University of Venda

MODELLING FLOOD HEIGHTS OF THE LIMPOPO RIVER AT BEITBRIDGE BORDER POST USING EXTREME VALUE DISTRIBUTIONS

By

Robert Kajambeu

SUBMITTED IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
UNIVERSITY OF VENDA
THOHOYANDOU, SOUTH AFRICA
OCTOBER 2016

UNIVERSITY OF VENDA

DEPARTMENT OF

STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematical and Natural Sciences for acceptance a thesis entitled "**Modelling flood heights of the Limpopo river at Beitbridge border post using extreme value distributions** " by **Robert Kajambeu** in fulfillment of the requirements for the degree of **Master of Science**.

Dated: October 2016

Supervisor: _____
Dr. Caston Sigauke

Co-Supervisor: _____
Dr. Alphonce Bere

_____

UNIVERSITY OF VENDA

Date: **October 2016**

Author: **Robert Kajambeu**

Title: **Modelling flood heights of the Limpopo river at Beitbridge border post using extreme value distributions**

Department: **Statistics**

Degree: **M.Sc.** Convocation: **October** Year: **2016**

Permission is herewith granted to University of Venda to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

Abstract

Haulage trucks and cross border traders cross through Beitbridge border post from landlocked countries such as Zimbabwe and Zambia for the sake of trading. Because of global warming, South Africa has lately been experiencing extreme weather patterns in the form of very high temperatures and heavy rainfall. Evidently, in 2013 traffic could not cross the Limpopo River because water was flowing above the bridge.

For planning, its important to predict the likelihood of such events occurring in future. Extreme value models offer one way in which this can be achieved. This study identifies suitable distributions to model the annual maximum heights of Limpopo river at Beitbridge border post. Maximum likelihood method and the Bayesian approach are used for parameter estimation.

The r -largest order statistics was also used in this dissertation. For goodness of fit, the probability and quantile- quantile plots are used. Finally return levels are calculated from these distributions.

The dissertation has revealed that the 100 year return level is 6.759 metres using the maximum likelihood and Bayesian approaches to estimate parameters. Empirical results show that the Fréchet class of distributions fits well the flood heights data at Beitbridge border post.

The dissertation contributes positively by informing stakeholders about the socio-economic impacts that are brought by extreme flood heights for Limpopo river at Beitbridge border post.
Keywords: Extreme value theory, Bayesian approach, r-largest order statistics.

Declaration

I declare that the thesis which is hereby submitted for the qualification of Master of science in Statistics at the University of Venda, is my own independent work and has not been handed in before for a qualification at/in another University/Faculty. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references. I further cede copyright of the thesis to the University of Venda.

Signature.....Date.....

Acknowledgements

In this dissertation I acknowledge so many people whose efforts has made this dissertation a success.

Above all, my supervisor Dr. Caston Sigauke for his excellent and perpetual support with relevant material and also for showing interest throughout the compilation of this project. I also extend many thanks to my Co-Supervisor Dr. Alphonce Bere for the guidance that he offered me throughout my research.

The chairman of the department of statistics Dr Kyei Kwabena cannot go unmentioned for the encouragement that led me to finally finish my studies. Mr Lebopa Victor Kgorompe my principal is thanked for giving me time for presentations during the proposal stage. Thank you sir for this unwavering support in my studies.

My friends and colleagues Emmanuel Gyamfi and Retang and all staff members of the statistics Department at the University of Venda for their moral support and encouragement which has helped me greatly.

Finally gratitude goes to my wife Shamiso and child Ruth whose loving support has made all the difference. My parents for their encouragement to learn since birth. Thank you for your support my father and my mother, finally I am there .

My wife Shamiso and child Ruth Nenyasha.

Table of Contents

List of Tables

List of Figures

Abbreviations

CDF	Cumulative Distribution Function
EVI	Extreme Value Index
EVT	Extreme Value Theory
GEVD	Generalized Extreme Value Distribution
GPD	Generalized Pareto Distribution
MLE	Maximum Likelihood Estimate
MCMC	Markov Chain Monte Carlo
PDF	Probability Density Function
POT	Peak Overthreshold
QQ	Quantile to Quantile

Symbols

$G_\gamma(x)$	extreme value distribution function for maxima
$W_\xi(x)$	generalized Pareto distribution function
x_p	quantile function
$\pi(\theta)$	prior distribution
$\pi(x\|\theta)$	likelihood function
$\pi(\theta\|x)$	posterior distribution
$\pi(x_{n+1}\|x)$	posterior predictive density of a future observation x_{n+1}
τ	a threshold
log	natural logarithm
γ	Extreme Value Index (EVI) for the generalized extreme value distribution
ξ	EVI for the generalized Pareto distribution
$F^{(\tau)}$	empirical exceedance distribution function of τ
Ψ	digamma function
γ gamma function	
$x_{0.5}$	symbol for median
x_m	symbol for mode

Distributions

Generalized Extreme Value Distribution, $\xi \neq 0$

Notation	$X \sim \text{GEVD}(\xi, \mu, \sigma)$
EVI	ξ
Distribution function	$G_\xi(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$
Density function	$g_\xi(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$
Range	$\left\{x\middle\vert 1+\xi\left(\frac{x-\mu}{\sigma}\right)>0\right\}$
Parameters	$\xi \neq 0, -\infty < \mu < \infty, \sigma > 0$

Generalized Extreme Value Distribution, $\xi = 0$

Notation	$X \sim \text{GEVD}(0, \mu, \sigma)$
EVI	0
Distribution function	$G_\xi(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$
Density function	$g_\xi(x) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$
Range	$-\infty < x < \infty$
Parameters	$-\infty < \mu < \infty, \sigma > 0$

Generalized Pareto Distribution, $\xi \neq 0$

Notation	$X \sim \text{GPD}(\xi, \sigma)$
EVI	ξ
Distribution function	$W_\xi(x) = 1 - \left(1 + \frac{\xi(x-\tau)}{\sigma}\right)^{-\frac{1}{\xi}}$
Density function	$w_\xi(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\tau)}{\sigma}\right)^{-(1+\frac{1}{\xi})}$
Range	$x > \tau$, for $\xi > 0$
	$\tau < x < \tau - \frac{\sigma}{\xi}$, for $\xi < 0$
Parameters	$\xi \neq 0, \sigma > 0$

Chapter 1

Introduction

Floods have become a common natural disaster in Southern Africa. Due to its geographical position, South Africa is one of the countries that have faced challenges in terms of floods. The Limpopo river is the second largest river in Southern Africa after the Zambezi and drains to the Indian ocean. Unlike other large rivers such as the Zambezi which have large dams, the Limpopo river does not have any large dams indicating that the flow is regulated (Maposa *et al.*, 2014). The hydrology of the Limpopo river is famously known by a cycle of rainfall that extends from October of the past year until April of the coming year with peak monthly totals in February. The dry season runs from May to September (World Meteorological Organisation, 2012). In this dissertation extreme value models are applied to the occurences and magnitudes of the flood heights of the Limpopo river at Beitbridge border post.

Coles (2001) defines extreme value theory (EVT) as the study of very rare events. According to Coles (2001); Gilli and Kellezi (2006), there are two fundamental methods used to identify extremes in (EVT) namely the block maxima approach and the peaks-over-threshold (POT) approaches. The first method solely concentrates on the distribution of block maxima whereas the latter identifies realised large values over

a high threshold that is called (GPD). EVT is applicable in the field of finance, hydrology and insurance only to mention a few.

1.1 Statement of the Problem

The Beitbridge is constructed on the Limpopo river close to Beitbridge Border Post. The bridge is important because it enhances transportation of goods from South Africa to the landlocked countries to the north and from these countries to South Africa.Cross border traders use this bridge to cross to and from South Africa.

With global warming and climate change becoming more and more of a reality, South Africa is experiencing a gradual, yet steady, change in climate. Temperatures have risen significantly over the last 60 years, and are predicted to continue this rising trend, with a rise in temperature of 1-2^{0}C expected in coastal regions, and 3-4^{0}C expected in interior regions by 2050 (Sithole and Murewi, 2009). An increase of 3-4^{0}C in coastal regions is predicted and 6-7^{0}C in interior regions is predicted by 2 100, (Sithole and Murewi, 2009). Rainfall patterns are also shifting, although this is more variable and unpredictable.

In the recent past, Southern Africa has experienced a lot of cyclones which bring with them a lot of rainfall. The tropical cyclones that have occurred so far in Southern Africa are Astrid (January 1958), Claude (January 1966), Caroline (February 1972), Eugenie (February 1972), Danae (January 1976), Emilie (February 1977), Kolia (March 1980), Justine (March 1982), Domoina (January 1984), Imboa (February 1984), Eline (February 2000) and Japhet (February- March 2003), (Sithole and Murewi, 2009).

According to a newspaper article (ZOUNET 30 January 2013) the traffic failed to cross the Beitbridge because of floods in 2013. Considering the changing climate,

there is need to model the flood heights of Limpopo river at Beitbridge Border Post with the aim of assessing the chances of extreme future flood heights. Such an analysis will be used to compute the return period. A return period is the average length of time in years for an event (e.g. flood or river level) of given magnitude to be equaled or exceeded once. A return level is defined as a value that is expected to be equaled or led or exceeded on average once every interval of time (T) with a given probability p.

1.2 Aim and objectives

1.2.1 Aim

The main aim of this study is to model the annual maximum flood heights of the Limpopo river at Beitbridge Border Post, using extreme value distributions.

1.2.2 Objectives

The main objectives of the study are to:-

(i) identify a suitable generalised extreme value (GEVD) distribution to model the annual maximum flood heights of the Limpopo river at Beitbridge,

(ii) use generalised Pareto distribution (GPD) to model the flood heights,

(iii) estimate the parameters of the identified GEVD distribution using maximum likelihood estimation and Bayesian estimation,

(iv) estimate the return levels of the flood heights.

1.3 Significance of the study

The study provides stake holders such as Government agencies from South Africa and Zimbabwe with a flood risk assessment. By estimating return periods and return

levels, preparedness about flood forecasting is enhanced.

1.4 Dissertation structure

The rest of the dissertation is structured as follows: **Chapter 2**, presents a literature survey on the application of extreme value distributions. The methods used in the modelling of flood heights for the Limpopo river at Beitbridge Border Post are discussed in **Chapter 3**. In **Chapter 3** we also give theory that is behind estimation of parameters using the method of maximum likelihood and Bayesian approach. In the same Chapter, the r largest order statistics and their properties are outlined. The analysis and research findings have been summarised in **Chapter 4**. **Chapter 5** presents the conclusions. We used the R package for data analysis and the R codes used in the entire of this analysis have been put on appendix.

Chapter 2

Literature survey

2.1 Introduction

This section reviews relevant studies that have been previously conducted.

Tadesse (2011) applies both the GEVD and the generalised Pareto distribution (GPD) to model floods in Southern Africa catchments. The countries covered are Namibia, Malawi, Zimbabwe and South Africa. In the study, 459 stations from these selected countries were considered. Tadesse uses the method of L moments for estimating parameters and for threshold selection, he uses the mean excess plot. In this research the generalised Pareto distribution and the and the Gumbel class of the generalised extreme value distributions were found to be suitable models for floods in Southern Africa catchments.

Kamwi (2005) proposes and fits the extreme value distributions to the flood heights at Katima Mulilo using the method of maximum likelihood. The study seeks to examine the uncertainty of the estimated parameters and evaluate the goodness of fit of the model that has been identified. The study uses 39 annual maximum heights of the

river flow at Katima Mulilo. The empirical results from this dissertation suggested that a Gumbel class of the generalised extreme value distributions model provided a good fit.

In another study, De Waal (2012) assesses the changes in extreme rainfall values over time across the Western Cape, South Africa. By using a GPD, the study examines the changes in return levels across the Western Cape region for the periods 1900-1954 and 1955-2010. De Waal's study uses daily rainfall extracted from South Africa Weather Services data. Of one hundred and thirty seven stations which are investigated, sixty two percent show an increase in 50-year return level, twenty two percent show a decrease in 50-year return levels. Sixteen percent of the stations display little change in rainfall intensity over time. The results also indicate an increase in frequency of intense rainfalls in the later half of the 20th and early 21st century.

Kagoda (2006) carried out a study in South Africa based on daily rainfall data for fifteen rain gauge stations selected across the country. The author uses the Bayesian approach to data analysis. The (GPD) is used to model the excesses over a threshold chosen from a mean residual life plot of the rainfall data. The shape and scale parameters of the GPD are represented by a joint distribution which is sequentially modified by the rainfall data resulting in a posterior distribution from which a Markov chain is generated using the Gibbs sampler.

A study similar to ours on EVT is by Singo *et al.* (2012)on the Luvuvhu river in Limpopo province in South Africa. The authors select eight stations with fifty years hydrological data to analyse flood frequencies in this catchment. From this study, the Gumbel class of distribution and log Pearson type 3 provide the best fit.

Kartz (2010) apply the block maxima and the peaks over threshold approach in their study. The authors use data for the period 1900-1999 from Fort Collins in United States of America. This data shows annual maximum daily precipitation amount derived from the original daily data. The Gumbel distribution is found to be the best

fitting distribution to the data.

Melice and Reason (2007) use the longest available daily rainfall series at George (from 1941 to 2006), in the vicinity of which the most destructive floods are observed, together with an extreme value model to estimate the return period of such an extreme event. According to this model, the greatest annual maximum daily rainfall of 230mm observed at the town on the first of August 2006 has a return period of 1222 years, whereas the second largest observed annual maximum daily rainfall (132mm in September 1964) has a return period of 23 years. This shows that the August 2006 extreme rainfall at George can be considered as a particularly rare event.

Salarpour *et al.* (2012) use the annual flood peaks selected from the maximum flow in each water year (June-July). The study was carried out from Rantau Panjang gauging station. The annual flood peaks that were used are from Johor river in Malaysia. Five probability distributions namely Gamma, Weibull, Exponential, Gumbel and Generalised Extreme Value were fitted to model the distribution of flood parameters. In testing the goodness of fit, the Kolmogorov Smirnov and chi squared tests were used. The conclusion that is derived by the authors indicate that at five percent level of significance, all the models mentioned above can be used to model the distribution of peak flow, duration and volume. However, the Exponential distribution was the most suitable model when tested with the Kolmogorov Smirnov test where as the GEVD was found to be the best fit after performing a chi squared test.

Pocernich (2002) examines the use of extreme value statistics in predicting the probability of rare flood events. The researcher uses the GPD to model high flows as well as fitting the GEVD. He had to fit a linear model whereby μ is allowed to vary with respect to time. The empirical results indicate that the parameter estimates did not vary significantly with time.

There is a wide gap between our study and all the studies mentioned above in the sense that all the researchers mentioned above have not done modelling of the annual

maximum flows of Limpopo river before. This is a new station that has never been examined by other researchers even though we are using the same methodology in analysing our data. Notably, most researchers used the generalised extreme value distributions and only a few used the GPD. In this study we fit the generalised extreme value distributions that is Gumbel, Fréchet and Weibull also known as type 1, type 2, type 3 respectively and the GPD to estimate the return levels of Limpopo river at Beitbridge Border Post.

