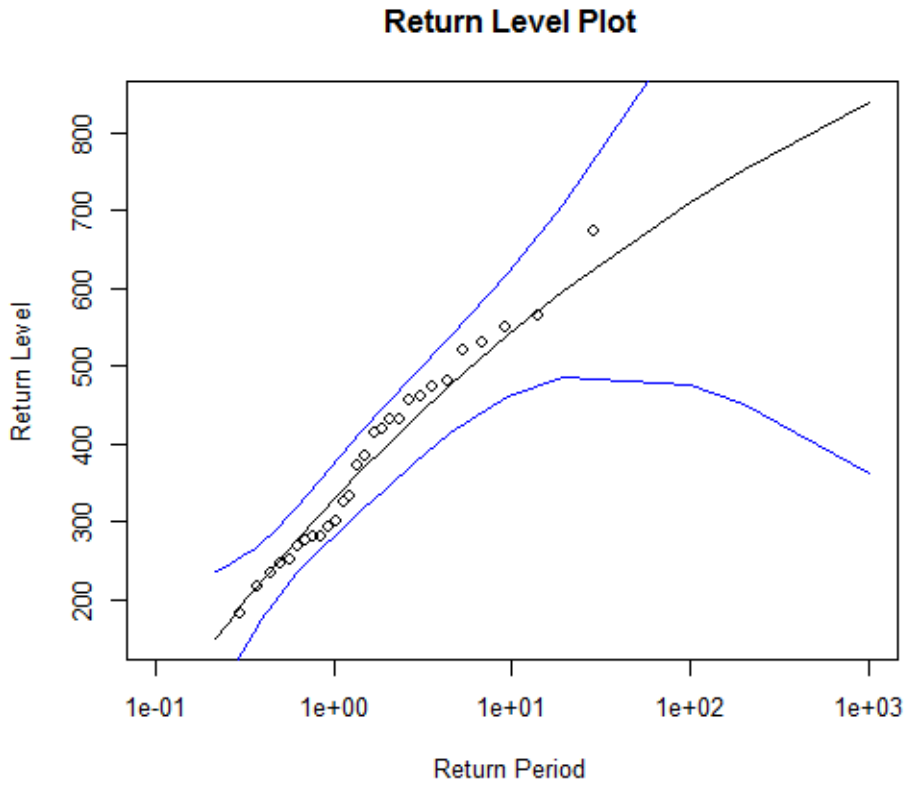

Figure 1.1: Return levels for Jesmond Dene.

As you can see from Figure 1.1, the R command uses the existing data you have to estimate r-year return level. It is important to point out that this does not show what we expect to happen once every r years, rather the probability of this occuring in the next year is $1/r$.

Chapter 2

UK rainfall data

2.1 Overview

The data we will be working with consists of 1281 sites from across 7 regions divided into 19 sub-regions of England and Wales:

- Anglian: Anglian Central(AC),Anglian Eastern(AC) and Anglain Northern(AN);
- Midlands: Midlands Central (MC), Midlands East(ME) and Midlands West(MW);
- North-East: North-East North-East(NENE) and North-East Yorkshire (NEY);
- North-West: North-West-North (NWN) and North-West-South(NWS);
- South-East: South-East Kes (SEK), South-East Net(SEN), South-East Ssd(SES) and South-East WT(SEW);
- South-West: South-West D&G(SWD) and South-West Wessex(SWW);
- Wales: Wales Northern(WN), Wales South-East(WSE) and Wales South-West (WSW).

Each of these have hourly rainfall measured in tenths of millimeters. For consistency all sites begin at midnight on 1st January 1949 and run through till midnight on 31st December 2011 so there is a total of 63 years of data. However, most sites don't actually have complete recordings–they begin later and finish earlier-as shown in Figure 2.1.

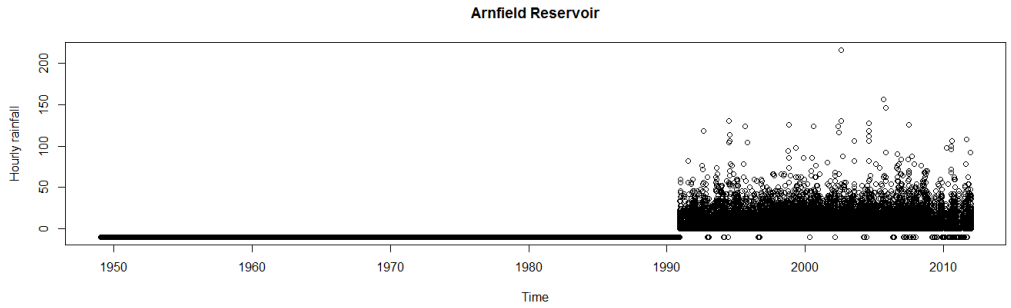

Figure 2.1: Hourly rainfall for Arnfield Reservoir.

This particular site is Arnfield Reservoir and as you can see the data starts around 1990 and runs through till 2011. You will also notice that when the data is not recorded the missing value is denoted -10, but this is only to aid the understanding of Figure 2.1, in the actual data they are denoted -999. This is fine when working with hourly values, but becomes problematic when we sum over 24 values to make daily totals. This issue is that if a day has large values in it but also missing values, then the day may not be recorded as the maxima even though it may be. This problem will be addressed further in Section 2.2. The data is sorted into regions and sub-regions to make analysis easier so for the majority of the work we will be focusing on the North-West-South sub region.

2.2 GEV fit to annual maxima of daily totals

To begin the analysis of annual maxima of daily totals we must first prepare the data to fit the model. This will first start with summing over 24 values and determining the appropriate method to deal with missing data, then finding the max value per year-which will also involve building a model to incorporate leap years. We must also decide the cut off point for when a year has too many missing values to be considered valid. This is just arbitrary so we will decide that if a year is missing 1/6 of its values then it is not valid and hope that the data is robust.

So first, we will set up a function to deal with leap years. The following code will set up vector with each value the first day of each year and the final value is the last day of the data:

```
> leap=c(rep(c(365,365,366,365),15),365,365,364)
> func=function(x){
+   y=vector()
+   y[1]=1
+   for(i in 1:length(x)){
```

```
+     y[i+1]=y[i]+x[i]
+   }
+   return(y)
+ }
> leaps=func(leap)
> leaps
 [1]     1   366   731  1097  1462  1827  2192  2558  2923
[10]  3288  3653  4019  4384  4749  5114  5480  5845  6210
[19]  6575  6941  7306  7671  8036  8402  8767  9132  9497
[28]  9863 10228 10593 10958 11324 11689 12054 12419 12785
[37] 13150 13515 13880 14246 14611 14976 15341 15707 16072
[46] 16437 16802 17168 17533 17898 18263 18629 18994 19359
[55] 19724 20090 20455 20820 21185 21551 21916 22281 22646
[64] 23010
```

Next we will find the number of missing values are in each year and remove any years that we deem void, this is done with a function that sums up the number of values denoted by -999 and removes the year if there is more than 1460 (1/6 of 8760). If the site passes this test then we will classify the remaining missing values as 0. This choice is again arbitrary as we don't know the cause of the missing value (it could be human error, damage due to extreme weather etc.). We have also decided that if a site has less that 5 years of valid data, then it will be void. If all of these are complied with, then we shall fit the GEV. The code to calculate this is quite lengthy so will not be placed here. Figure 2.2 shows the diagnostic plots of GEV fitted to Arnfield Reservoir's annual maxima. We have achieved convergence and considering we only have 17 data points in the particular site, the quantile plot shows that the data fits reasonably well to a straight line suggesting a reasonable fit.