2.2 Concluding remarks

Modelling of flood heights have been reviewed in this chapter. Models applied by different researchers have been also reviewed in this chapter. The tests conducted by different authors also have been discussed. Sometimes, it is very important for us to move away from the average thinking and we then try by all means to exploit the information that we get from extreme value distributions modelling. This knowledge that we get helps people in the field of hydrology to make informed decisions through flood heights modelling.

Chapter 3

Extreme value modelling

3.1 Introduction

This chapter introduces, the EVT techniques that will be used in this study. Stationary and non stationary GEVD for homogeneous sequences and time varying parameters respectively are going to be discussed for r largest order statistics where $(r \geq 1)$. For the case $r = 1$, we have the usual block maxima. There are two approaches that exist for practical extreme value analysis namely the block maxima approach and the peaks-over-threshold approach. Block maxima approach is defined by Anna Ferreira and De Haan (2015) as an approach in extreme value theory that consists of dividing the observation period into non-overlapping periods of equal size. The approach restricts attention to the maximum observation of each period. Coles (2001) defines the peaks-over-threshold as the method whereby we extract the peak values from a continuous record for any period during which values exceed a certain threshold. We normally resort to the use of r largest statistics in extreme value analysis due to limited amount of data. Estimation of parameters, model checking and diagnosis are

also going to be discussed.

3.2 Generalised extreme value distribution for block maxima

In this section we are going to discuss the generalised extreme value distribution for block maxima.

3.2.1 Generalised extreme value distribution for block maxima

The model focuses mainly on the statistically behaviour of $M_n = \max\{X_1, ..., X_n\}$, where $X_1, ..., X_n$ is a sequence of independent random variables having a common distribution F. If n is the number of observations in a year, then M_n corresponds to the annual maximum. In theory the distribution of M_n can be derived exactly for all values of n:

$$F^n = P_r\{M_n \leq z\} = P_r\{X_1 \leq z, ..., X_n \leq z\} = P_r\{X_1 \leq z\} \times ... \times P_r\{X_n \leq z\} = \{F(z)\}^n \tag{3.2.1}$$

One possibility is to use standard statistical techniques to estimate the value of F from observed data and substitute in equation (3.2.1), above. Unfortunately very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n where $X_1, X_2, ..., X_n$ is a sequence of independent, identically distributed random variables with common distribution function F which can be estimated on the basis of extreme data only. This is the same as the usual way of approximating the distribution of sample means by the way of normal distribution as justified by central limit theorem. We proceed by looking at the behaviour of F^n as $n \to \infty$. But this alone is not

enough: for any $z < z_+$ where z_+ is the upper end point of F, $F^n(z) \to 0$ as $n \to \infty$, therefore the distribution of M_n degenerates to a point mass of z_+. This difficulty is avoided by allowing a linear renormalisation of the variable M_n:

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{3.2.2}$$

for sequences of constants $\{a_n > 0\}$ and $\{b_n > 0\}$. According to Fisher and Tippet (1928), in their extremal types theorem (ETT), for an appropriate selection of$\{a_n\}$ and $\{b_n\}$, the distribution of properly normalised maxima of a sequence of independent and identically distributed random variables converges to one of the three extreme value distribution (EVD) families depending on the shape parameter α. The three cumulative distribution functions namely the Fréchet, Gumbel and Weibull have been given below.

Definition

Let $G(z) = P(Z \leq x)$ be the distribution of a random variable Z. Then the sequence (i)is said to converge weakly towards Z if the distribution functions of the random variables $\frac{M_n - b_n}{a_n}$ converge towards $G(x)$ at all points of continuity of $G(x)$.

Theorem 1

Let $\{a_1, a_2...\}$ and $\{b_1, b_2...\}$ with $a_i > 0$ be two sequences of real numbers with property that the sequence of the linearly transformed random maxima′

$$\left\{ \frac{M_n - b_n}{a_n}; n = 1, 2, ... \right\} \tag{3.2.3}$$

converges weakly towards a random variable Z with positive variance (i.e. Z is not degenerated). Then the weakly convergence of the sequence (i) towards Z is equivalent to the limit relation:

$$\lim_{n\to\infty}\left(F\left(\frac{z}{a_n}+b_n\right)\right)^n = G(x)$$

for all points x of continuity of $G(x)$ and $G(x)$ belongs to any of the functions listed under (3.2.4)below.

$$G(z) = \begin{cases} \exp\{-\exp[-\left(\frac{x-b}{a}\right)]\} & \text{if } -\infty < x < \infty; \\ \exp\{-\left(\frac{x-b}{a}\right)^{-\alpha}\} & \text{if } x > b; \\ \exp\{-[-\left(\frac{x-b}{a}\right)^{\alpha}]\} & \text{if } x \geq b, \end{cases} \tag{3.2.4}$$

In this case the first family is the Gumbel, the second is the Fréchet and lastly the third family is the Weibull distribution.

Von Mises (1936) unified the three distributions into a single generalised distribution with location parameter μ, scale parameter $\sigma > 0$ and shape parameter ξ of the form:

$$G_\xi(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}. \tag{3.2.5}$$

This is valid for $\{x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$, where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$.

3.2.2 Properties of the GEVD when $\xi = 0$

If we differentiate with respect to x we get the probability density function:

$$g(x) = \frac{1}{\sigma}\exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right]\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right] \tag{3.2.6}$$

In order to find the characteristic function, the Gumbel random variable X is first transformed into a standard random variable using $Z = \frac{X-\mu}{\sigma}$. The density of Z is then given by :

$p(z) = e^{-e^{-z}}e^{-z}$ that is for $-\infty < z < \infty$. The characteristic function of Z is determined to be :

$$\varphi z(t) = E[e^{itZ}] = \int_{-\infty}^{\infty} e^{itz} e^{-e^{-z}} e^{-z} dz = \int_{0}^{\infty} u^{-it} e^{-u} du = \Gamma(1 - it)$$

where $i = \sqrt{-1}$ and the characteristic function of $X = \mu + \sigma(Z)$, which follows the Gumbel distribution, is therefore given by:

$\varphi_X(t) = E[e^{itX}] = e^{it\mu}\Gamma(1 - it\sigma)$.

If we evaluate these derivatives at the interval $t = 0$, the first two moments of X are given as :

$E[X] = \mu - \sigma\Gamma^{(1)} = \mu + \sigma\gamma$

$E[X^2] = \mu^2 - 2\sigma\mu\Gamma^{(1)}(1) + \sigma^2\Gamma^{(2)}(1) = (\mu + \sigma\gamma)^2 + \sigma^2\frac{\pi^2}{6}$

Mathematically γ is approximated as 0.5772156649.... which is referred as the Euler's constant. It then follows up from above derivation that the variance is equal to:

$$Var(X) = \sigma^2\Psi'(1) = \sigma^2\frac{\pi^2}{6} \tag{3.2.7}$$

whereby $\Psi'(1)$ is the first derivative of the digamma function. In a more general case, the normal expected value of the random random variable $Z = \frac{X-\mu}{\sigma}$ which is given by Smith (1986) as:

$$E[Z^m \exp(cZ)] = (-1)^m\Gamma^{(m)}(1 + c) \tag{3.2.8}$$

This binds for $c > -1$. Furthermore, if we set $c = 0$ in equation (3.2.8), above, then it means that it is possible to obtain moments for the standard Gumbel random variable Z, that is:

$E[Z^m] = (-1)^m\Gamma^{(m)}(1)$,

which then can be eventually used to find the moments of the Gumbel random variable

X. For completeness sake, the median of the Gumbel distribution is given by:

$$\exp\left[\text{-exp}\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right] = 0.5 \tag{3.2.9}$$

which yields $x_{(0.5)} = \mu - \sigma\left(\ln(\ln 2)\right)$

The mode of X is solution of the equation (3.2.6)above which yields μ. For completeness, the mode of the GEVD is given as:

$$\frac{d}{dx}\left[\frac{1}{\sigma}\exp\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right]\right]\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right] = 0 \tag{3.2.10}$$

A summary of the Gumbel distribution is illustrated in table (3.1).

Table 3.1: Summary for the properties of a Gumbel distribution

Distribution Function	$G_\gamma(x)$=exp$\left[\text{-exp}\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right]$ for $\gamma = 0$
Density Function	$g_\gamma(x)$=$\frac{1}{\sigma}$exp$\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right]\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]$ for $\gamma \neq 0$
Characteristic Function	$e^{it\mu}\Gamma(1-it\sigma)$
Expected Value	$\mu + \sigma\gamma$
Variance	$\sigma^2\frac{\pi^2}{6}$
Median	$\mu - \sigma\left(\ln(\ln 2)\right)$
Mode	μ

3.2.3 Properties of the GEVD when $\xi \neq 0$

The shape parameter ξ has an impact on the support and tail behaviour of the GEVD. The support of GEVD is given by all x satisfying $\left\{x, 1+\xi(\frac{x-\mu}{\sigma}) > 0\right\}$.

When $\xi > 0$, the distribution function is skewed to the right and the support is

bounded to the left with $\mu - \frac{\sigma}{\xi} < x$. On the other hand if $\xi < 0$, the distribution is skewed to the left and the support is bounded to the right with $x < \mu - \frac{\sigma}{\xi}$. If we differentiate the cumulative density function for the GEVD $G(x)$ in equation (3.2.5) above we get the probability density function as:

$$g(x) = \frac{1}{\sigma}\exp\left\{ -\left[1+\xi(\frac{x-\mu}{\sigma})\right]^{\frac{-1}{\xi}}\right\}\left[1+\xi(\frac{x-\mu}{\sigma})\right]^{(\frac{1}{\xi}+1)}. \tag{3.2.11}$$

with support of the GEVD which is given by all x satisfying $\left\{x, 1+\xi(\frac{x-\mu}{\sigma}) > 0\right\}$ and the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. One way of calculating the moments for the GEVD is to first transform the Generalised extreme value random variable X into a standard Gumbel random variable. We achieve this by setting:

$$e^W = \left[1+\xi(\frac{X-\mu}{\sigma})\right]^{\frac{-1}{\xi}} \tag{3.2.12}$$

which eventually implies that:

$$W = \frac{1}{\xi}\ln\left[1+\xi(\frac{X-\mu}{\sigma})\right] \tag{3.2.13}$$

and

$$X = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}e^{W\xi} \tag{3.2.14}$$

Now after the transformation, it is easily seen that W is a standard Gumbel random variable.

Now the moment generating function of standard Gumbel distribution can be used to determine the moments of the GEVD. This moment generating function is of the form:

$M_W(t) = \Gamma(1-t)$.

We calculate the first two moments as follows:

$$E[X] = E\left[\mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}e^{W\xi}\right] \tag{3.2.15}$$

$$= \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}M_W(\xi) = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}\Gamma(1-\xi) \tag{3.2.16}$$

and also

$$E[X^2] = E\left[\left(\mu - \frac{\sigma}{\xi}\right)^2 + 2\left(\mu - \frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}e^{W\xi} + \frac{\sigma^2}{\xi^2}e^{2W\xi}\right] \tag{3.2.17}$$

$$= \left(\mu - \frac{\sigma}{\xi}\right)^2 + 2\left(\mu - \frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}M_W(\xi) + \frac{\sigma^2}{\xi^2}M_W(2\xi) \tag{3.2.18}$$

$$= (\mu - \frac{\sigma}{\xi})^2 + 2\left(\mu - \frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}\Gamma(1-\xi) + \frac{\sigma^2}{\xi^2}\Gamma(1-2\xi). \tag{3.2.19}$$

Higher moments can be calculated in a similar way. Note that the first moment of X only exists if $\xi < 1$, and that its second moment only exists if $\xi < \frac{1}{2}$. Generally, the n^{th} moment of X only exists if $\xi < \frac{1}{n}$.

Therefore, if $\xi < 0$, then GEVD has always finite moments, and if $\xi > 0$, then X has only moments of order less than $\frac{1}{\xi}$.

By using the first two moments, the variance of the generalised extreme value variable X can simply be calculated as follows:

$$\text{var X} = E[X^2] - (E[X])^2 = \frac{\sigma^2}{\xi^2}\left[\Gamma(1-2\xi) - (\Gamma(1-\xi))^2\right]. \tag{3.2.20}$$

For $b\epsilon\mathbb{R}$ and positive integers c and m, Tawn (1998) gives the normal expected value

for a standard Generalised extreme value random variable $Z = \frac{X-\mu}{\sigma}$ as:

$$E[Y] = (-\xi)^{c-m} \sum_{p=0}^{m} (-1)^P \binom{m}{p} \Gamma^{(c)}(2 + b\xi - p\xi) \tag{3.2.21}$$

In this case $Y = Z^m(1+\xi Z)^{-(\frac{1}{\xi}+b)}(\ln(1+\xi Z))^c$. The normal expected value that is given in equation (3.2.21) is used in order to determine the expectations of the second derivative of the likelihood function.

Generally, the expectation given in equation (3.2.21) above might also be used to determine the moments of the Generalised extreme value distribution. If for example we set $c = 0$ and $b = \frac{-1}{\xi}$ this implies that $Y = Z^m$ and

$$E[Y] = E[Z^m] = (-\xi)^{-m} \sum_{p=0}^{m} (-1)^p \binom{m}{p} \Gamma(1 - p\xi). \tag{3.2.22}$$

For completeness, the median of the GEVD is given as:

$$\exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}} \right\} = 0.5 \tag{3.2.23}$$

which yields

$$x_{(0.5)} = \mu + \frac{\sigma}{\xi}[(\ln 2)^{-\xi} - 1] \tag{3.2.24}$$

and the mode is derived as:

$$\frac{d}{dx}\left[\frac{1}{\sigma}\exp\left\{ -\left[1 + \xi(\frac{x-\mu}{\sigma})\right]^{\frac{-1}{\xi}} \right\}\left[1 + \xi(\frac{x-\mu}{\sigma})\right]^{(\frac{1}{\xi}+1)}\right] = 0 \tag{3.2.25}$$

which yields

$$\mu + \frac{\sigma}{\xi}[(1+\xi)^{-\xi} - 1] \tag{3.2.26}$$

Table (3.2) summarises the properties for the GEVD as explained above:

The corresponding p quantile function of a GEVD is given by:

Table 3.2: Summary of the properties for the GEVD

Distribution Function	$G_\gamma(x)=\exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}\right\}$ for $\gamma \neq 0$
Density Function	$g_\gamma(x)=\frac{1}{\sigma}\exp\left\{-\left[1+\xi(\frac{x-\mu}{\sigma})\right]^{\frac{-1}{\xi}}\right\}\left[1+\xi(\frac{x-\mu}{\sigma})\right]^{(\frac{1}{\xi}+1)}$ for $\gamma \neq 0$
Expected Value	$\mu-\frac{\sigma}{\xi}+\frac{\sigma}{\xi}\Gamma(1-\xi)$
Variance	$\frac{\sigma^2}{\xi^2}\left[\Gamma(1-2\xi)-\left(\Gamma(1-\xi)\right)^2\right]$
Median	$\mu+\frac{\sigma}{\xi}[(\ln 2)^{-\xi}-1]$
Mode	$\mu+\frac{\sigma}{\xi}[(1+\xi)^{-\xi}-1]$

$$x_p = \begin{cases} \mu-\frac{\sigma}{\xi}[1-\{-\log(1-p)\}^{-\xi}], & \text{for } \xi \neq 0; \\ \mu-\sigma\log(1-p)\}, & \text{for } \xi = 0, \end{cases} \tag{3.2.27}$$

3.2.4 Estimation of parameters for the GEV

In this study we will use both the maximum likelihood and Bayesian methods to estimate parameters.