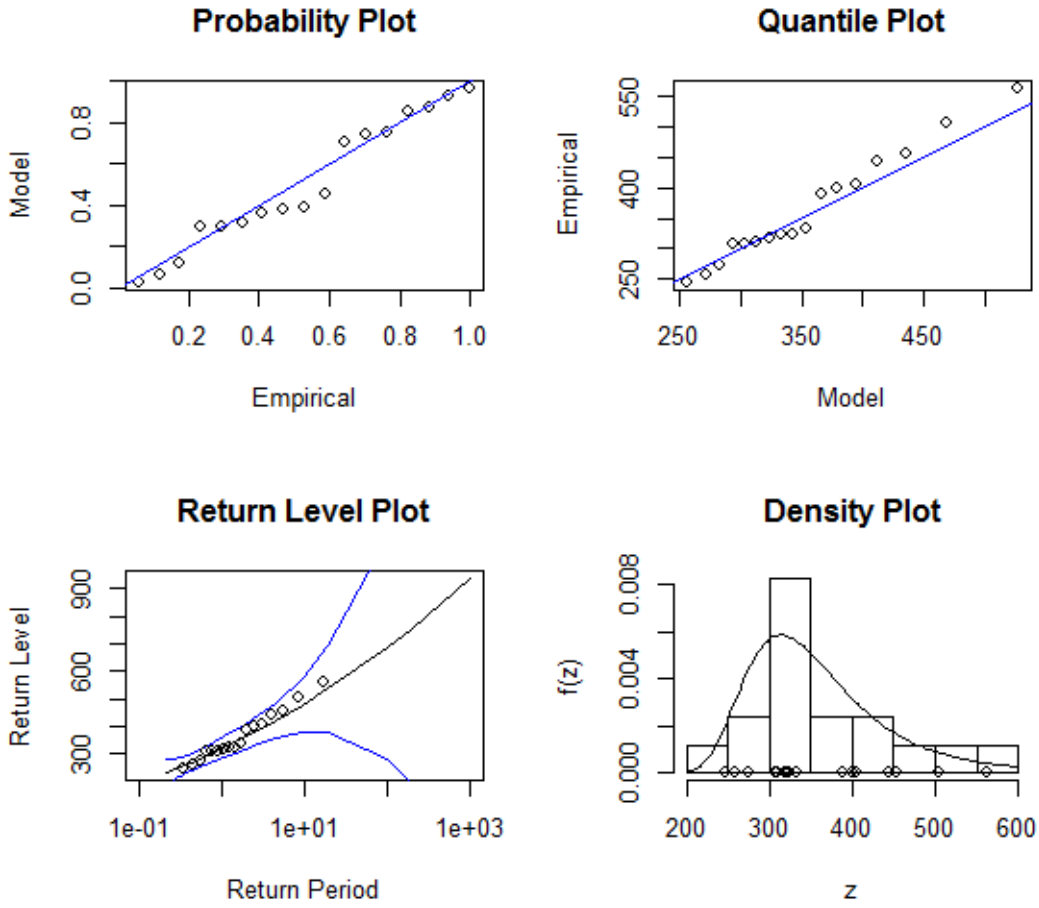

Figure 2.2: Diagnostic plots indicating the goodness-of-fit of the GEV to Arnfield Reservior rainfall.

2.3 GEV with trend

In the previous sections we have assumed that there is no time dependence in this data. Considering there are up to 63 years worth of data it is pretty naive of us to not consider it. The procedure to check if time has an influence is pretty straightforward once we have the previous code. We will now create a vector with standardised years from 1949 to 2011 and input that into the GEV model when fitting the data. This will now give us slope and intercept parameters instead of a location parameter. Applying this method to a whole sub-region and looking at the slope parameter will help us determine whether of not the sub-region is changing over time. For this i will use North-West-South and show my findings in a histogram along with the 95% confidence interval for the slope:

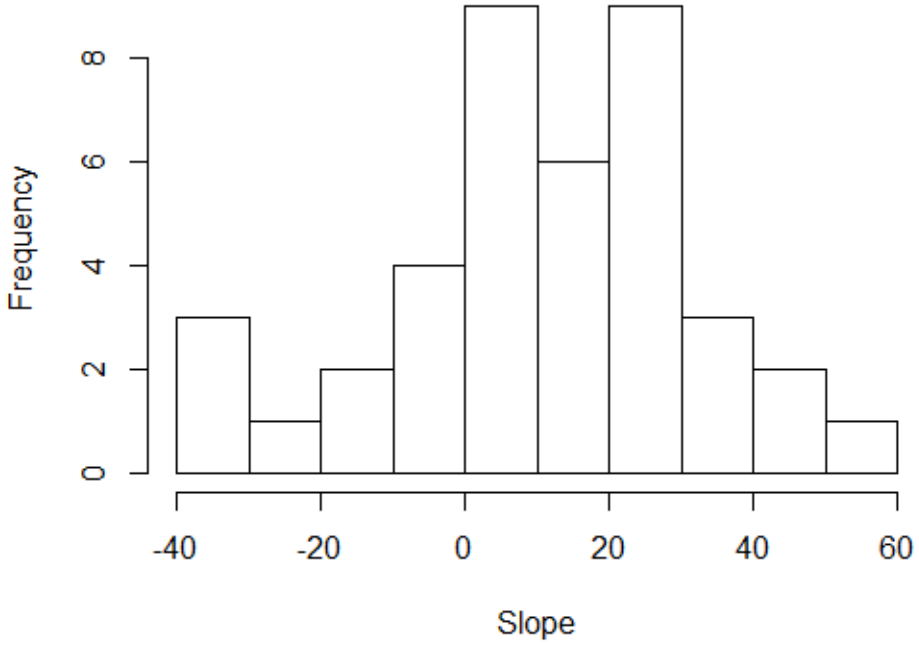

Figure 2.3: Histogram of slope parameters from North-West-South.

```
> c(mean(matout[,3])-1.96*sd(matout[,3])/length(matout[,3]),
+   mean(matout[,3])+1.96*sd(matout[,3])/length(matout[,3]))
[1]  9.291649 11.367545
```

So the North-West-South region has an average slope of 10.33 (9.29,11.37) so it is clear from both the confidence interval and Figure 2.3 that this particular region has a positive slope on average which suggests that the extreme weather is increasing over time. This theory will be further examined in Secion 3.4. The final code for this section is in Appendix A

Chapter 3

Generalised Pareto Distribution

3.1 GPD derivation

When working with extreme values we have to determine what we consider extreme. A simple way of determining this is to set a threshold and all values greater than this value can be considered, but to determine this threshold efficiently we must consider the asymptotic theory that would be appropiate for this problem.

Recall Sections 1.3 and 1.4 where we showed that $Pr\big(M_n \leq z\big) \approx G(z)$, where $G(z)$ is defined in Equation (1.4). For a large enough threshold u, the distribution function of (X-u) conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\bar{\sigma}}\right)^{-1/\xi}, \tag{3.1}$$

defined on $y > 0$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{3.2}$$

However, just like with Equation (1.4), this is not defined when $\xi = 0$ so we take the limit $\xi \to 0$, giving

$$H(y) = 1 - \exp\big(-\frac{y}{\tilde{\sigma}}\big), y > 0; \tag{3.3}$$

i.e. an exponential distribution with rate $1/\tilde{\sigma}$ Any distribution that is definded by Equation 3.1 is a member of the *Generalised Pareto family* and the distribtion itself is known as the Generalised Pareto Distribution (GPD). Picklands [1975]

3.2 Threshold selection

We know from the *threshold stability property* Coles [2001] of the GPD that if a chosen threshold allows for the GPD to be a valid model then it is also valid for all values above the threshold. The expected value of our threshold is given by

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

where u_0 is some threshold and σ_{u_0} is the GPD scale parameter for excess over the threshold u_0. This is better described using the *mean residual life* (MRL) plot citetcoles01, where a plot is produced modelling the data above each potential threshold and the aim is to find where the data is most stable.