Method of Maximum likelihood

Likelihood based methods of estimating parameters of the EVT models are more reliable compared to other methods for various reasons that include the adaptability to model change, Coles (2001).

Given a series of observed sample block maxima $x_1, x_2, ..., x_N$ which are assumed be realizations of independent random variables each having a GEV distribution with probability density function $f(x, \mu, \sigma, \xi)$, the likelihood function is given by:

$$L(\mu, \sigma, \xi | x_1, x_2, ..., x_N) = f(x_1, x_2, ..., x_N; \mu, \sigma, \xi) = \prod_{i=1}^{N} f(x_i, \mu, \sigma, \xi) \tag{3.2.28}$$

and the log likelihood function by:

$$\ell(\mu,\sigma,\xi|x_1,x_2,...,x_N) = \ln L(\mu,\sigma,\xi|x_1,x_2,...,x_N) = \sum_{i=1}^{N} \ln f(x_i;\mu,\sigma,\xi) \qquad (3.2.29)$$

For convenience sake, the log-likelihood function is denoted by $\ell(\mu,\sigma,\xi)$. The maximum likelihood estimators $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ are the parameter values that will maximise the log likelihood function. In the case of the Generalised extreme value distribution when $\xi \neq 0$,the log likelihood is given as:

$$\ell(\mu,\sigma,\xi) = \sum_{i=1}^{N}\left(-\ln(\sigma) - \left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}} - \left(\frac{1}{\xi}+1\right)\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]\right) \qquad (3.2.30)$$

with the restriction that $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ for $i = 1,2,3,...,N$. If we differentiate equation (3.2.30), above with respect to μ,σ, and ξ , this yields likelihood equations for the Generalised extreme value distribution as follows:

$$\frac{\partial\ell(\mu,\sigma,\xi)}{\partial\mu} = \sum_{i=1}^{N}\left(\frac{-1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-(\frac{1}{\xi}+1)} + \frac{1}{\sigma}\left(1+\xi\right)\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1}\right) = 0 \qquad (3.2.31)$$

$\frac{\partial\ell(\mu,\sigma,\xi)}{\partial\sigma} = \sum_{i=1}^{N}\left(\frac{-1}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-(\frac{1}{\xi}+1)} + \left(\frac{1+\xi}{\sigma}\right)\left(\frac{x_i-\mu}{\sigma}\right)\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1} - \frac{-1}{\sigma}\right) = 0$ $\frac{\partial\ell(\mu,\sigma,\xi)}{\partial\sigma} = \sum_{i=1}^{N}\left(-\frac{1}{\xi^2}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{\frac{-}{\xi}} \ln\left(1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right) + \frac{1}{\xi}\left(\frac{x_i-\mu}{\sigma}\right)\left[1+\right.\right.$
$\left.\left.\xi\left(\frac{x_i-\mu}{\sigma}\right]^{-(\frac{1}{\xi}+1)} + \frac{1}{\xi^2}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \left(\frac{1}{\xi}+1\right)\left(\frac{x_i-\mu}{\sigma}\right)\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1}\right) = 0$

However, as mentioned earlier on this chapter, when $\xi = 0$, this would be a Gumbel distribution family. The Gumbel log likelihood function is given by:

$$\ell(\mu,\sigma,0) = -N\ln(\sigma) - \sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{N}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right] \qquad (3.2.32)$$

and likelihood equations are:

$$\frac{\partial\ell(\mu,\sigma,0)}{\partial\mu} = \sum_{i=1}^{N}\frac{1}{\sigma}\left\{1-\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right]\right\} = 0 \qquad (3.2.33)$$

$$\frac{\partial \ell(\mu,\sigma,0)}{\partial \sigma} = \frac{1}{\sigma}\sum_{i=1}^{N}\left\{ -1 + \frac{x_i-\mu}{\sigma}\left(1-\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right]\right)\right\} = 0 \qquad (3.2.34)$$

However, these equations for both the Generalised extreme value distribution and Gumbel case have no analytical solutions. Therefore, inorder to find the maximum likelihood estimates $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ we resort to the use of numerical optimisation algorithms such as Quasi-Newton numerical maximisation.

Bayesian Estimation

A trivariate normal prior, which is a non informative prior is discussed in Coles and Powell (1996); Stephenson and Ribatet (2006) is used in this dissertation. Let the prior distribution be denoted by $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = \{\sigma, \mu, \gamma\}$. Let

$$\boldsymbol{\theta}' = \{\log\sigma, \mu, \gamma\}$$

then the prior distribution on $\boldsymbol{\theta}$ is defined as Stephenson and Ribatet (2006).

$$\pi(\boldsymbol{\theta}') \propto \frac{1}{\sigma}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}'-\boldsymbol{\nu}\right)^T \Sigma^{-1}\left(\boldsymbol{\theta}'-\boldsymbol{\nu}\right)\right\} \qquad (3.2.35)$$

where $\boldsymbol{\nu}$ is the mean vector and Σ is the covariance matrix. The likelihood function is

$$\pi(x|\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n}\frac{1}{\sigma}\left[1+\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}-1}\exp\left\{-\left[1+\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}}\right\} \qquad (3.2.36)$$

The joint posterior density is then given as:

$$\pi(\boldsymbol{\theta}|x) \;\propto\; \pi(\boldsymbol{\theta})\pi(x|\boldsymbol{\theta}) \qquad (3.2.37)$$

which reduces to

$$\pi(\boldsymbol{\theta}|x) \propto \frac{1}{\sigma^{n+1}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}'-\boldsymbol{\nu}\right)^T\Sigma^{-1}\left(\boldsymbol{\theta}'-\boldsymbol{\nu}\right)-\sum_{i=1}^{n}\left[1+\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}}\right\}$$

$$\times \prod_{i=1}^{n}\left[1+\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}-1} \tag{3.2.38}$$

3.2.5 Stationary and non Stationary models

The following models defined below will be considered in this study. Nonstationary models have characteristics which change randomly with a deterministic component over time. In this context of environmental issues, non stationarity is apparent because of seasonal effects most probably due to the change of climatic patterns in different months or in the form of trends, possibly due to long term climate changes.

$$X(t) \sim \text{GEVD}(\mu(t), \sigma(t), \xi(t)) \tag{3.2.39}$$

Using the notation $\text{GEV}(\mu, \sigma, \xi)$ to denote the GEV distribution with parameters μ, σ, ξ it follows that a suitable model for $X(t)$, the annual flood heights of Limpopo river at Beitbridge border post in year t as given in equation (3.2.39) above.
where $\mu(t) = \beta_0 + \beta_1(t)$ for parameters β_0 and β_1. In this way the variation through time in the observed process are modelled by a linear trend in the location parameter of the appropriate extreme value model which in this case is the GEV distribution.

$$M_0 : \mu(t) = \mu, \sigma(t) = \sigma, \xi(t) = \xi \tag{3.2.40}$$

M_0 is considered as a benchmark model of which all models listed there after are

dependent to M_0.

$$M_1 = \begin{cases} \mu_1(t) = \mu_{01} + \mu_{02}t \\ \sigma_1(t) = \exp(\sigma_{01} + \sigma_{02}t) \\ \xi(t) = \xi \end{cases}$$

Non stationarity can also be expressed in terms of other extreme value parameters for example $\sigma_1(t) = \exp(\sigma_{01} + \sigma_{02}t)$ where the exponential function is used to ensure that the positivity of σ is respected for all values of t. This then justifies the use of model M_1 in this research as given above.

$$M_2 = \begin{cases} \mu(t) = \mu_{03} + \mu_{04}(t) + \mu_{05}t^2 \\ \sigma_2(t) = \sigma_{03} + \sigma_{04}(t) + \sigma_{05}t^2 \\ \xi(t) = \xi \end{cases}$$

When there are more complex changes in the modelling of flood heights for Limpopo river at Beitbridge border post, we resort to the use of a quadratic model $\mu(t) = \mu_{03} + \mu_{04}(t) + \mu_{05}t^2$ as given in the model M_2 above.

$$M_3 = \begin{cases} \mu_3(t) = \mu_{06} + \mu_{07}(t) + \mu_{08}\text{SOI}(t) \\ \sigma_3(t) = \sigma_{06} + \sigma_{07}(t) + \sigma_{08}\text{SOI}(t) \\ \xi(t) = \xi \end{cases}$$

A different scenario may arise whereby the extremal behaviour of one series is related to that of another variable referred to as a covariate. A similar phenomenon is whereby the annual flood heights of Limpopo river at Beitbridge border post are modelled by $X(t) \sim \text{GEVD}(\mu(t), \sigma(t), \xi(t))$ whereby $\mu_3(t) = \mu_{06} + \mu_{07}(t) + \mu_{08}\text{SOI}(t)$ and $\text{SOI}(t)$ denotes the Southern Oscillation Index in year t. In this research, this is justified by the use of models M_3 above and M_4 below.

$$M_4 = \begin{cases} \mu_4(t) = \sigma_{09}(t) + \sigma_{10}\text{SOI}(t) \\ \sigma_4(t) = \sigma_{09} + \sigma_{10}\text{SOI}(t) \\ \xi(t) = \xi \end{cases}$$

3.2.6 Deviance Statistics

Generally, maximum likelihood estimation of nested models leads to a simple test procedure of one model against the other. With models $M_0 \subset M_1$, then we will define the deviance statistic as:

$$D = 2\{\ell_1(M_1) - \ell_0(M_0)\} \tag{3.2.41}$$

where $\ell_1(M_1)$ and $\ell_0(M_0)$ are maximised log likelihood under models M_1 and M_0 respectively. The asymptotic distribution of D is given by the χ^2_κ distribution with κ degrees of freedom, where κ is the difference in dimensionality of M_1 and M_0; thus, calculated values of D can be compared to critical values from χ^2_κ, where large values of D suggest that the model M_1 explains substantially more of the variation in the data than M_0.

3.2.7 Augmented Dickey-Fuller test

The Dickey-Fuller test, was used in this research to test whether flood heights for the Limpopo river at Beitbridge border post are stationary. The Dickey-Fuller test is estimated under the following three null hypothesis.

(i) Y_t is a random walk : $\Delta Y_t = \delta Y_{t-1} + u_t$.

(ii) Y_t is a random walk with drift:$\Delta Y_t = \beta_1 + \delta Y_{t-1} + u_t$.

(iii) Y_t is a random walk with drift around a stochastic trend: $\Delta Y_t = \beta_1 + \beta_2 t \delta Y_{t-1} + u_t$

where t is the time or trend variable. In each of the above cases, we have $H_0 : \delta = 0$, that is, there exists a unit root and the time series is non stationary. The alternative hypothesis states that: $H_1 : \delta < 0$ which implies the time series is stationary. If H_0

is rejected, then Y_t is a stationary time series with zero mean in the first case. In the second case, it is implied that Y_t is stationary with non zero mean and lastly, Y_t is stationary around a deterministic trend. The Augmented Dickey Fuller test is conducted by augmenting the equations above by adding the lagged values of the dependent variable ΔY_t. We are concerned in estimating the following regression equation.

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-i} + \epsilon_t \tag{3.2.42}$$

where ϵ_t is a pure white noise error term and where $\Delta Y_{t-1} = (Y_{t-1} - Y_{t-2}), Y_{t-2} = \Delta Y_{t-2} - Y_{t-3}$ and so forth. The number of lagged differences to be added is determined empirically where the main idea is to include enough terms so that the error term is serially uncorrelated. In the Augmented Dickey Fuller test, we test $\delta = 0$. The same critical values are used since the Augmented Dickey Fuller test follows the same asymptotic distribution as the Dickey Fuller statistic.

3.2.8 Model diagnostics

Before estimating return levels, we should check the goodness of fit of our model which allows for a linear trend in μ. The lack of homogeneity in the distributional assumptions for each observation, however, mean some modification of the standard procedures (for example probability plots and quantile plots) are required to do this model diagnostics. We define the residuals as the difference between the actual values and the fitted values of a model: $\hat{e} = y_i - \hat{y}_i$. Before we make conclusions about any model, we normally need to carry out an assessment to whether the model satisfies the model assumptions. Amongst the assumptions to be tested is the assumption of normality on residuals as well as the independence assumption. Moreso, we test whether the residuals have a constant variance or not.

3.2.9 Modelling the R largest order statistics

We normally resort to the use of r largest statistics if and only if we have limited data. It is well known that extremes are scarce by nature, therefore the issue has motivated the search for characterizations of extreme value behavior that would enable us to model the data other than the use of one method of block maxima. In solution to this, we then resort to the use of r largest order statistics within each block that is for small values of r.

Consider the vector of r largest order statistics in a block (say a year), rescaled in the same way as the maxima $\tilde{M}_n^{(1)}$ with $\mu = \tilde{M}_n^{(r)}$, also consider $X_1, X_2, ..., X_n$ a sequence of independent and identically distributed random variables. Define $M_n^{(k)} = k^{th}$ largest of $\{X_1, ..., X_n\}$. If there exists sequences of constants $\{a_n > 0\}$ and $\{b_n > 0\}$ such that: $P\{(m_n - b_n)/a_n \leq z\} \to G(z)$ as $n \to \infty$ for some non degenerate distribution from G, then, for fixed r, the limiting distribution as $n \to \infty$ of $\tilde{M}_n^{(r)} = \left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n}\right)$ falls within the family having joint probability density function

$$f\left(z^{(1)}, ..., z^{(r)}\right) = \exp\left\{ -\left[1 + \xi\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]^{\frac{-1}{\xi}} \right\} \times \prod_{k=1}^{r} \frac{1}{\sigma}\left[1 + \xi\left(\frac{z^{(k)} - \mu}{\sigma}\right)\right]^{\frac{-1}{\xi}-1} \tag{3.2.43}$$

where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty; z^{(r)} \leq z^{(r-1)} \leq ... \leq z^{(1)}$; and $z^{(k)} : 1 + \xi\left(\frac{z^{(k)} - \mu}{\sigma}\right) > 0$ for $k = 1, 2, ..., r$, (Coles, 2001).

3.3 Generalised Pareto Distribution

3.3.1 Introduction

In (EVT) there are times when we are not interested solely in modelling maxima or minima but with observations which are above a high threshold. Modelling flood heights resembles one example. Balkema and de Haan (1974) and Pickands (1975) show that the distribution function of the excesses above a high threshold converges to a Generalised Pareto Distribution (GPD) as the threshold tends to the right end point. The GPD is a peaks- over-threshold (POT) distribution whose distribution is as follows:

$$W_{\xi}(x) = 1 - \left(1 + \frac{\xi(x-\tau)}{\sigma}^{\frac{-1}{\xi}}\right) \text{if } \xi \neq 0 \tag{3.3.1}$$

3.3.2 Threshold Selection

Three plots are applicable in selecting the threshold level τ and these are quantile-quantile plot, threshold stability plot and mean excess plot.