Here we will use Arnfield Reserviour once again as an example and I will use the MRL plot to determine the best threshold. Again the R package *ismev* provides handy code for this and also lets us determine a confidence interval:

```
mrl.plot(ar,umin=0,conf=.95)
```

We have decided to use the 95% confidence interval as is the norm and since there are missing values in the data (denoted -999) we have set the minimum to be zero as we don't want a negative threshold. The rough idea is to determine the left most point of the Figure such that a straight line can be drawn out to the end of the data without crossing the confidence intervals. For Figure 3.1 we have already drawn the line on to show what we beleive is the best threshold. [Coles, 2001, p. 78-80]

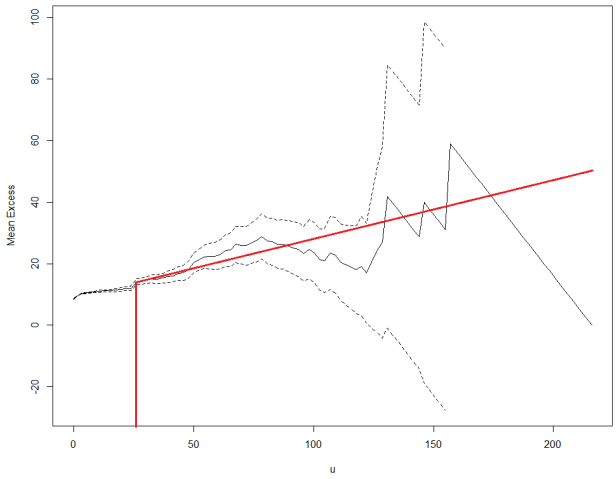

Figure 3.1: Mean residual life plot of Arnfield Reservior.

Looking at Figure 3.1 we have determined that the optimal threshold is 35 (3.5mm), but also it is important to note that the right hand side of the plot is unreliable, the data is limited so the variability is very high. when it is clearly visible that the variability is high we can disregard the data. We can repeat this method for several sites in the region to determine a constant threshold for this sub region:

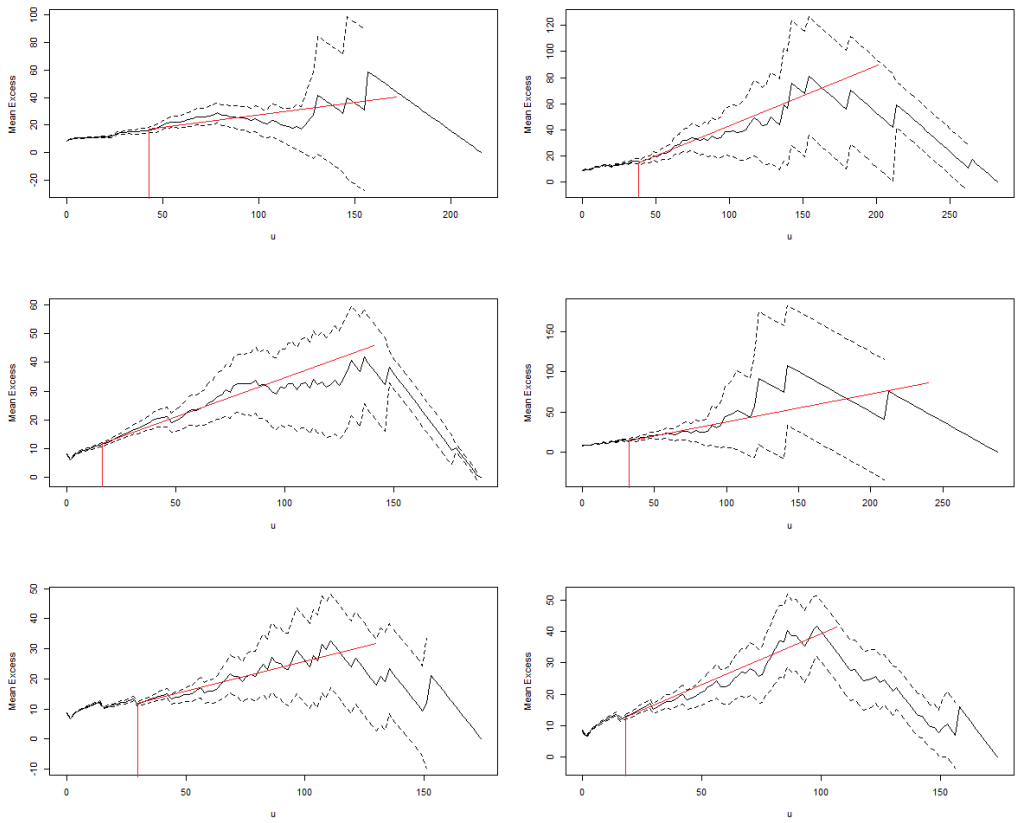

Figure 3.2: MRL plots from first 6 sites in North-West-South.

Obviously this method is time comsuming and there is no method to efficiently and effectively determine a threshold so for the purposes of this project we will consider two different thresholds for each site which we fix at the 95^{th} and 99^{th} percentile of each site.

3.3 Fitting the GPD

To fit the GPD we will once again find the maximum likelihood estimates for the parameters. We can do this by again using R package *ismev* but with

the threshold being decided by a percentile of the data, we must remove the initial and final sections of the data where no recording takes place. The method of doing this and the output are given below:

```
> ar=read.table("Arnfield_Reservoir_LOG.txt")[,7]
> start=which(ar!=-999)[1]  #find where data starts beings recorded
> missing=which(ar!=-999)   #locate all missing value
> finish=missing[length(missing)] #find when data ends
> ar1=(ar[start:finish])
> threshold=quantile(ar1,.95)
> fit=gpd.fit(ar1,threshold)
$threshold
95%
  8

$nexc
[1] 7597

$conv
[1] 0

$nllh
[1] 26173.71

$mle
[1] 10.95005532  0.05183838

$se
[1] 0.165932584 0.009914643
```

So for this particular site we have $\hat{\sigma} = 10.950$ and $\hat{\xi} = 0.052$. Once again we can test the model adequacy:

```
gpd.diag(fit)
```

Figure 3.3 indicates that the threshold exceedences have a reasonable fit to the GPD, but as with the GEV we need to determine whether or not there is a time dependence in the data. This will involve adding a time vector to the fit but as this distribution has no locaton parameter, we will add it to the scale parameter.