Quantile- Quantile Plots

The distribution of normalised interexceedance times is a mixture distribution. If we estimate $\hat{\theta}$ by using the intervals estimator, we can check now the goodness of fit of this underlying model: $\hat{\theta}$ estimates the proportion of interexceedance times which correspond to inter cluster times and whose normalised values should be well approximated by the exponential distribution with mean $\frac{1}{\hat{\theta}}$. The remaining normalised interexceedance times are intra cluster times and should be small relative to the inter cluster times.

The parameter stability plots

We exploit another useful property of the extremal index estimator to derive a further tool for threshold selection. This property is that of threshold stability which is used in threshold selection for many other extreme value models. The threshold stability of the extremal index estimator refers to its invariance to change in threshold above a suitably high threshold. Simply put, once a threshold is high enough, raising the threshold further should not dramatically change the estimated value of θ. As the threshold increases, the sample size used for estimation will fall so sampling uncertainty will increase, but the estimated values of $\hat{\theta}$ should not alter above this anticipated sampling variation.

Extremal mixture models

Extremal mixture models are currently used to determine threshold in extreme value theory. Modellers use them to compare what they get from using such similar methods of determining threshold such as the graphical techniques like the mean excess plot and Pareto quantile plot. The extremal mixture models currently in use comprise of parametric bulk model and the non parametric model for the bulk component. Under the former category we have the Normal GPD mixture distribution, Normal GPD with single continuity constraint mixture distribution, Gamma GPD mixture distribution, Gamma GPD with single with continuity constraint mixture distribution, Weibull GPD mixture distribution, Lognormal GPD mixture distribution and under non parametric we do have Kernel GPD distribution, Kernel GPD with single continuity constraint distribution, two tailed Kernel GPD distribution and Boundary correction Kernel distribution.

Mean Excess plot

Beirlant *et al.* (2004) define the mean excess function with the finite expectation $E(X) < \infty$ as:

$e(t) = E\Big((X - t|X > t\Big)$

where for a random sample $x_1, x_2, ..., x_n$ the mean excess function is given by:

$$\hat{e}_n(t) = \frac{\sum_{i=1}^{n} x_i 1(t,\infty)(x_i)}{\sum_{i=1} n 1_{(t,\infty)}(x_i)} - 1 \tag{3.3.2}$$

In this case the indicator function $1_{(t,\infty)}$ is defined such that:

$$1_{(t,\infty)} = \begin{cases} 1 & \text{if } x_i > t \\ \\ 0 & \text{otherwise} \end{cases}$$

3.3.3 Declustering

In order to reduce the dependencies of time series data, we normally decluster the exceedances using the method of declustering which Ferro and Sergers (2003) has discussed. The intervals estimator method that was proposed by Ferro and Sergers (2003) is given as:

$$\eta_u = \frac{2\Big[\sum_{n-1}^{N-1}(T_i - 1)\Big]^2}{(N-1)\sum_{i=1}^{N-1}(T_i - 1)(T_i - 2)} \tag{3.3.3}$$

In this case, u is a sufficiently high threshold and T_i denote the exceedance times. According to Smith (1989) and Coles (2001), the extremal index measures the amount of clustering and $0 \leq \eta \leq 1$, where define $\frac{1}{\eta_u}$ as the limiting mean cluster size.

To check for goodness of fit of the Ferro and Sergers(2003) model for the estimated extremal index η_u, we use a Quantile- Quantile (QQ) plot of interexceedance times against standard exponential quantiles.

Since the mid 1990s, various methods for declustering a series of extremes, to extract a set of independent extremes, have been discussed in the literature. The most natural, commonly used method of declustering is that of runs declustering. This is how it works:

(i) Choose an auxiliary declustering parameter.(which we call κ)

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

(iii) Go through the entire series identifying clusters in this way.

(iv) The maximum observation from each cluster is then extracted and the GPD is then fitted to the cluster of peaks excesses.

Although the method is quite easy to implement, there are issues surrounding the choice of κ; if:

(i) κ is too small the cluster peaks will not be far enough apart to safely assume independence.

(ii) κ is too large, there will be too few cluster exceedances on which to form our inference.

The identification of clusters makes use of the characterisation of interexceedance times as either inter cluster or intra cluster. Under the Ferro and Segers model, the extremal index θ arises as the proportion of intereceedance times which are also inter cluster times.
This means that interexceedance times can be easily categorised in these two groups as a natural consequence of this model.

Assuming that we have observed N threshold exceedances, then the θN largest interexceedance times are assumed to be approximately independent inter cluster times, which separate the remaining exceedances times into intra cluster times.

Practically, the estimated value of θ is used to identify the critical cut off interexceedance time that distinguishes inter from intra cluster times.

Once clusters have been identified, the cluster characteristics can be examined, or the original threshold exceedances can be thinned to give approximately independent cluster maxima. Inference on these cluster maxima can then be carried out such inference is much more straight forward than the inference on the original dependent sequence, although arguably this approach is wasteful of information as it discards all but the largest observation in each cluster.

3.3.4 Parameter Estimation for the Generalised Pareto Distribution

After determining a threshold, the parameters of the GPD can be estimated by maximum likelihood method. Suppose that the values $x_1, ..., x_k$ are the k excesses of a threshold u. For $\xi \neq 0$ then the log likelihood is derived as:

$$\ell\left(\sigma, \xi\right) = -k \log\left(\sigma\right) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log\left(1 + \frac{\xi x_i}{\sigma}\right) \tag{3.3.4}$$

provided that $\left(1 + \sigma^{-1}\xi x_i\right) > 0$ for $i = 1, ..., k$, otherwise $\ell\left(\sigma, \xi\right) = -\infty$. In the case when we have $\xi = 0$ the log likelihood for the GPD is as follows:

$$\ell\left(\sigma\right) = -k \log(\sigma) - \sigma^{-1} \sum_{i=1}^{k} x_i \tag{3.3.5}$$

More so, analytical maximisation of the log likelihood is highly impossible, so we then

resort to numerical techniques are once again resorted to in this respect to get the parameter estimate.

However there are some irregularities in using the method of maximum likelihood to estimate parameters of extreme value distributions (EVT). Smith (1985) emphasize the regularity conditions pertaining to the use of maximum likelihood estimation method in estimating the shape parameter. The regularity conditions that Smith (1985) emphasised are as follows:

(i) When the shape parameter ξ is greater than -0.5, then the maximum likelihood estimators are regular, in the sense of having the usual asymptotic properties.

(ii) When $-1 < \xi < -0.5$, the maximum likelihood estimators are generally obtainable, but do not have the standard asymptotic properties.

(iii) When $\xi < -1$, maximum likelihood estimators are unlikely to be obtainable.

(iv) When $\xi < -0.5$, corresponds to distributions with very short bounded upper tail.

In our study, we will use both the maximum likelihood estimation and Bayesian methods to estimate parameters. According to Coles (2001) and Smith (1987), the problem of regularity assumptions in the limiting behavior of the maximum likelihood estimation method is encountered if the shape parameter, $\xi < -\frac{1}{2}$. Therefore, the Bayesian based estimation approach gains much preference over maximum likelihood estimation and the probability weighted methods due to its independence of any regularity conditions Beirlant *et al.* (2004).

It is of great importance to show the relationship between the Generalised extreme value distribution and the Generalised Pareto distribution.

The following derivation outlines how the two methods are related.

Let Y follows a distribution over a high threshold u and y is the excess $(x-u)$.

$$\begin{aligned} Pr\{Y > u+y \mid Y > u\} &= \frac{1-Pr(Y<y+u)}{1-Pr(Y<u)} \\ &\approx \frac{\left[1+\xi(u+y-\mu)/\sigma\right]^{\frac{-1}{\xi}}}{\left[1+\xi(u-\mu)/\sigma\right]^{\frac{-1}{\xi}}} \\ &= \left[1+\frac{\xi(u+y-\mu)/\sigma}{1+\xi[(u-\mu)/\sigma]^{\frac{-1}{\xi}}}\right] \\ &= \left[1+\frac{\xi y}{\hat{\sigma_u}}\right]^{\frac{-1}{\xi}} \end{aligned} \tag{3.3.6}$$

3.3.5 Model diagnostics

Several standard statistical model diagnostic plots are used for checking model fit and the suitability of the threshold choice. Such plots include probability plots, return level plots, extremal mixture models and empirical versus fitted comparison plots. These checks based on a chosen threshold, so they are post calculation diagnostic plots. They provide an alternative way of assessing the model performance.

3.4 Estimating uncertainty using bootstrapping

The uncertainty in the way we estimate θ and in the subsequent cluster identification is estimated using a bootstrap algorithm. This bootstrap algorithm maintains the within cluster dependence structure, and exploits the independence between clusters by resampling clusters and inter cluster times rather than individual observations from the original dependent sequence.

The resulting estimates of θ and any other parameters associated with cluster characteristics represent a bootstrap estimate of the sampling distribution of these estimators.

Chapter 4

Results and discussion

4.1 Introduction

EVT is defined as a statistical tool that is normally used to give a thorough description of very rare events. This is applicable in the field of finance, floods, insurance, Embrechts (1999),engineering, Castillo (2012) and risk management, McNeil (1999)only to mention but a few.

However, in this dissertation our sole objective is to estimate return levels of Limpopo river flow at Beitbridge border post. This chapter goes on to explore an application of generalised extreme value distribution in modelling the upper tail heights of Limpopo river at Beitbridge border post.

4.1.1 Flood height data

In this dissertation, the historical monthly and annual maximums data for Limpopo river from Beitbridge border post station was used. Both the monthly and yearly data from this station is for the period 1992 to 2014. The annual maximum data was

used to model the GEV model while the monthly was used for the modelling using POT. This is verified data by South African weather services that was obtained from its website http://www.dwa.gov.za.

4.1.2 Descriptive statistics

Table 4.1: Table showing descriptive statistics for flood height.

	N	Range	Min	Max	1stQu	Median	Mean	3rdQu	std dev	variance
Floodheight	23	6.00	0.77	6.78	1.48	1.9	2.4	2.8	1.56	2.44

Table (4.1) shows the descriptive statistics of annual maximum flood heights for Limpopo river at Beitbridge border post. The maximum flood height was 6.78 metres where as the minimum flood height was 0.77 metres. The variance was 2.44 metres. We got this from an uninterrupted series of twenty three observations from this site. The range was approximately 6 metres.

4.1.3 Time series plot, density plot, normal QQ plot and Box plot for Limpopo river at Beitbridge Border post

Figure (4.1) illustrates the time series plot,density plot, normal QQ plot and box and whisker plot for the flood height data for Limpopo river at Beitbridge border post.

4.2 Stationary GEV models

4.2.1 Fitting the model without trend

Maximization of the GEV log likelihood for this data of flood heights leads to the estimates and standard errors as given in the table below.The results in table (4.2)

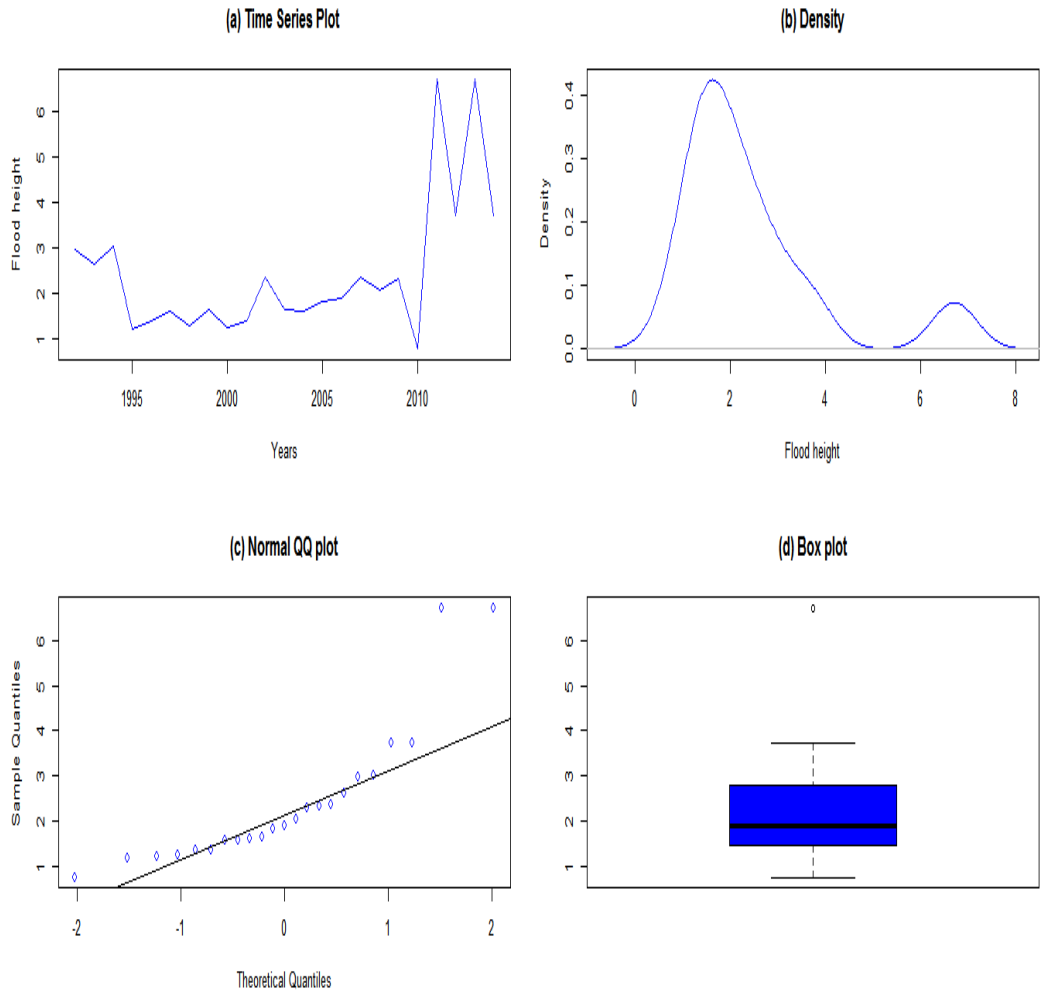

Figure 4.1: Top Panel (a)Time series plot (b) density plot Bottom Panel (c) Normal Quantile- Quantile plot (d) Box and Whisker plot.

indicate that the flood heights can be modelled using a Fréchet class of distributions since the shape parameter $\hat{\xi} > 0$. Combining the estimates and standard errors, the 95% confidence intervals for ξ, σ and μ are summarised in table (4.3) The 95%

Table 4.2: Table showing a summary of parameters and standard errors.

$\hat{\mu}(se)$	$\hat{\sigma}(se)$	$\hat{\xi}(se)$	ℓ
1.688 (0.176)	0.747 (0.149)	0.315(0.175)	33.8

Table 4.3: 95% confidence intervals for stationary GEVD model.

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
(1.343;2.031)	(0.4548;0.1.040)	(-0.0293;0.658)

confidence intervals for the shape parameter extends well above zero. This suggests that the flood heights for Limpopo river at Beitbridge border post are modelled by a Fréchet class of distributions. However, better accuracy of confidence interval bands are obtained by the use of profile likelihood of the shape parameter as shown below.