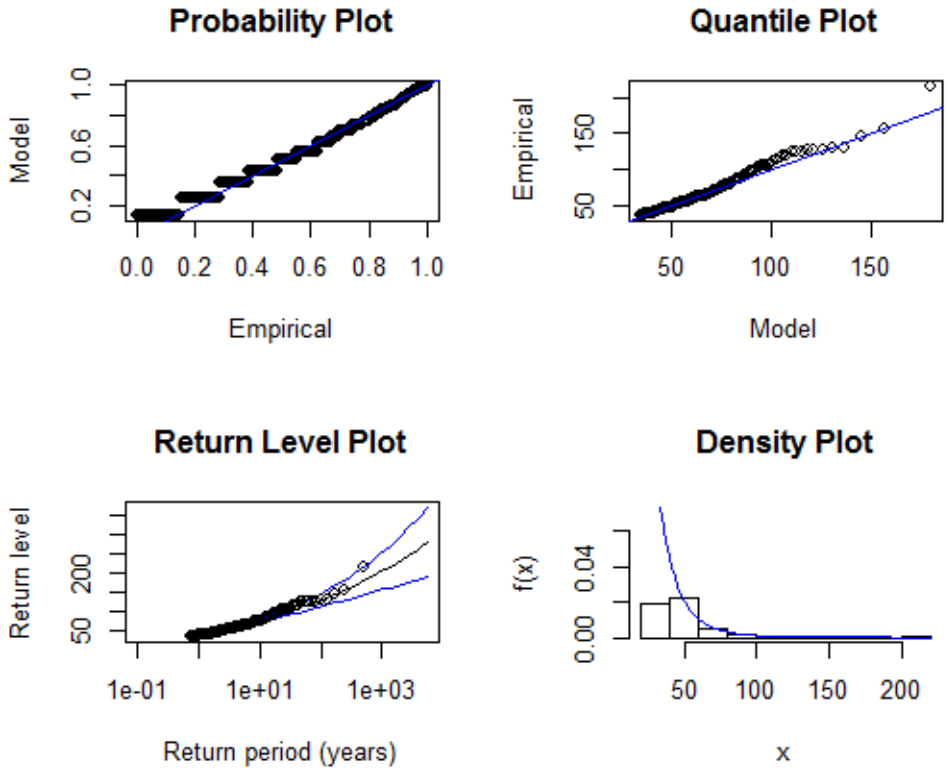

Figure 3.3: Diagnostic plots indicating goodness of fit of the GPD to Arnfield Reservior

3.4 GPD with trend

This part is very similar to that of Section 2.3 where we will have an intercept and slope parameter and leave the shape parameter alone. The difference will be that instead of using standardised years we will make each time increment 1/8766 (8766 being the average number of hours in a year) and run this from the beginning before we cut the useless parts of the data away. Applying this method to the whole sub-region for thresholds at the 95% and 99% levels we can determine whether they have a time dependence.

```
> ci95
[1] 0.03683099 0.04134837
> ci99
[1] 0.06727490 0.07924716
```

So again, with the 95th percentile confidence interval at (0.037,0.041) and the 99th percentile confidence interval at (0.067,0.079), along with Figure 3.4, we can see that this particular region shows an increase in the amount of extreme

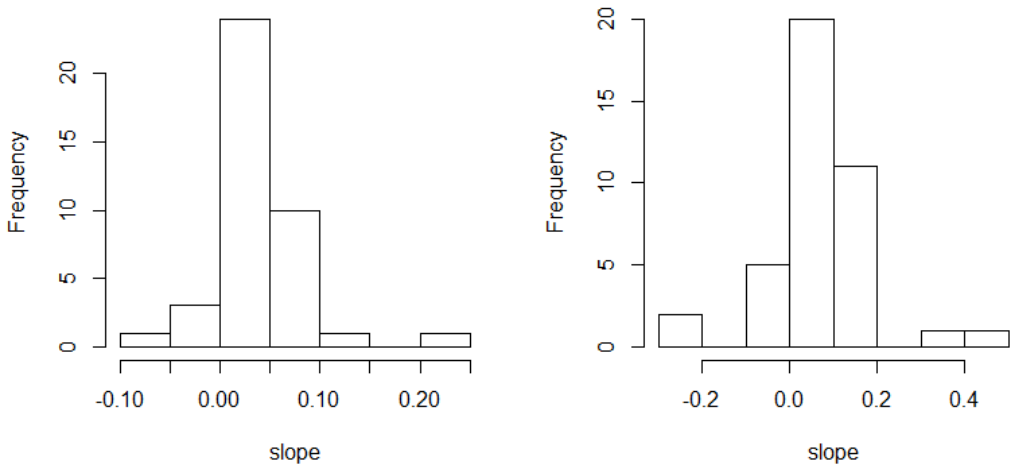

Figure 3.4: Histogram of slope parameters from North-West-South with threshold at 95% (left) and 99% (right)

rainfall over time. However, with GPD and thresholds, it is important to check extreme dependence.

3.5 Dependence in extreme values

When values exceed the threshold, the chances are that they will exceed several times in quick succession because there happens to be a storm. This raises the issue of the observations no longer being independent of each other as they were deemed to be in Section 1.4 and therefore their standard errors will be inaccurate. There are three commonly used methods to deal with these issues:

1. Fit all exceedences with GPD and ignore the dependence but adjust the standard errors accordingly.

2. Model the dependence in the process.

3. Filter the exceedences to achieve approximate independence.

We shall focus on the third approach, which is the most commonly used, referred to as declustering. This involves picking a declustering parameter κ for which a storm is deemed to have ended if there are κ consecutive values below the threshold after the threshold has been exceeded and then we record the maximum value from that storm. This works as it is regarded that exceedences from different storms are independent of each other. The aim in choosing κ is to make sure it is large enough so that we can assume independence but small enough so that there are sufficient cluster exceedences to form the inference (the Goldilocks principle).

The value of κ is often chosen arbitrarily based on inferences from the data so we have $\kappa = 10$ for hourly data (i.e. one storm is considered to have ended and another has began if there are 10 consecutive points below the threshold). The R code for this has been taken from Fawcett [2013, p. 66] and is given below:

```
cluster10values=function(dataset,threshold){
    x=list()
    z=list()
    j=1
{
  for(i in (11):length(dataset)){
    if(dataset[i-10]>threshold
       & dataset[i-9]<=threshold & dataset[i-8]<=threshold
       & dataset[i-7]<=threshold & dataset[i-6]<=threshold
       & dataset[i-5]<=threshold & dataset[i-4]<=threshold
       & dataset[i-3]<=threshold & dataset[i-2]<=threshold
       & dataset[i-1]<=threshold
       & dataset[i]<=threshold){
      x=max(dataset[j:i])
      ifelse(i !=length(dataset), j<-i+1, NA)
      z=c(z,x)}}}
```

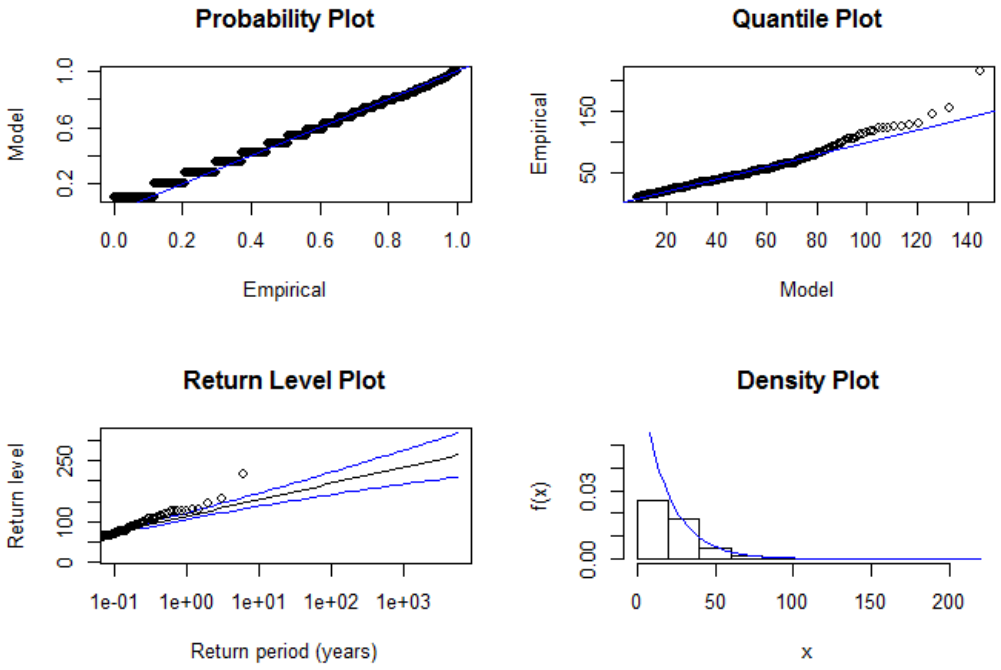

Figure 3.5: Diagnostic plots indicating goodness of fit of the GPD to Arnfield Reservior with declustering)

Of course this code isn't very efficient, but for the purposes of what we want to do it is sufficient and can be easily edited if we wish to change the value of κ. Now if we combine this with our previous code we can we can check the goodness of fit and compare it to our previous code. Looking at this Q-Q-plot, the data appears to fit worse to the line in Figure 3.5 than it does in Figure 3.3 which is not what we expected to see. The fit should have improved. This could be accounted to the lack of data after declustering.