4.2.2 Profile Likelihood for the shape parameter

We can constantly change the range when plotting the profile likelihood until we see the confidence line. From the profile likelihood plot in figure (4.2) the confidence intervals for the shape parameter are approximately (-0.0293;0.658)

4.2.3 Diagnostic Plots for the GEV Stationary model

The various diagnostic plots in Figure (4.3) for assessing the accuracy of the GEV model fitted to the Limpopo river data at Beitbridge border post have been shown. Neither the probability plot nor the quantile- quantile plot give cause to doubt the validity of the model fitted: each set of plotted points is near linear. In the same manner, the return level curve asymptotes to a finite level. The estimate is close to

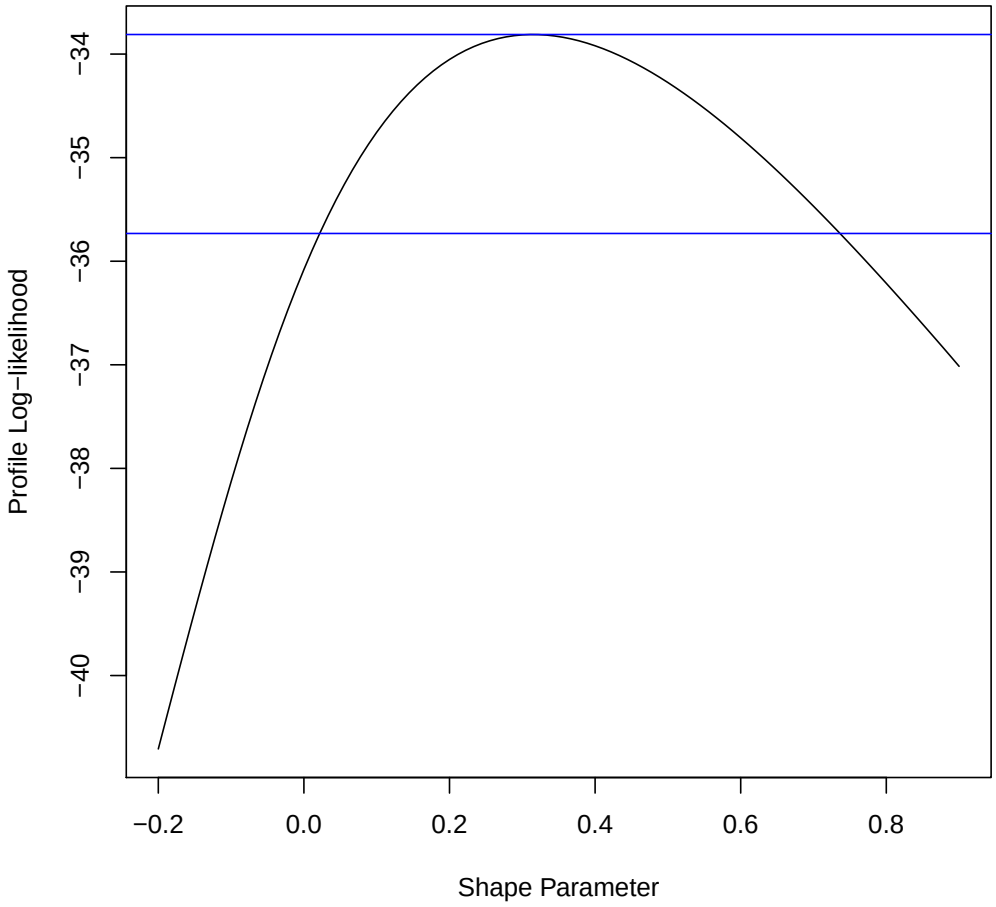

Figure 4.2: profile likelihood for the shape parameter.

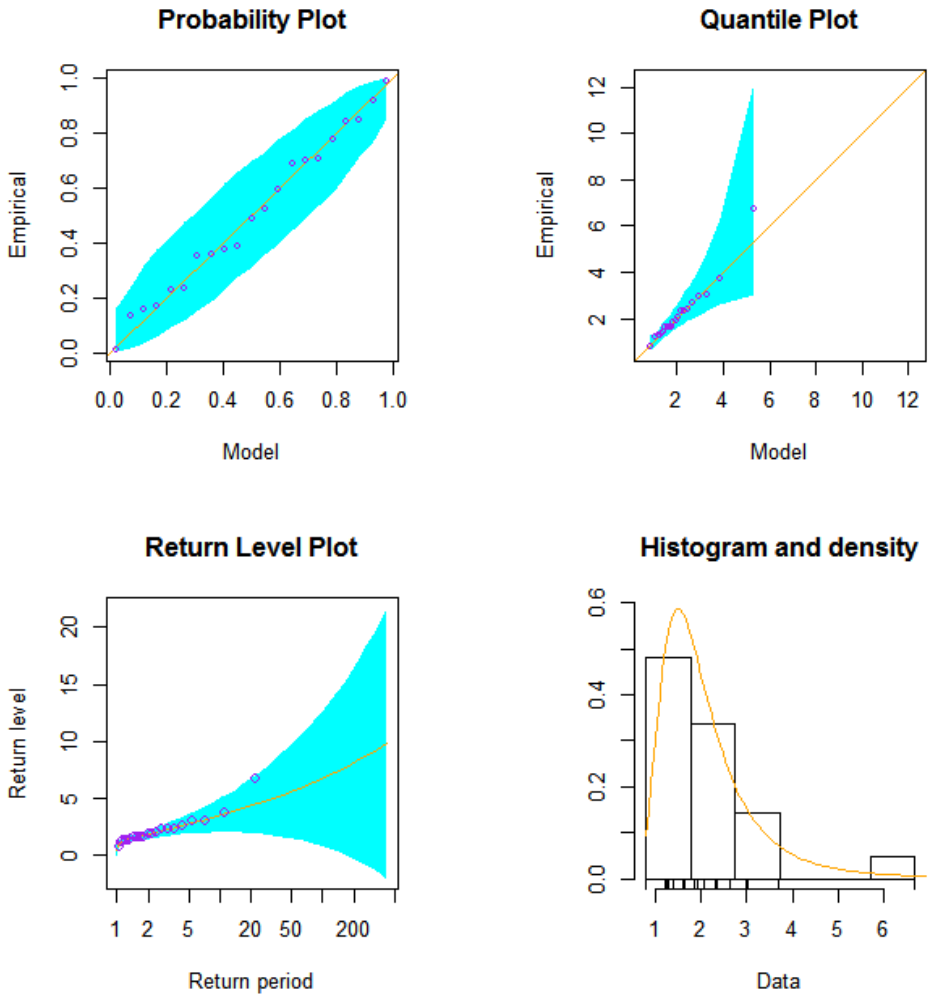

Figure 4.3: Diagnostic plots illustrating the fit of the maxima of flood height data for Limpopo river at Beitbridge border post to the GEVD, (a) P-P plot (top left panel), (b) Q-Q plot (top right panel), (c) Return level plot (Bottom left panel and (d) Bottom right panel).

zero and also the estimated curve is close to linear. More so, the curve provides a satisfactory representation of the empirical estimates, especially once we take into consideration sampling variability.

Finally, the density plot seems consistent with the histogram of the data. Consequently, all four plots support the GEV model.

4.2.4 Test for Stationarity using Augmented Dickey Fuller test

H_0: Flood heights data for Limpopo river at Beitbridge border post is stationary.

H_0: Flood heights data for Limpopo river at Beitbridge border post is not stationary.

Decision rule: We reject H_0 if $p < \alpha$

Since our $p = 0.3811 > \alpha = 0.05$ we fail to reject H_0 and concludethat the yearly flood heights data are stationary at 5% level of significance.

4.3 Non Stationary GEV models

The time series plot above for annual maximum flood heights seems for Limpopo river seems to exhibit some trends. There seems to be rather strong evidence for a positive trend over the years, and a substantial part of the variability in the data will probably be explained by a systematic variation in flood heights over years.

One way of investigating this trend is by allowing the generalised extreme Value Distribution location parameter μ to vary across time. A simpler linear trend in time seems plausible for the flood heights X, and this would lead us to the model:

$X_t = \text{GEV}(\mu(t), \sigma, \xi)$ where $\mu(t) = \beta_0 + \beta_1 t$ and t represents an indicator of year. In this way, variations over time are modelled as a linear trend in the location parameter of the GEV. In this case β_1 represents the slope in this case the maximum flood heights

of Limpopo river at Beitbridge border post. Therefore, the time homogeneous model is a special case of this time dependent model, with $\beta_1 = 0$; since this is nested within the model which allows for a time dependence, the deviance statistic can be used to formally compare these models.

After allowing for a linear trend in the data we get the estimates as summarised on the table below. This gives an estimated decrease in flood height of about -0.0170metres.

Table 4.4: Summary of estimates after allowing for a linear trend.

$\hat{\beta}_0(se)$	$\hat{\beta}_1(se)$	$\xi(se)$	$\sigma(se)$	ℓ
1.82(0.24)	-0.017(0.023)	0.70(0.17)	0.44(0.30)	33.63

This might be caused by effects of global warming . The confidence interval extracted after fitting this model is [-0.06226;0.02824] which includes a zero justifying the fact that the downward trend is insignificant. From the probability plot and quantile plot given in Figure (4.4), neither of the two cause a doubt that the model that allows for a linear trend in μ is adequate for our data points are all near linear and not deviating much from the reference line.

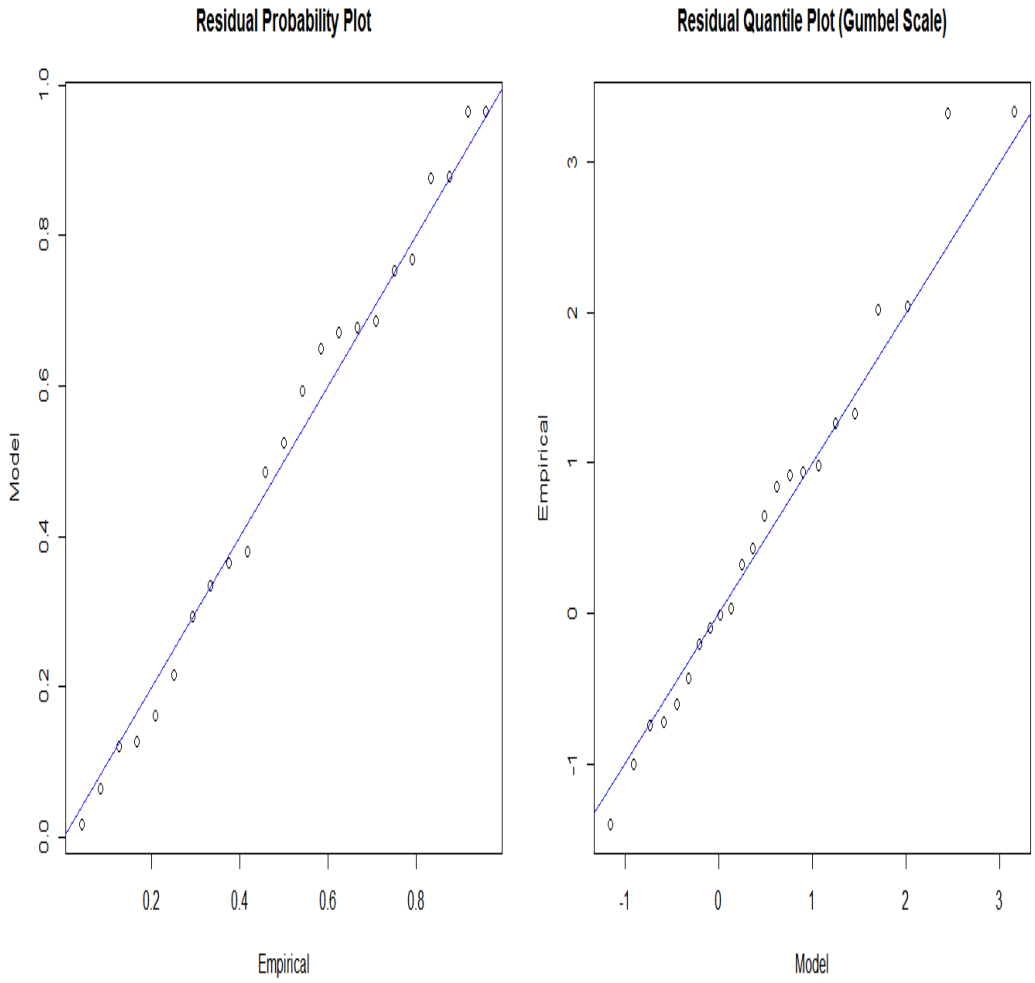

Figure 4.4: left panel: Residual probability plot Right Panel: Residual quantile plot for a GEV model allowing a linear trend.

4.3.1 Deviance statistics

We use the likelihood ratio test to check again for the significance of this decreasing trend. This would bring 2(33.81127-33.6301)=0.3618 which is small compared to χ_1^2= 3.841. Therefore allowing for a linear trend in time does not improve on our model which does not allow a linear trend therefore we reject the model $\mu(t) = \beta_0 + \beta_1(t)$ and conclude that there is no significance at all to allow for a linear trend.

4.3.2 Modelling flood height with inclusion of Southern Oscillation Index data

A different scenario which uses the same approach as that in the previous section is where the extremal behaviour of a series is related to another variable instead of time. For example studies have since revealed a link between annual maximum flood height and the mean value of the Southern Oscillation Index (SOI) which is an indicator of meteorological volatility due to the effects such as El Nino. In this section of analysis we try again to explore the results of flood heights with the inclusion of (SOI). Thus the following model for X_t, the annual flood heights for Limpopo river at Beitbridge border post in year t, might be suitable. The model we refer to here is of the form: $X_t \sim \text{GEV}(\mu(t), \sigma, \xi)$, whereby $\mu(t) = \beta_0 + \beta_1 \text{SOI}(t)$, where SOI(t) denotes the mean value of the (SOI) t years.

However, the time series plot reveals a possible trend in flood levels of Limpopo river at Beitbridge border post through time, which suggests that $\mu(t) = \beta_0 + \beta_1(t)$, where $t = 1, 2, ...,$ However, we are allowed to combine $\mu(t) = \beta_0 + \beta_1 \text{SOI}(t)$ and $\mu(t) = \beta_0 + \beta_1(t)$ in order to allow for dependence on time and (SOI) by letting: $\mu(t) = \beta_0 + \beta_1 \text{SOI}(t) + \beta_2(t)$. This section of analysis tries to check whether or not any of the equations mentioned above give a significant improvement over the stationary model. By incorporating (SOI)and time one variable at a time yields the

following results:

Comparing to the stationary model,we have D=226.51-26.21=0.6060

0.606 is less than 3.84 at $\chi_1^2(0.05)$, which suggests that there is no significance improvement over the stationary model.

The parameters obtained after inclusion of (SOI) are summarised as below: This

Table 4.5: Summary of estimates with inclusion of SOI.

$\hat{\beta}_0(se)$	$\hat{\beta}_1(se)$	$\xi(se)$	$\sigma(se)$	ℓ
1.45(0.27)	-0.017(0.02)	0.65(0.12)	0.18(0.16)	26.2

implies that when we add SOI data we get $\mu(t) = 1.45 - 0.017(t)$, $\hat{\sigma} = 0.65$ and $\hat{\xi} = 0.178$ and the log likelihood is 26.21.