3.6 Dependence in extreme values with trend

Now we must combine the previous methods from Sections 3.4 and 3.5 to find if there is a trend in the data now that it has been declustered. All this requires is for us to decluster the associated times of the data and fit the these to the GPD. Again we shall check if there is a trend using a confidence interval and histogram. With the confidence intervals at (0.05,0.11) and (0.03,0.16) along with Figure 3.6 it is clear that there is a positive trend in the data suggesting that when extreme rainfall occurs, it is likely to be more extreme than in previous years. The final code is located in Appendix B

```
> ci95
[1] 0.04981116 0.11165395
```

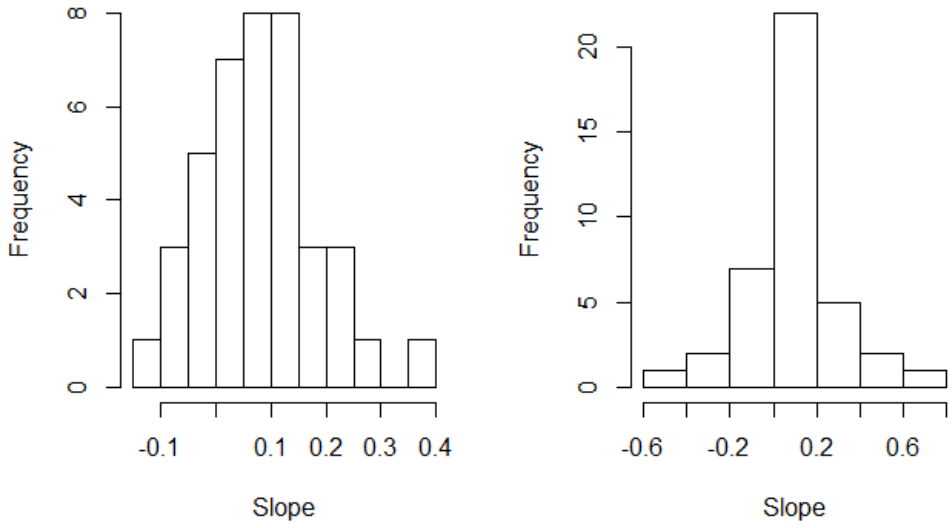

Figure 3.6: Histogram of slope parameters from North-West-South with threshold at 95% (left) and 99% (right) with declustered data

Chapter 4

Full Data Analysis

4.1 Hourly Data

Now we shall run the final function from Section 3.6 for the entire data and determine how rainfall have behaved across the country for the 63 years. The means and confidence intervals are given below:

	Hourly 95% threshold Mean		Hourly 99% threshold Mean	
AC(a)	-0.141	(-0.234,-0.047)	-0.230	(-0.401,-0.058)
AE(b)	0.011	(-0.074, 0.095)	0.029	(-0.061, 0.119)
AN(c)	-0.159	(-0.259,-0.060)	-0.292	(-0.460,-0.124)
MC(d)	-0.201	(-0.330,-0.072)	-0.050	(-0.165, 0.064)
ME(e)	-0.012	(-0.196, 0.171)	0.032	(-0.182, 0.247)
MW(f)	-0.290	(-0.457,-0.122)	-0.249	(-0.511, 0.013)
NENE(g)	-0.018	(-0.168, 0.133)	0.060	(0.120, 0.240)
NEY(h)	-0.110	(-0.278, 0.058)	-0.263	(-0.611, 0.084)
NWN(i)	0.117	(0.077, 0.158)	0.100	(0.042, 0.159)
NWS(j)	0.081	(0.050, 0.112)	0.095	(0.030, 0.160)
SEK(k)	0.018	(-0.042, 0.079)	0.002	(-0.079, 0.083)
SEN(l)	-0.071	(-0.119,-0.022)	0.003	(-0.065, 0.070)
SES(m)	0.011	(-0.164, 0.186)	0.049	(-0.185, 0.283)
SEW(n)	-0.104	(-0.194,-0.013)	-0.007	(-0.083, 0.069)
SWD(o)	-0.657	(-0.852,-0.462)	-0.506	(-0.745,-0.267)
SWW(p)	-0.381	(-0.594,-0.169)	-0.075	(-0.283, 0.133)
WN(q)	-0.268	(-0.449,-0.087)	-0.253	(-0.515, 0.009)
WSE(r)	-0.397	(-0.644,-0.150)	-0.389	(-0.807, 0.029)
WSW(s)	-0.239	(-0.639, 0.160)	-0.162	(-0.939, 0.615)

It is clear that using the North-West region has been somewhat misleading in the results we expected to what have actually been found. While the North-West suggests there has been a positive trend in the data, the rest of the country suggests that there is in fact no trend or a negative one.

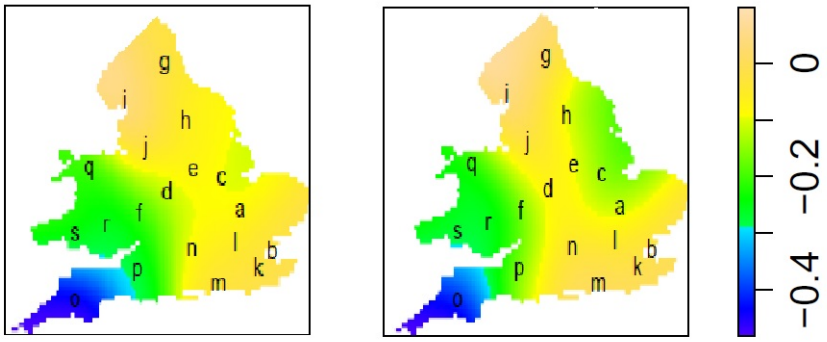

Figure 4.1: Colour density plots of Hourly values at 95% threshold (left) and 99% threshold (right)

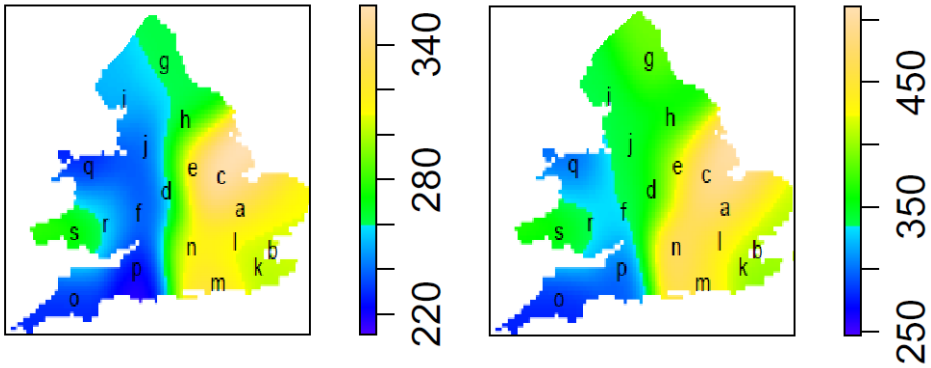

Figure 4.2: 100 year return levels of Hourly values at 95% threshold (left) and 99% threshold (right)