4.3.3 Scatter plots

Here we are not looking for linear regression type relationship but whether the scatter (scale) or tail behaviour (shape) of flood height variable depend on the candidate explanatory variable. The Oscillation index variable looks like a possible candidate since the largest values of (SOI) are much more scattered for small values of flood heights than for large. This is depicted by the scatter plots as shown by Figure (4.5) below.

4.3.4 Correlation coefficient values for different months

Table 4.6: The correlation coefficients r.

Months	r
November	0.012
December	0.096
January	0.299
February	0.33
March	0.278
October	-0.123

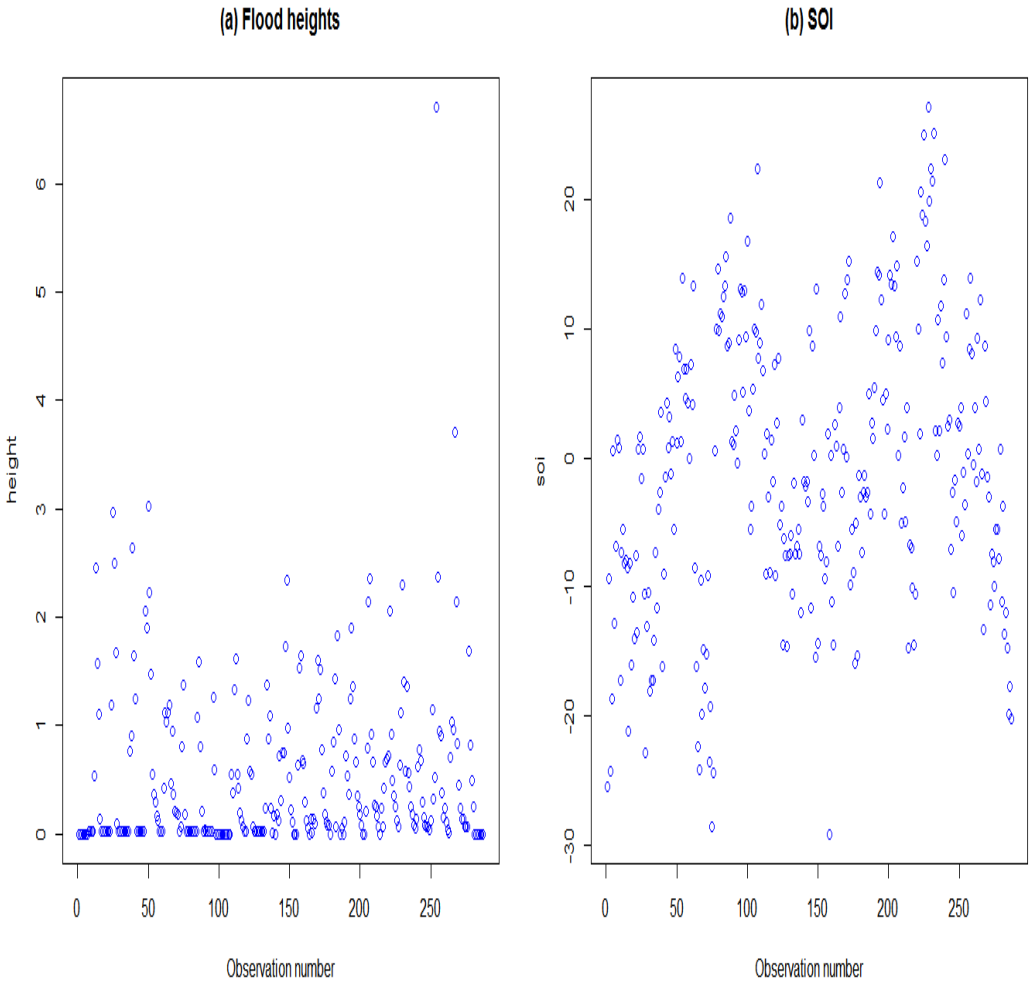

Figure 4.5: Top panel from left (a) Scatter plot showing flood heights (b) Scatter plot for Southern Oscillation index data.

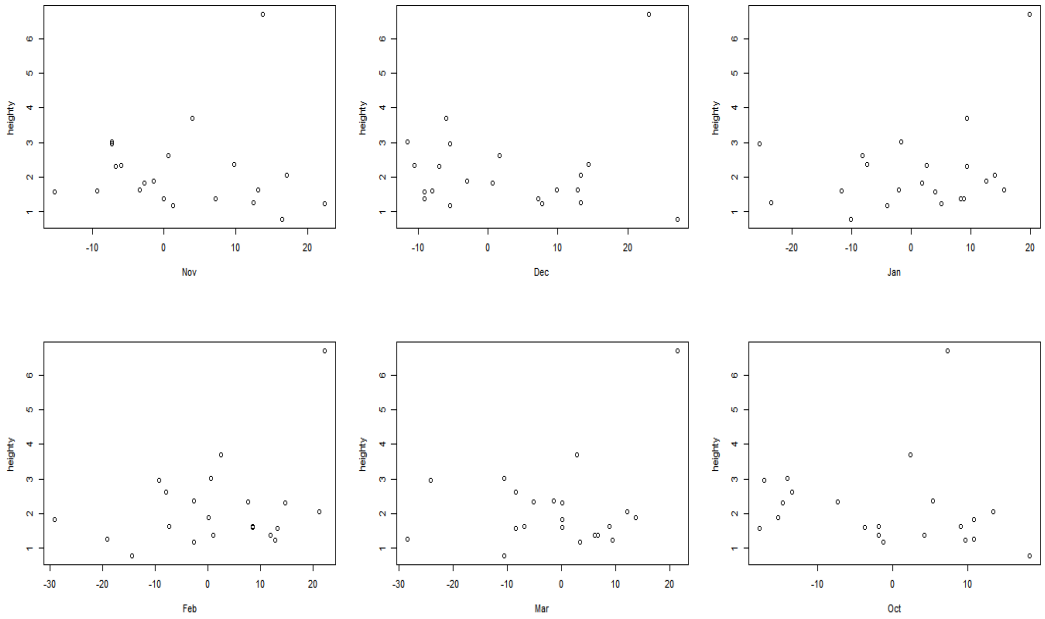

Figure 4.6: correlations plots (a) Correlation plots showing the relationship between flood heights and SOI for the months October-April.

From the plots in Figure (4.6) it is evident that February was a raining month because the flood height is positively correlated as portrayed by the correlation plots as well as by the value of r in Table (4.6).

4.3.5 Modelling the r largest order statistics

The table in Figure (4.7) illustrates the maximum likelihood estimates, standard errors and the 95% confidence intervals. Using the r largest order statistics method, the analysis was done for $r = 1$ to $r = 10$. Maximum likelihood estimates of the three GEV parameters and the associated standard errors (SEs) were calculated as shown in Figure (4.7). The standard errors associated with the shape parameters ξ for various values of r are shown in Figure (4.7). The variability decreases with an increase in r up to 2, but there is no appreciable change in standard error for r between 2 and 10. Moreso, a decreasing trend in standard errors is seen with the

Table 4.7: Maximum log likelihoods parameter estimates, confidance intervals and standard errors in parentheses of r largest order statistics model fitted to the Limpopo river data at Beitbridge border with different values of r.

r	ℓ	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	95%Confidence Interval for ξ
1	32.09	1.49(0.16)	0.72(0.12)	0.14(0.14)	(-0.13;0.13)
2	41.63	1.68(0.16)	0.85(0.09)	-0.01(0.07)	(-0.16;0.13)
3	39.32	1.65(0.15)	0.83(0.09)	0.01(0.07)	(-0.14;0.14)
4	21.98	1.68(0.14)	0.8(0.08)	0.04(0.07)	(-0.13;0.15)
5	7.37	1.63(0.14)	0.81(0.09)	0.04(0.08)	(-0.11;0.20)
6	-22.94	1.58(0.14)	0.82(0.11)	0.12(0.10)	(-0.08;0.32)
7	-62.16	1.54(0.15)	0.91(0.17)	0.28(0.14)	(-0.55;0.55)
8	-112.80	2.16(0.64)	2.75(1.41)	1.22(0.34)	(0.56;1.89)
9	-191.25	4.09(-)	8.03(-)	1.96(-)	(1.96;1.96)
10	-219.89	1.57(0.20)	1.18(0.24)	0.55(0.11)	(0.36;0.76)

modest increase in r, which levels off for $r > 2$. Therefore an optimum choice of r is expected to be between close to 2 and 3. Therefore we have chosen a fixed value of r to be 2 based on the standard errors and Figure (4.7) and also the plots in Figure (4.9)in our subsequent analysis.

4.3.6 Diagnostic plots for r=2 rlargest order statistics

As seen in Figure (4.8) all the plots lend support to the goodness of fit of the data for Limpopo river at Beitbridge border post. Neither of the probability nor the quantile-quantile plot can cause some doubts about the fitted model.

Each of the points is near linear in the above plots. The return level curve is close to linear too. Finally, the corresponding density estimate seems consistent with the histogram of the data. Consequently, all four plots do lend support to the fitted model using the r largest order statistics.

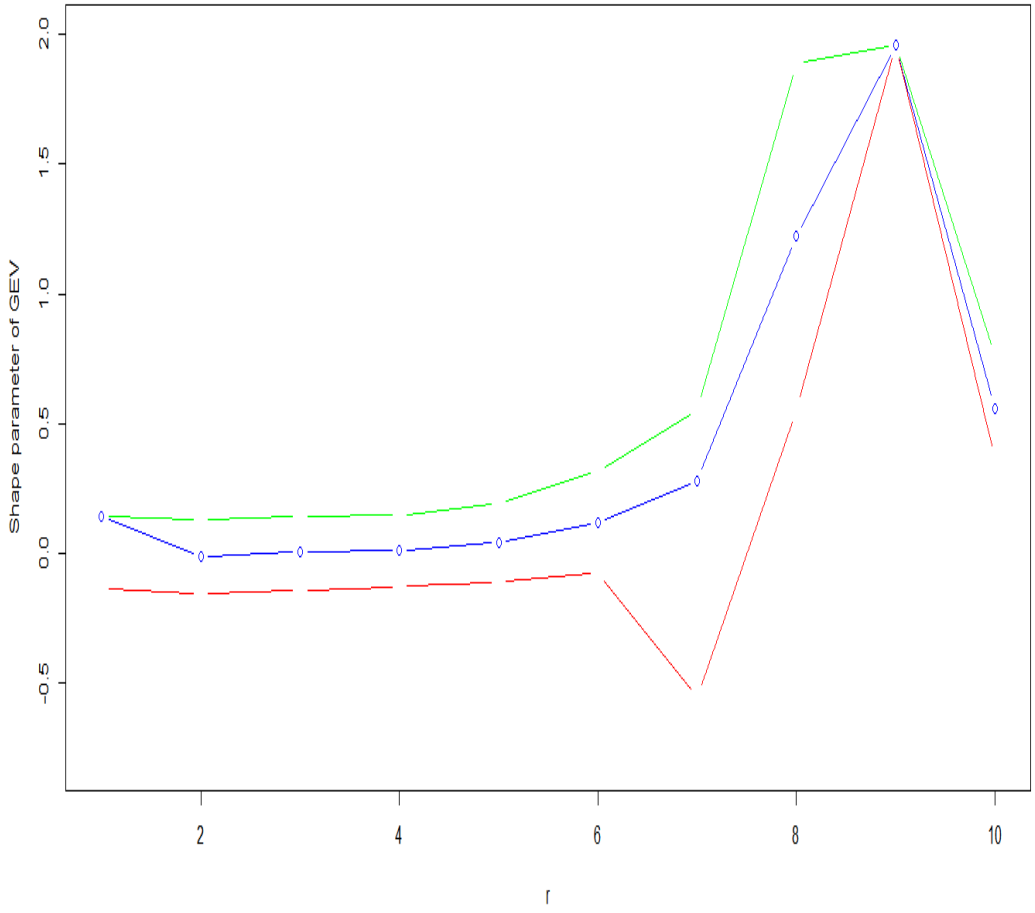

Figure 4.7: The GEV shape parameter with 95% confidence interval.

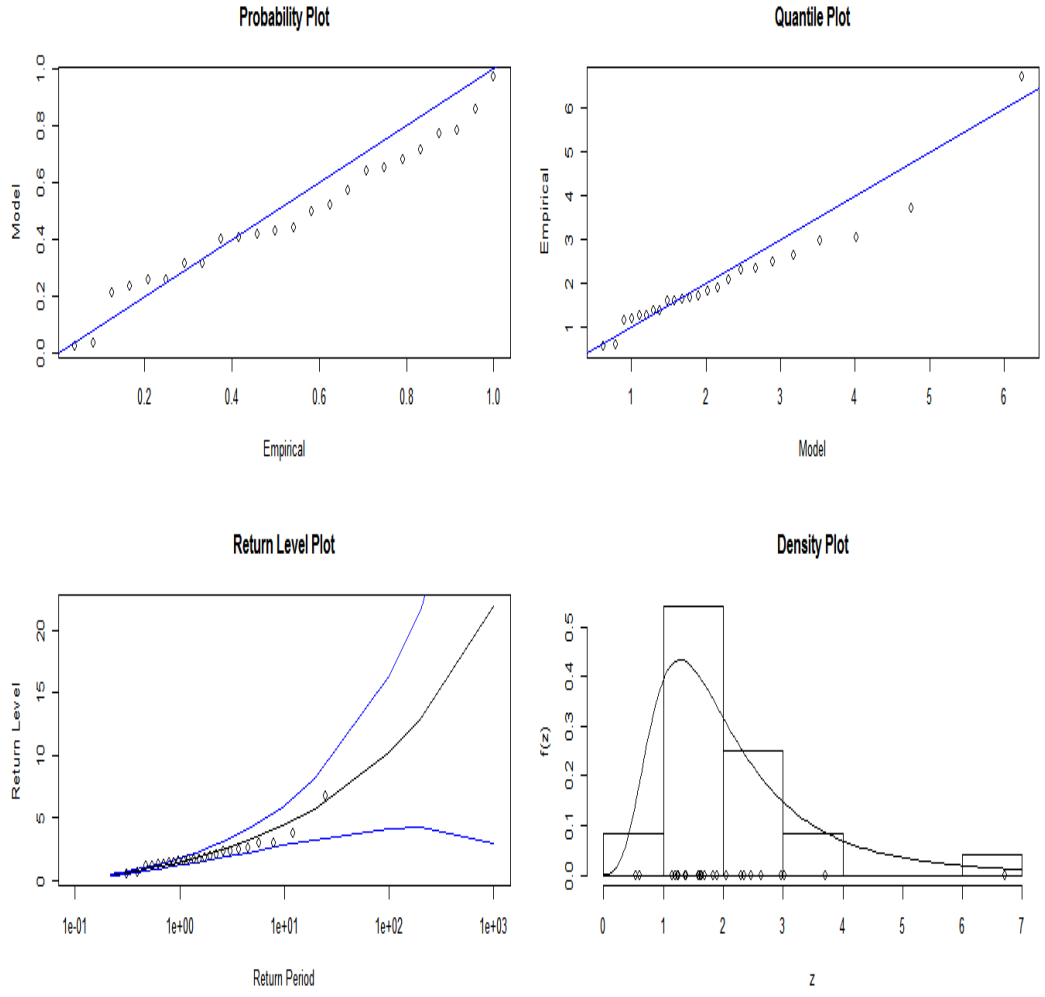

Figure 4.8: Diagnostic plots illustrating the fit of the data (Annual flood heights of Limpopo river at Beitbridge border post) to the GEVD for r largest order statistics model with $k = 2$, (a) P-P plot for $k = 2$ (Top left panel), (b) Q-Q plot for $k = 2$ (top right panel), (c) Return level plot for $k = 2$(Bottom left panel and (d)Density plot for $k = 2$ (bottom right panel).

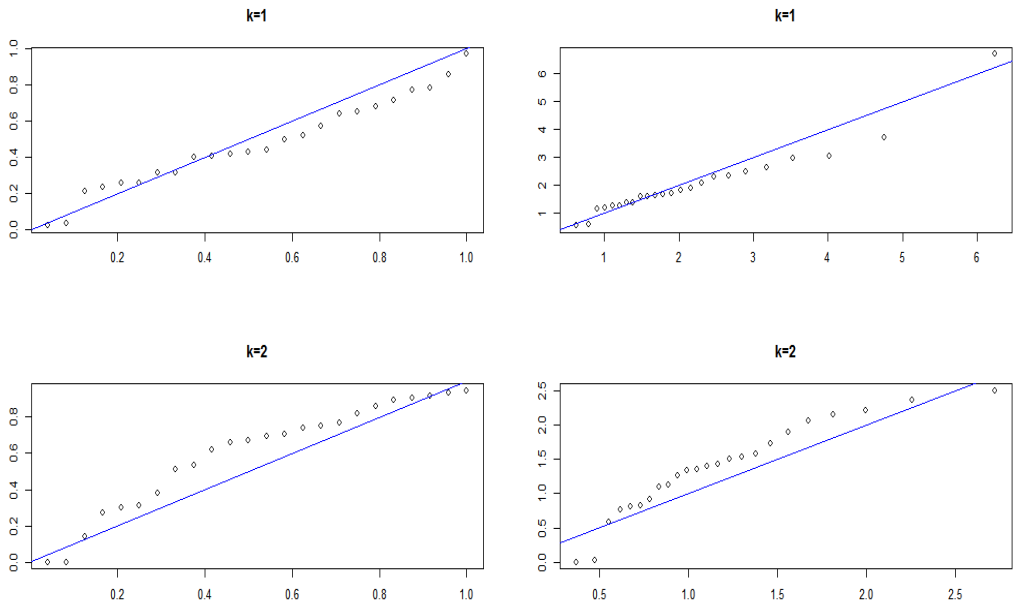

Figure 4.9: Diagnostic plots illustrating the fit of the data (Limpopo flood height data(at Beitbridge border post) to the GEVD for r largest order statistics model with $k = 1$ and $k = 2$)(a) P-P plot for $k = 1$(top left panel), (b) Q-Q plot for $k = 1$ (top right panel), (c) P-P plot for $k = 2$ (bottom left panel) and (d) Q-Q plot for $k = 2$ (Bottom right panel

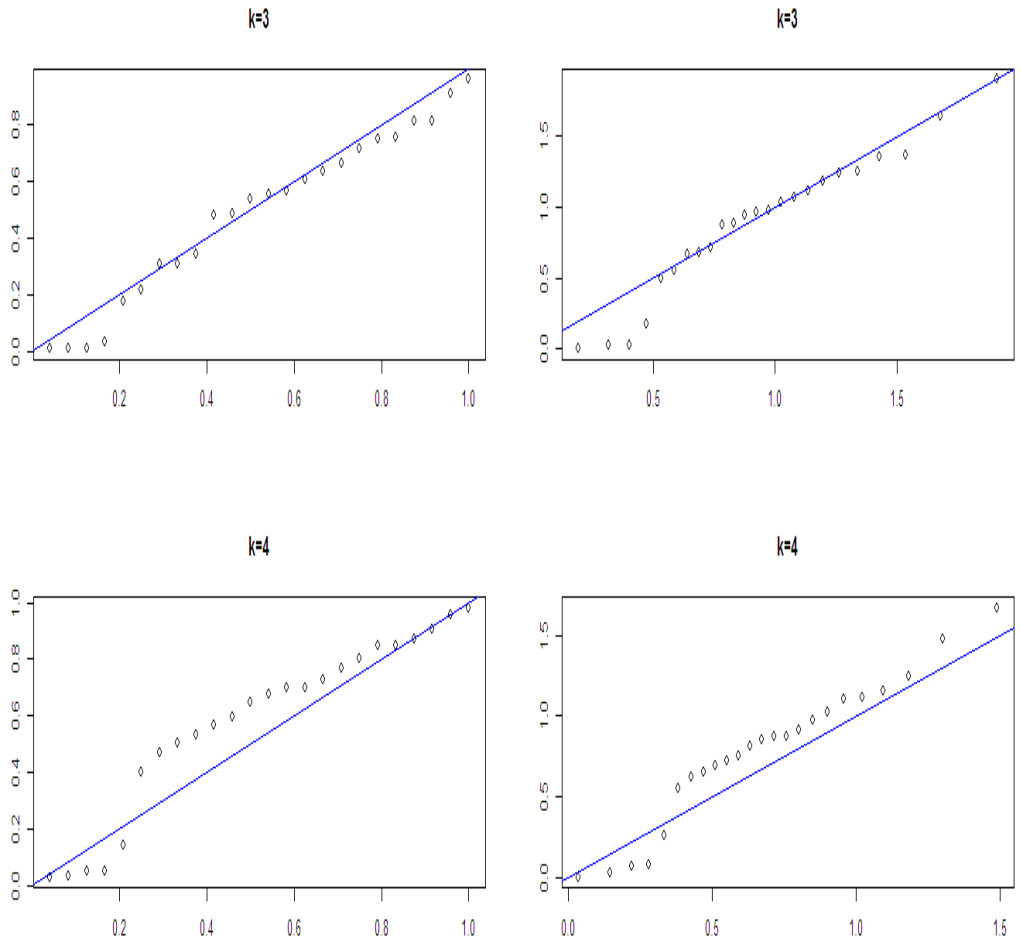

Figure 4.10: Diagnostic plots illustrating the fit of the data(Limpopo flood height data(at Beitbridge border post)to the GEVD for r largest order statistics model with $k = 3$ and $k = 4$)(a)P-P plot for $k = 3$(top left panel),(b)Q-Q plot for $k = 3$(top right panel),(c)P-P plot for $k = 4$(bottom left panel) and (d)Q-Q plot for $k = 4$(Bottom right panel.

4.3.7 Probability plots and quantile plots for r largest order statistics

From Figure (4.9) that show $r = 1$ and $r = 2$, neither of the probability plots nor the quantile- quantile plots creates doubt that the model fits within the range for all points are nearer linear. In other words points are not deviating so much from the reference line.

4.4 Generalised Pareto distribution

4.4.1 Introduction

In extreme value theory, sometimes we are not just concerned about modelling the maxima or minima but sometimes with observations that are above a high threshold. Flood heights of Limpopo river at Beitbridge border post is one such example. Balkema and de Haan (1974) and Pickands (1975) show that the distribution function of the excesses above a certain high threshold converges to a (GPD) as the threshold tend to the right end point

4.4.2 Declustering

The flood height variable for Limpopo river at Beitbridge border post is declustered whilst we retain the covariate structure of our data frame which holds the data. Firstly we select the threshold.

After having declustered the data, it suggests that we can not then use the threshold of below 1.33. The Figure (4.12) below illustrates the way we had chosen the threshold by taking into consideration the extremal index of flood heights for our data.

The extremal index of around 1.33 metres is chosen thereof with average clusters

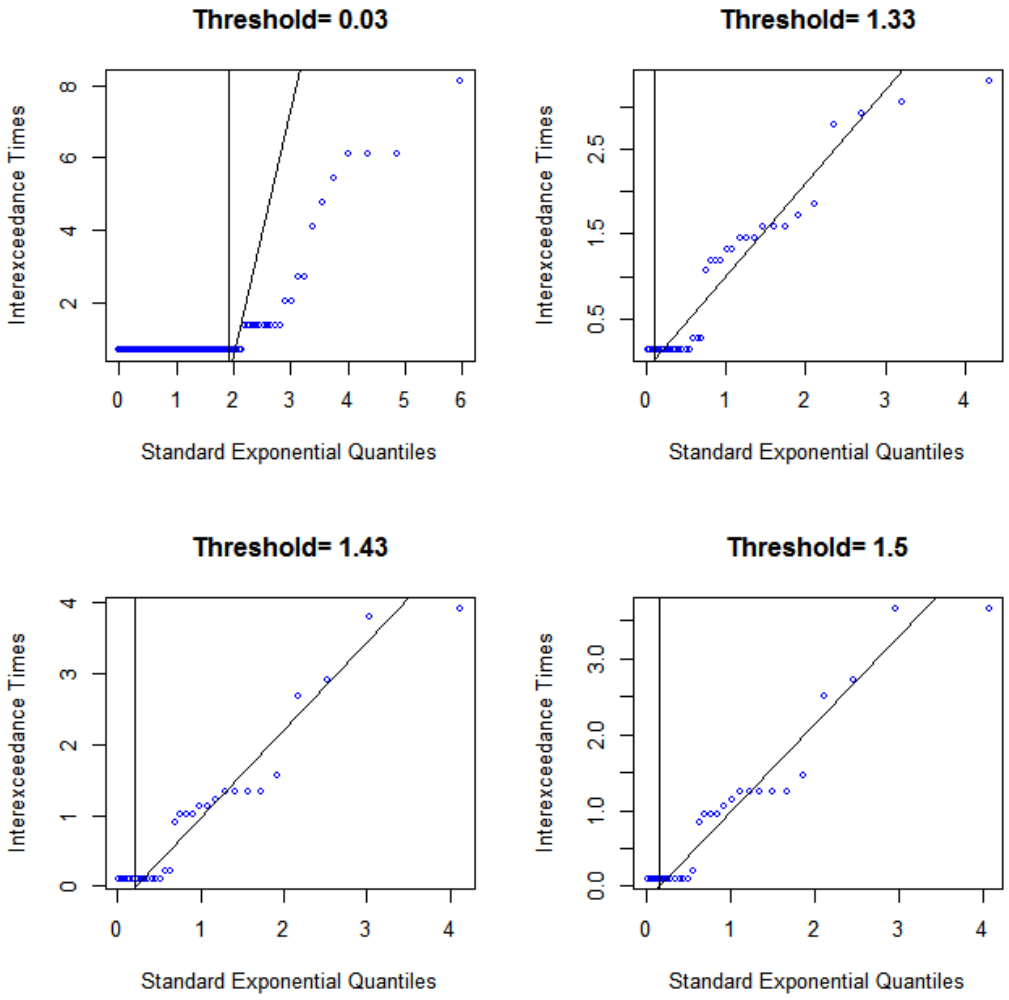

Figure 4.11: Quantile-Quantile plots of normalised exceedance times against standard exponential quantiles (vertical line shows the (1-$\hat{\theta}$))quantile slope; sloping line has gradient $\frac{1}{\hat{\theta}}$. The data is above 1.43 metres simulated from an autoregressive process with extremal index $\theta = 1.25$ and the corresponding $\frac{1}{\hat{\theta}}$ =0.8metres

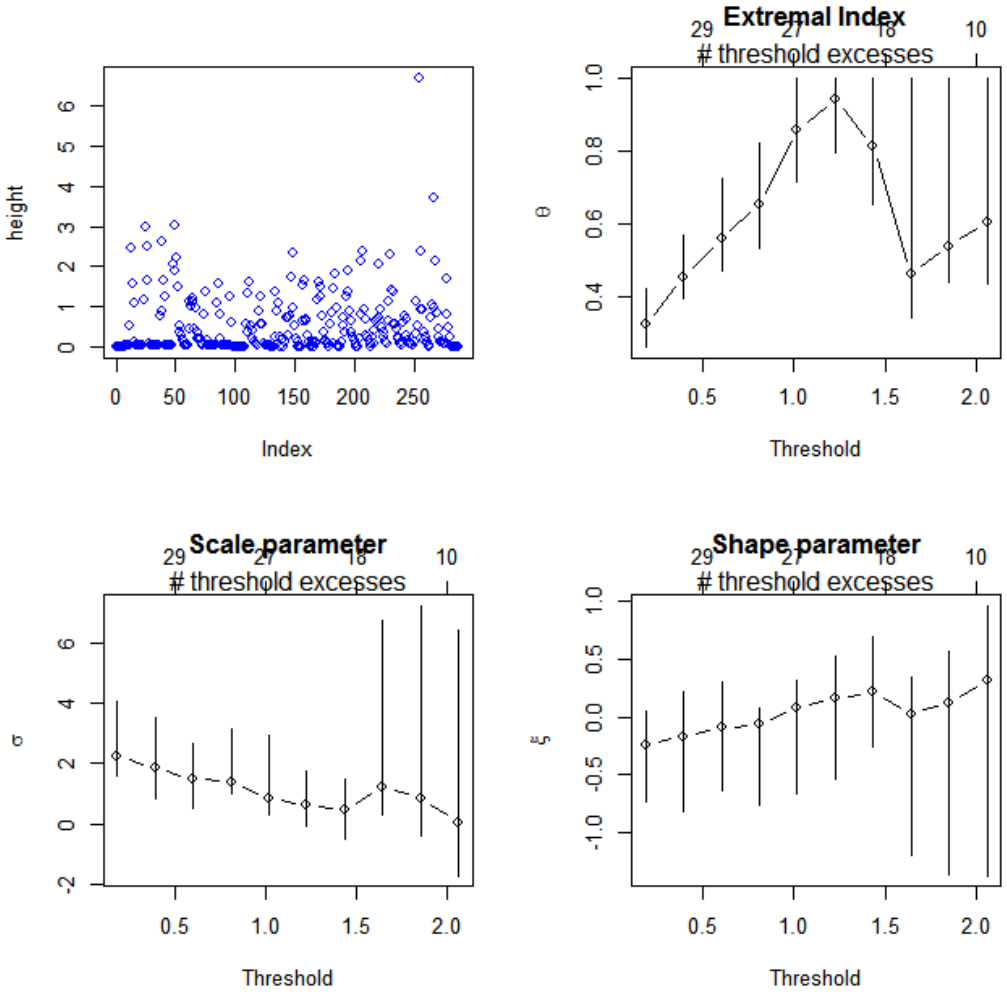

Figure 4.12: (a)Top four panels: The positive residuals (excesses)are on the top left panel. For the top right panel we have the threshold stability plots.

occurring with an average size of 1. We can therefore further on by declustering using this estimate of extremal index of 1.33 metres and a threshold of approximately 1.33metres. Therefore of the 287 threshold exceedances for our flood height data we have 23 clusters as illustrated in Figure (4.13) below.