Figure 4.1 shows the mean slope of the sub-region in a colour density plot, however the software available isn't advanced enough for a high resolution image, but it still shows to an extent what is happening. As you can see the south west of the uk has had the most change in rainfall over the years while Wales and the east coast also have a slight negative slope. We can also form a return level plot for these two thresholds and this is given in Figure 4.2. This shows the 100 year return levels from 2011. So in 2011 we would expected to see weather this extreme with probability 1/100. Looking at these plots it appears that the weather gets more extreme as we move from the west coast to the east coast and the least extreme weather is in the south-west. This fits in with Figure 4.1 as the south west is suggested to have have a decrease in extreme weather. We will now investigate what happens if we increase the the length of time for each recording.

4.2 Daily data

Now that we have evidence to suggest that the amount of extreme weather we expect to see ever hour changes over time, we want to investigate whether this change applies over a longer observation time. We will check this by summing over 24 hour periods and applying the GPD with time. this raises the problem of missing values. This wasn't a problem before because the threshold removed missing values as they were all denoted -999, but if a 24 hour period has a missing value it will surely be dragged below the threshold and excluded even if it has extreme values in it. We will therefore set a limit of how many values we will allow to be missing then set them all the 0 and scale by 24/(number of missing values). The number of missing values is again arbitrary so we have chosen 4 and the R code for this is given below:

```
data2=data1[data1[,1]>-999*4,]
# removes all data with more that 4 missing values
for(i in 1:length(data2[,1])){
  if(data2[i,1]>(-999)&data2[i,1]<0)data2[i,1]=
    ((data2[i,1]+999)*(24/23))
} #adds on 999 and scales by 24/23
for(i in 1:length(data2[,1])){
  if(data2[i,1]>=(2*-999)&data2[i,1]<(-999))data2[i,1]=
    ((data2[i,1]+2*999)*(24/22))
} #adds on 2*999 and scales by 24/22
for(i in 1:length(data2[,1])){
  if(data2[i,1]>=(3*-999)&data2[i,1]<(2*-999))data2[i,1]=
    ((data2[i,1]+3*999)*(24/21))
} #adds on 3*999 and scales by 24/21
for(i in 1:length(data2[,1])){
  if(data2[i,1]>=(4*-999)&data2[i,1]<(3*-999))data2[i,1]=
    (data2[i,1]+4*999)*(24/20)
} #adds on 4*999 and scales by 24/20
```

Now we will again find the confidence intervals when the threshold is at 95% and 99% along with a colour density plot. However this time we experianced a few problems, namely convergence and scale. When summing over 24 hours we will therefore have less data to work with and fewer still when the threshold is set. This can lead to a lack of data in some sites and the GPD wont converge. Since the function is built in to the *ismev* package we are unable to change the starting values to help achieve convergence so at this stage we must remove these sites to stop them distorting our confidence intervals. Also, with the colour density plots, we can no longer use the same scale for both graphs as there is a set colour chart and with larger values the spread of the slope means is greater leading to different values being

shown with different colours. This is shown in Figure 4.3 and the means and confidence intervals are given below:

	Daily 95% threshold Mean	Daily 99% threshold Mean
AC(a)	-0.461 (-1.164 , 0.242)	0.674 (-2.697 , 4.045)
AE(b)	-0.386 (-0.833 , 0.062)	-1.921 (-3.361 ,-0.480)
AN(c)	0.012 (-3.897 , 3.921)	-1.013 (-2.951 , 0.924)
MC(d)	0.339 (-0.003 , 0.681)	-1.882 (-5.060 , 1.295)
ME(e)	-0.846 (-1.989 , 0.297)	-0.541 (-2.360 , 1.277)
MW(f)	-0.961 (-2.792 , 0.869)	-0.185 (-1.681 , 1.311)
NENE(g)	0.110 (-1.417 , 1.636)	0.975 (-2.807 , 4.756)
NEY(h)	-2.574 (-4.282 ,-0.866)	1.411 (-2.497 , 5.319)
NWN(i)	1.107 (0.660 , 1.554)	0.621 (-0.880 , 2.122)
NWS(j)	0.769 (0.477 , 1.060)	0.850 (-0.711 , 2.411)
SEK(k)	0.372 (-0.125 , 0.869)	-0.286 (-1.514 , 0.941)
SEN(l)	-0.397 (-0.664 ,-0.131)	-1.068 (-1.872 ,-0.264)
SES(m)	0.848 (-1.020 , 2.715)	1.032 (-0.967 , 3.032)
SEW(n)	0.203 (-0.118 , 0.524)	0.883 (0.191 , 1.575)
SWD(o)	0.893 (-0.960 , 2.747)	6.764 (1.518 , 12.010)
SWW(p)	-6.827 (-17.746, 4.093)	-4.600 (-12.354, 3.154)
WN(q)	-1.667 (-7.870 , 4.535)	9.568 (-17.019, 36.156)
WSE(r)	-3.312 (-6.190 ,-0.435)	-5.502 (-12.791, 1.787)
WSW(s)	-4.832 (-19.458, 9.793)	-1.972 (-18.670, 14.726)

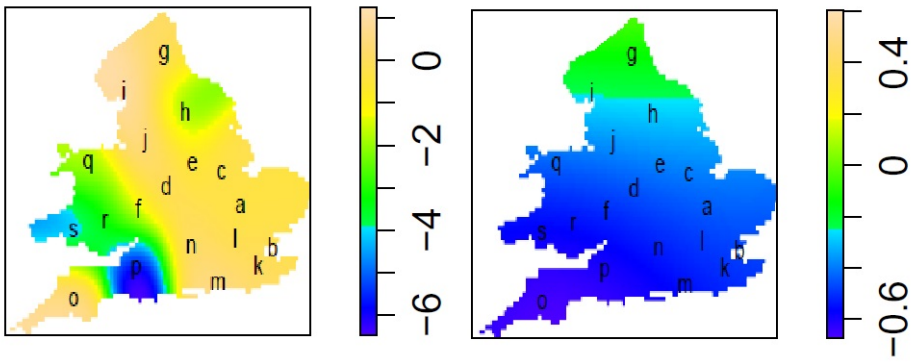

Figure 4.3: Colour density plots of Daily values at 95% threshold (left) and 99% threshold (right)

As you can see from Figure 4.3, the scales are very different and so could not be shown on the same scale effectively. Looking at what this figure actually tells us, at daily totals the data is showing more negative slopes on average but looking at the table, most confidence intervals contain zero so we cannot say with any confidence that the slope is changing over time. You

may have also noticed that the software to create these colour density plots isnt performing as well as we would have liked, with some of the colours not matching up to what the data tell us it should be (e.g. Wales Northen should be around 9 on the right plot but is shown to be nearer -0.5). Again this is just down to the lack of advancement in the software. It is interesting to see that the trend appears to be disappearing as the time interval increases. This suggests that although the level of extremes seems to be decreasing at the hourly level, we are not detecting this trend at the daily level. We shall investigate one more time interval later to see if the pattern continues.

Now we shall check the 100 year return plot for these daily totals. This is given in Figure 4.4. It is surprising to discover that now the east coast appears to have the least extreme weather while the west coast has the most extreme. So the east is expected to have the return of rain per hour but the least per day and the opposite for the west.

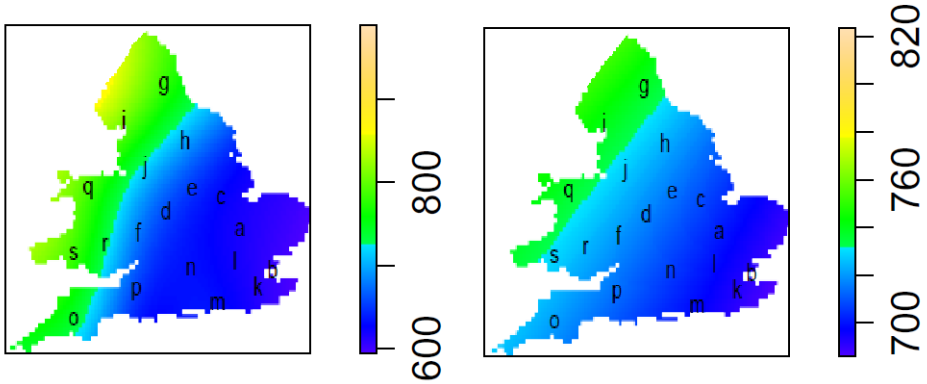

Figure 4.4: 100 year return level plots of Daily values at 95% threshold (left) and 99% threshold (right)

4.3 Aggregating over 5 days

The final time interval we have chosen to investigate is 5 days. This will be 120 hours of values so summing over this many values will potentially give us a large number of missing values so we have decided to discard any with more than 20 values missing (as this is 1/6 and that appears to be the magic number) and as before we will scale the values to show what we would have expected the totals to be. Once again the convergence is an issue so we will remove any that fail.

	5 Days 95% threshold Mean		5 Days 99% threshold Mean	
AC(a)	-0.518	(-3.476 , 2.440)	-10.548	(-28.674, 7.578)
AE(b)	-3.578	(-5.974 ,-1.181)	-4.999	(-11.611, 1.614)
AN(c)	-0.032	(-1.791 , 1.728)	-1.400	(-5.511 , 2.711)
MC(d)	0.027	(-2.169 , 2.223)	0.383	(-2.096 , 2.862)
ME(e)	-0.872	(-4.616 , 2.872)	-4.665	(-32.053, 22.722)
MW(f)	-0.007	(-4.051 , 4.038)	-13.977	(-33.373, 5.419)
NENE(g)	-0.007	(-7.707 , 7.693)	5.079	(-3.608 , 13.766)
NEY(h)	-3.523	(-8.713 , 1.667)	-0.839	(-19.418, 17.740)
NWN(i)	3.531	(1.301 , 5.761)	-8.763	(-43.753, 26.228)
NWS(j)	5.907	(0.227 , 11.587)	7.991	(-0.918 , 16.900)
SEK(k)	-1.410	(-4.362 , 1.542)	-4.948	(-11.313, 1.417)
SEN(l)	-1.756	(-2.871 ,-0.642)	1.689	(-6.018 , 9.396)
SES(m)	1.874	(-0.593 , 4.340)	2.583	(-4.530 , 9.695)
SEW(n)	2.227	(-1.064 , 5.517)	9.359	(0.655 , 18.062)
SWD(o)	-4.585	(-11.287, 2.117)	-1.052	(-11.553, 9.449)
SWW(p)	-13.516	(-24.720,-2.311)	-14.938	(-50.006, 20.129)
WN(q)	-1.841	(-11.019, 7.338)	-18.905	(-41.769, 3.958)
WSE(r)	-3.124	(-9.318 , 3.070)	-5.266	(-21.589, 11.057)
WSW(s)	-0.1535	(-12.754, 12.447)	9.359	(0.655 , 18.062)

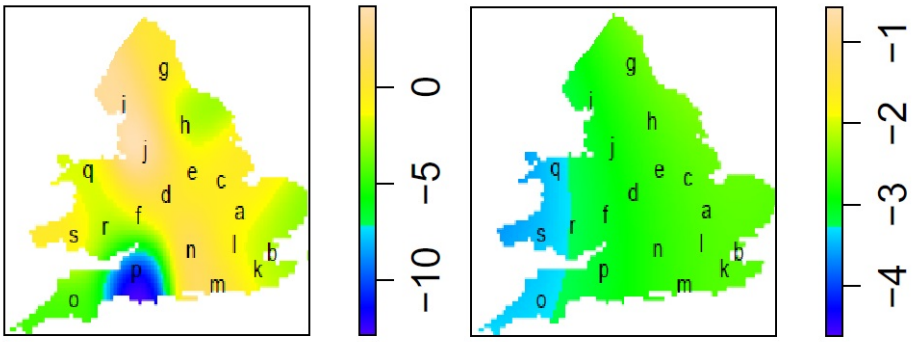

Figure 4.5: Colour density plots of 5 day aggregated values at 95% threshold (left) and 99% threshold (right)

Looking at the confidence intervals and Figure 4.5 the majority of sub regions now include 0 which suggests that there is virtually no change in slope over the period at the 5 days level. This confirms what was stated in Section 4.2, that although there appeared to be a change at hourly values, there is little to no change over longer periods. It is also worth noting that the confidence intervals are a lot larger than they were at the hourly level suggesting that there are huge ranges of values recorded and when there is little data at a site, the recorded values don't necessarily represent what is actually happening. Finally, Figure 4.6 shows the 100 year return level plots for these aggregated

5 day values. These two plots show contrasting results, with the threshold at 95% showing that the south-west is expected to have the least amount of rain and the amount gradually increases as you move towards the north-east. However, the 99% threshold plot shows the least in the east and inceasing as you move west. This shows that the choice of threshold is clearly very important.

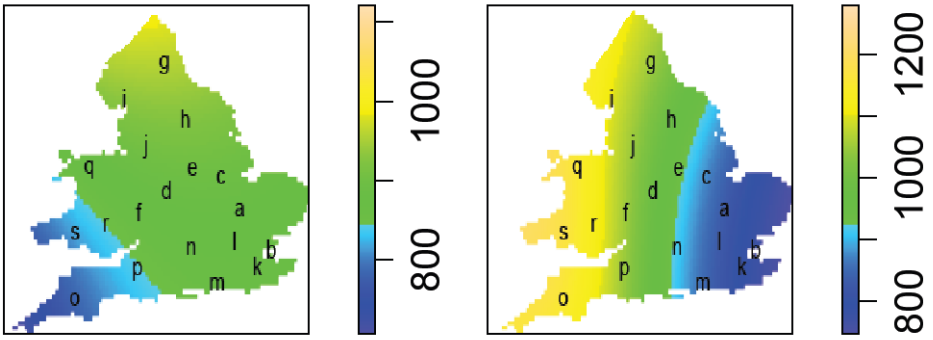

Figure 4.6: 100 year return level plots of 5 day aggregated values at 95% threshold (left) and 99% threshold (right)

Chapter 5

Conclusion

In conclusion, when working with extreme data we cannot use elementary methods such as simplistic counting as these lead to misleading results or in some cases fail to produce a result. This is why we use the Generalised Extreme Value distribution as it has been specifically designed to only deal with the maximum values and uses them to predict beyond the scope of the data. Also the Generalised Pareto distribution goes that one step further and allows us to work with all extreme values above a threshold while keeping the independence assumption through various methods, such as declustering. Each of these methods also allow to easily adapt from stationary to non-stationary to investigate the presence of trends.

The data set we have been working with is extremely large, with there being potentially 63 years worth of hourly values from over 1200 sites. This is actually larger than the data set used in the Flood Estimation Handbook. Of course we found out that many sites did not actually have all of this data and complications arose with uncooperative sites (such as failure to converge mainly down to lack of data) but there was still plenty to work with.

As for the results, we found some quite interesting patterns in regards to the final chapter. The return level plots told us that for hourly results, the east coast is expected to see larger rainfall than the west coast, but this is reversed (to an extent) at the daily and aggregated 5 days totals. As for the trends in the data, the data has shown us that when the threshold is exceeded, the values tend to be lower per hour now than they were in the past, suggesting a negative trend. However when the length of the recordings are increased to 24 and 120 hours, this negative trend becomes less noticeable and in some case disappears completely. This implies that although the hourly rainfall is less extreme now, the length of these "'storms"' has increased so on average we see the same amount of rainfall. This theory is known by Haby [2010] as a change in precipitation from convective (short and intense)

to dynamic (long and gentle) and is one way NASA describes how global warming will change the planet. [Riebeek, 2010]

further investigation into this area could be to test the robustness of the methods. Throughout the analysis we worked with values chosen on ac hoc (such as amount of data we allowed to be missing and size of threshold) and didn't investigate whether or not these arbitrarily chosen values affected the behavior of the data. So testing different values for these would have helped to confirm our calculations. We also skimmed over goodness-of-fit and the convergence issue, where we could have investigated further about choice of threshold and even created our own function to fit the distributions which would allow us to choose our own starting values. Another idea would be to take a Bayesian approach to the dataset, perhaps looking at hierarchical models and random effects as well as Monte-Carlo Markov Chain methods.

Appendix A

GEV code

```
hourgev=function(x){
  print(x)
  data=read.table(x)[,7]
  totmiss=matrix(0,ncol=2,nrow=length(data)/24)
  maxmiss=matrix(0,ncol=2,nrow=63)
  max=vector()
  j=1
  for(i in 1:(length(data)/24)){
    totmiss[i,2]=sum(data[seq((24*i)-23,(24*i))]>=0)
  }
  for(i in 1:length(data)){
    if(data[i]<0)data[i]=0
  }
  for(i in 1:(length(data)/24)){
    totmiss[i,1]=sum(data[seq((24*i)-23,(24*i))])
  }
  for(i in 1:63){
    maxmiss[i,1]=max(totmiss[seq(leaps[i],leaps[i+1]-1),1])
    maxmiss[i,2]=sum(totmiss[seq(leaps[i],leaps[i+1]-1),2])
    if(maxmiss[i,2]>7300){
      max[j]=maxmiss[i,1]
      j=j+1
    }
  }
  if(length(max)>=5){
    fit=gev.fit(max)
    gev.diag(fit)
  }
  else if(length(max)<5)
  {return(1:9)}
}
hourgev("Arnfield_Reservoir_LOG.txt")
```

```
yearfit=function(x){
  data=read.table(x)[,7]
  totmiss=matrix(0,ncol=2,nrow=length(data)/24)
  maxmiss=matrix(0,ncol=2,nrow=63)
  max=vector()
  years=vector()
  year=1948
  j=1
  k=1
  for(i in 1:(length(data)/24)){
    totmiss[i,2]=sum(data[seq((24*i)-23,(24*i))]>=0)
  }
  for(i in 1:length(data)){
    if(data[i]<0)data[i]=0
  }
  for(i in 1:(length(data)/24)){
    totmiss[i,1]=sum(data[seq((24*i)-23,(24*i))])
  }
  for(i in 1:63){
    year=year+1
    maxmiss[i,1]=max(totmiss[seq(leaps[i],leaps[i+1]-1),1])
    maxmiss[i,2]=sum(totmiss[seq(leaps[i],leaps[i+1]-1),2])
    if(maxmiss[i,2]>7300){
      max[j]=maxmiss[i,1]
      j=j+1
      years[k]=year
      k=k+1
    }
  }
  standyears=(years-mean(years))/sd(years)
  fit=gev.fit(max,ydat=matrix(standyears,ncol=1),mul=c(1))
  return(c(fit$conv,fit$mle,fit$se))
}
```

Appendix B

GPD code

```
fit=function(dataset){
  full=read.table(dataset)
  time=seq(1,length(full[,7]))
  data=cbind(full,time)
  start=which(data[,7]!=-999)[1]
  missing=which(data[,7]!=-999)
  finish=missing[length(missing)]
  data1=subset(data[start:finish,])
  threshold=quantile(data1[data1[,7]!=(-999),7],.95)
  cluster10values=function(dataset,threshold){
    x=list()
    z=list()
    j=1
{
  for(i in (11):length(dataset)){
    if(dataset[i-10]>threshold
       & dataset[i-9]<=threshold & dataset[i-8]<=threshold
       & dataset[i-7]<=threshold & dataset[i-6]<=threshold
       & dataset[i-5]<=threshold & dataset[i-4]<=threshold
       & dataset[i-3]<=threshold & dataset[i-2]<=threshold
       & dataset[i-1]<=threshold
       & dataset[i]<=threshold){
      x=max(dataset[j:i])
      ifelse(i !=length(dataset), j<-i+1, NA)
      z=c(z,x)}}}
return(z)}
cluster.peaks.values=as.numeric(cluster10values(data1[,7],
                                                threshold))
cluster10times=function(dataset,threshold,time){
  x=list()
  z=list()
  t=list()
  j=1
```

```
{
  for(i in (11):length(dataset)){
    if(dataset[i-10]>threshold
       & dataset[i-9]<=threshold & dataset[i-8]<=threshold
       & dataset[i-7]<=threshold & dataset[i-6]<=threshold
       & dataset[i-5]<=threshold & dataset[i-4]<=threshold
       & dataset[i-3]<=threshold & dataset[i-2]<=threshold
       & dataset[i-1]<=threshold
       & dataset[i]<=threshold){
      x=max(dataset[j:i])
      t=max(time[j:i])-10
      ifelse(i !=length(dataset), j<-i+1, NA)
      z=c(z,t)
    }
  }
}
return(z)
}
cluster.peaks.times=as.numeric(cluster10times(data1[,7],
                                        threshold,data1[,8]))
cluster.peaks=cbind(cluster.peaks.values,cluster.peaks.times)
fit = gpd.fit(cluster.peaks[,1],threshold,ydat=matrix(
  cluster.peaks[,2]/8766,ncol=1),sigl=c(1))
return(c(fit$conv,length(data1[,7])/8766,fit$thresh,
         fit$nexc,fit$mle[1],fit$se[1],fit$mle[2],
                        fit$se[2],fit$mle[3],fit$se[3]))
}
```

Bibliography

Stuart Coles. *An Introduction to Statistical of Extreme Values*. Springer Series in Statistics, 2001.

Lee Fawcett. Topics in statistics: Environmental extremes. Course Module, Newcastle University, 2013.

Jeff Haby. Dynamic precip vs. convective precip, March 2010. URL `http://www.theweatherprediction.com/habyhints/336/`.

A.F. Jenkinson. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171, April 1955.

Richard Von Mises. La distribution de la plus grande de n valeurs. *American Mathematical Society*, II:271–294, 1954.

James Picklands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131, 1975.

Holli Riebeek. How will global warming change earth?, June 2010. URL `http://earthobservatory.nasa.gov/Features/GlobalWarming/page6.php`.

Ronald Fisher & Leonard Tippett. On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24(02):180, 1928.