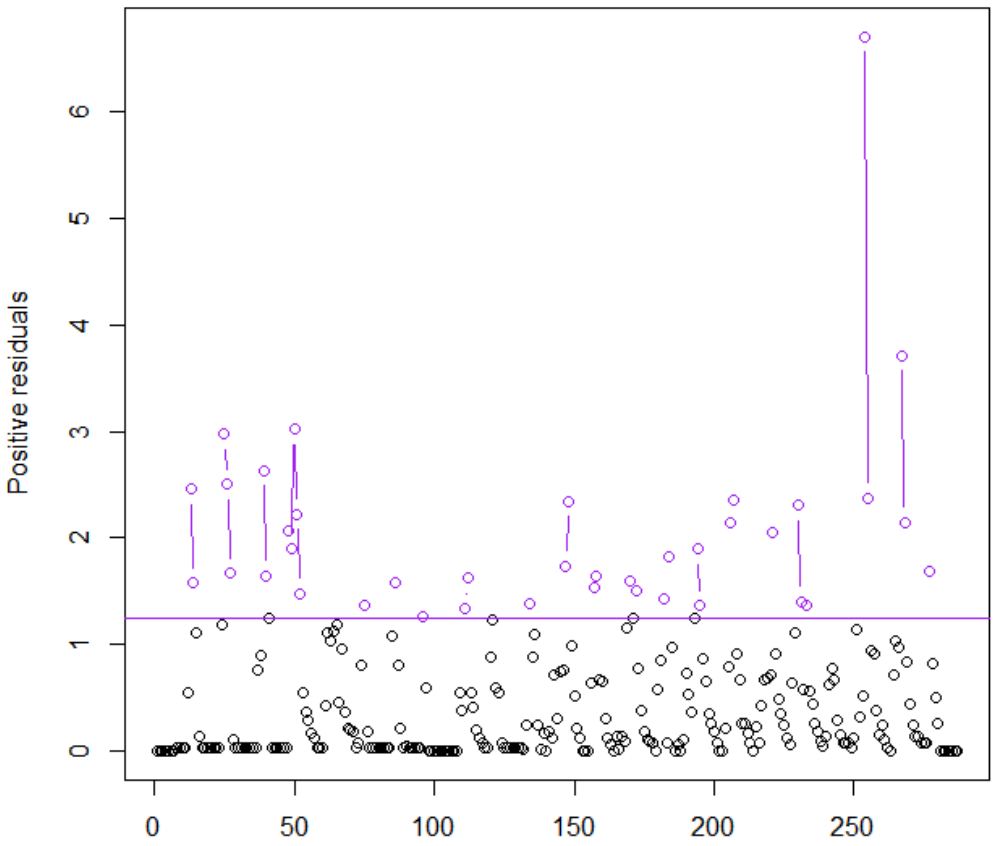

Figure 4.13: GPD fitted to cluster maxima (excesses)for the flood height data for Limpopo river at Beitbridge border post.

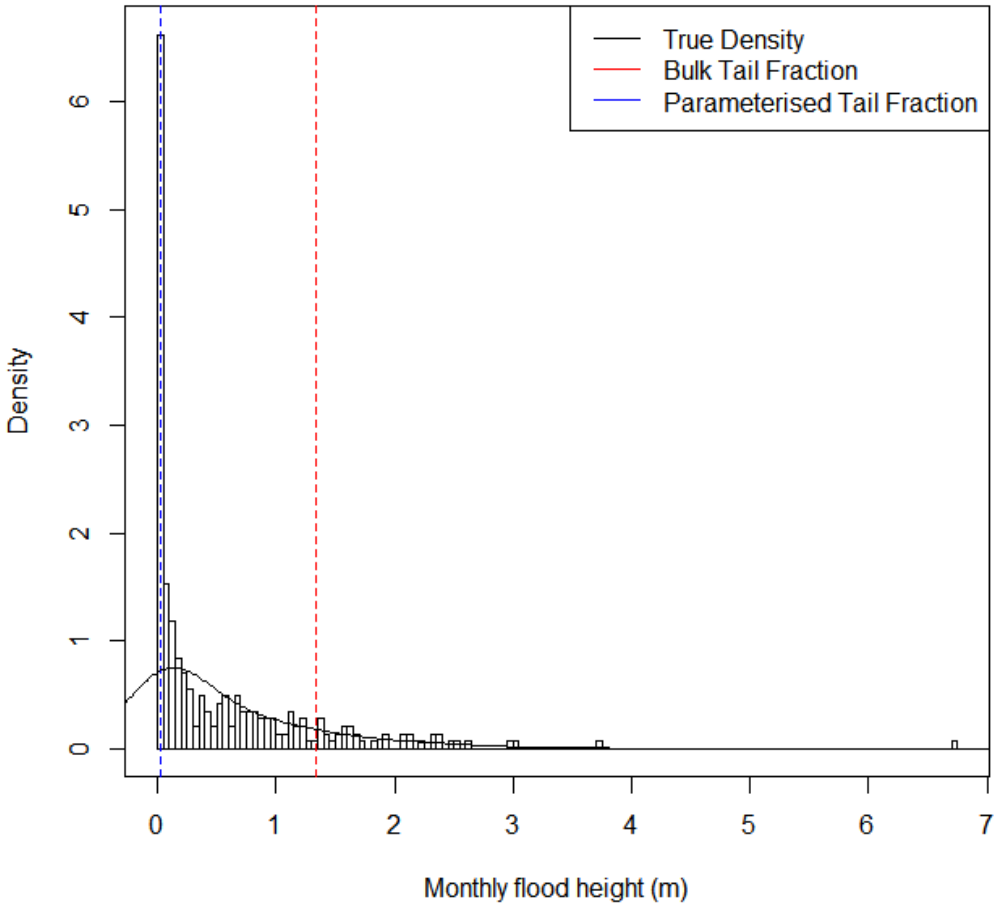

Figure 4.14: The threshold estimation using a parametric mixture model with a truncated Weibull distribution fitted to the bulk and a GPD to the upper tail. The tail fraction that is based on the bulk model is shown in red whereas the parameterised tail fraction is indicated in blue. The corresponding thresholds are respectively indicated by vertical dashed lines.

4.4.3 Use of extremal mixture model to select threshold

From the plot in Figure (4.14) we determine the two thresholds by bulk tail fraction and by parameterised tail fraction. It is evident from this plot that the threshold is 1.33 metres by the bulk tail fraction method and 0.025metres by parameterised tail fraction. Therefore we would then take our threshold to be 1.33metres because the mean excess plot gave the same threshold as this one. The parameters of scale and shape of the bulk tail fraction are 0.607 and 0.177 respectively where as the scale and shape parameters by the use of parameterised tail fraction are 0.64 and 0.13 respectively.

4.4.4 Examining the fit of threshold over a range of threshold

We can be reassured by these plots below that a threshold of 1.33 metres cause no doubt in terms of fit of the model to the flood heights data of Limpopo river at Beitbridge border post. For this choice of threshold, the estimate of the extremal index is about 1.25 metres and the corresponding $\frac{1}{E(I)} = 0.8$. The mean residual life plot is one method used to determine the threshold level. The interpretation of the mean residual life plot is not always simple practically. From the mean residual life plot in figure (4.15) it is better to conclude that there is some evidence for linearity above $\hat{\tau} = 1.33$ and therefore we work initially with threshold set at $\hat{\tau} = 1.33$.

We can now go on to estimate the parameters of the Generalised Pareto Distribution used to describe the conditional distribution of a cluster maximum given that it exceeds the threshold used for declustering.

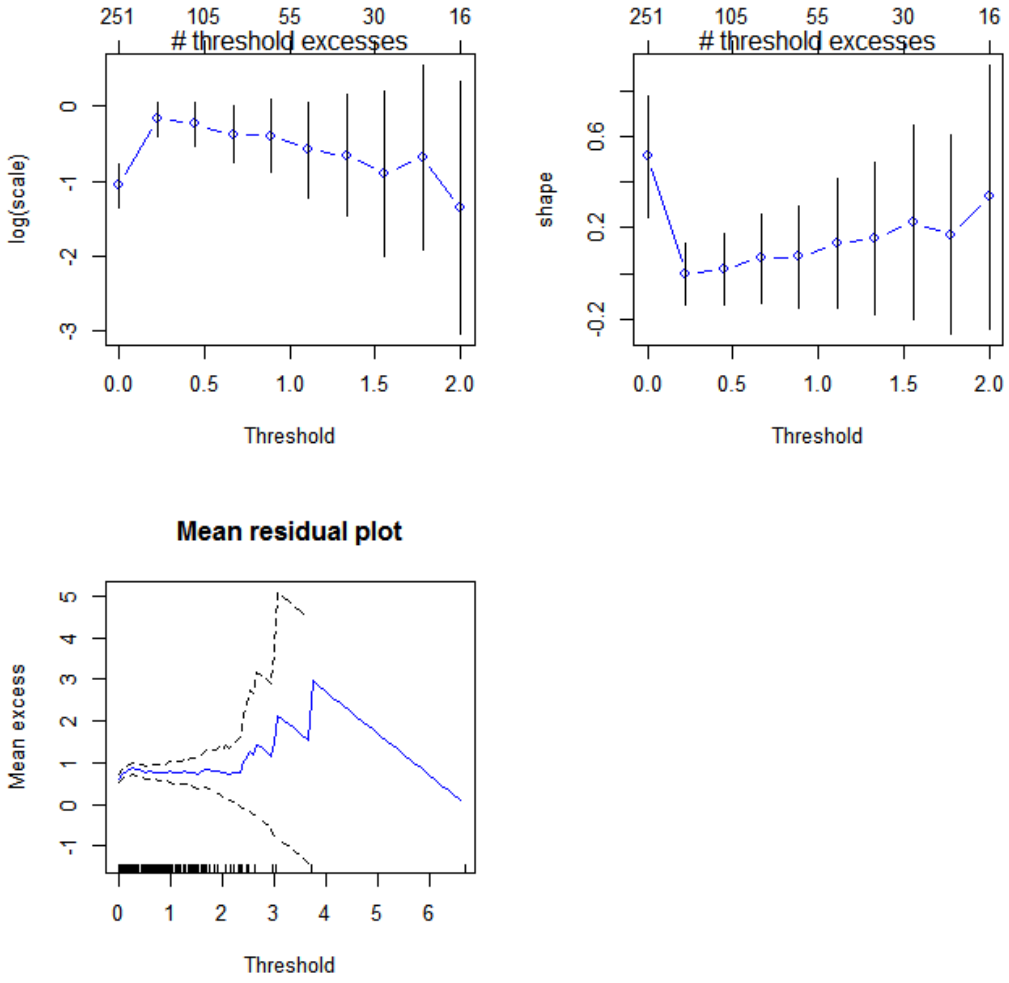

Figure 4.15: Threshold diagnostic plots ,(a)Scale parameter threshold stability plot (top left panel),(b) Shape parameter threshold stability plot(top right panel) and (c) Mean residual life plot (bottom panel).For the flood height data for Limpopo river at Beitbridge border post.)

Table 4.8: Estimates and standard errors for GPD fitting.

$\hat{\sigma}(se)$	$\hat{\xi}(se)$
0.719(0.21)	0.097(0.145)

4.4.5 Estimates and standard errors for GPD fitting

4.4.6 Model diagnostics of Generalised Pareto Distribution

These plots give minor cause for concern: Just fewer points at the top of the quantile plot lie slightly outside the tolerance intervals for this plot. This slight deviation from the model fit is not improved particularly by increasing the threshold.

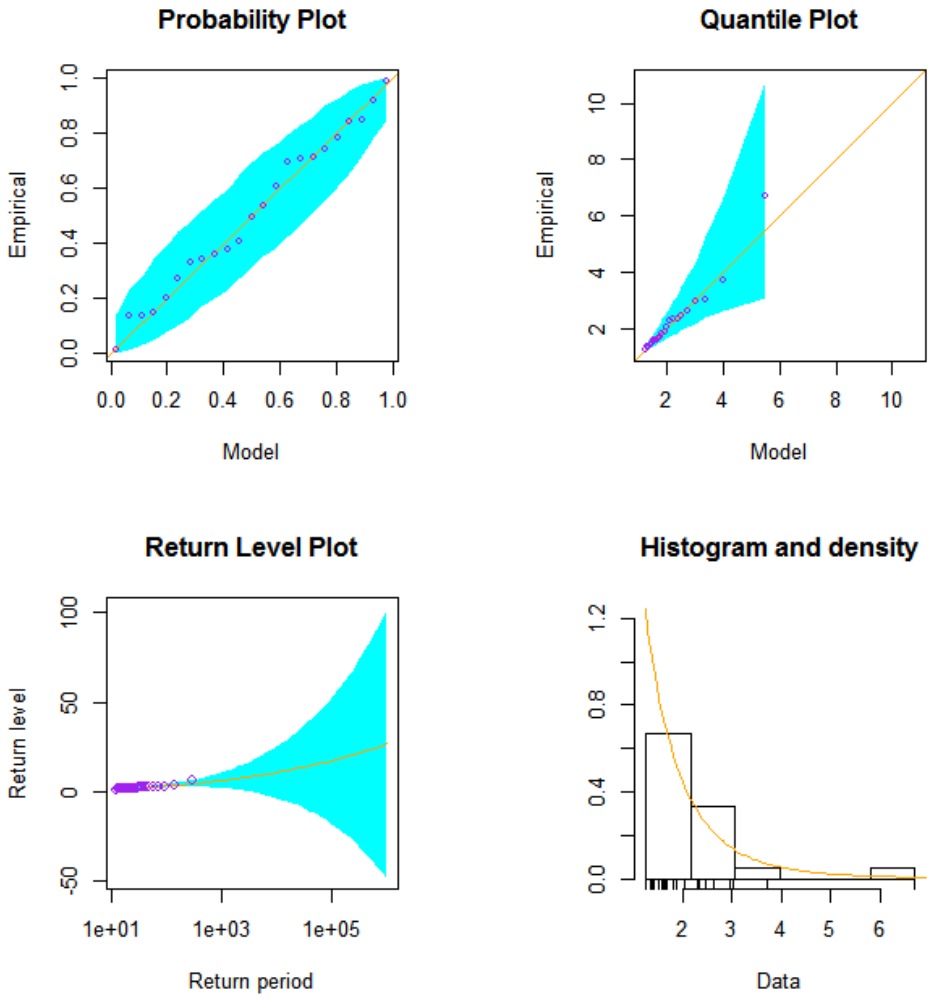

Figure 4.16: Diagnostic plots illustrating the fit of the cluster maxima of flood height data for Limpopo river at Beitbridge border post to the GPD, (a) P-P plot (top left panel),(b) Q-Q plot (top right panel),(c) Return level plot(Bottom left panel and (d) Bottom right panel) .

4.5 Bayesian Inference of the annual maxima of flood heights

4.5.1 Introduction

In this section we give an analysis of the annual flood heights for the rainfall data using the Markov Chain Monte Carlo method. The GEV scale parameter σ was reparameterised as $\log(\sigma)$ in order to retain the positivity of this scale parameter. We used different starting points in order to check whether the chains had converged to the correct place and all the chains are converged well.

4.5.2 Parameter estimates

Because of the uncertainties that surround the flood heights for Limpopo river at Beitbridge Border Post, the increase in flood heights in due course might be between 1.0metres and 2.5metres. In other words, this is very highly unlikely that the increase will be less than 1.0 metres for the flood height and on the same note that the increase might be above 2.5metres for the flood height in due course. But basing on the flood height of 2013, when the flood height was 6.707, it seems that the frequentist properties of the Bayesian inferences of the model based on the prior are not adequate.

Table 4.9: Bayesian estimates for the GEVD.

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1.61	-0.44	0.22

4.5.3 Bayesian Analysis

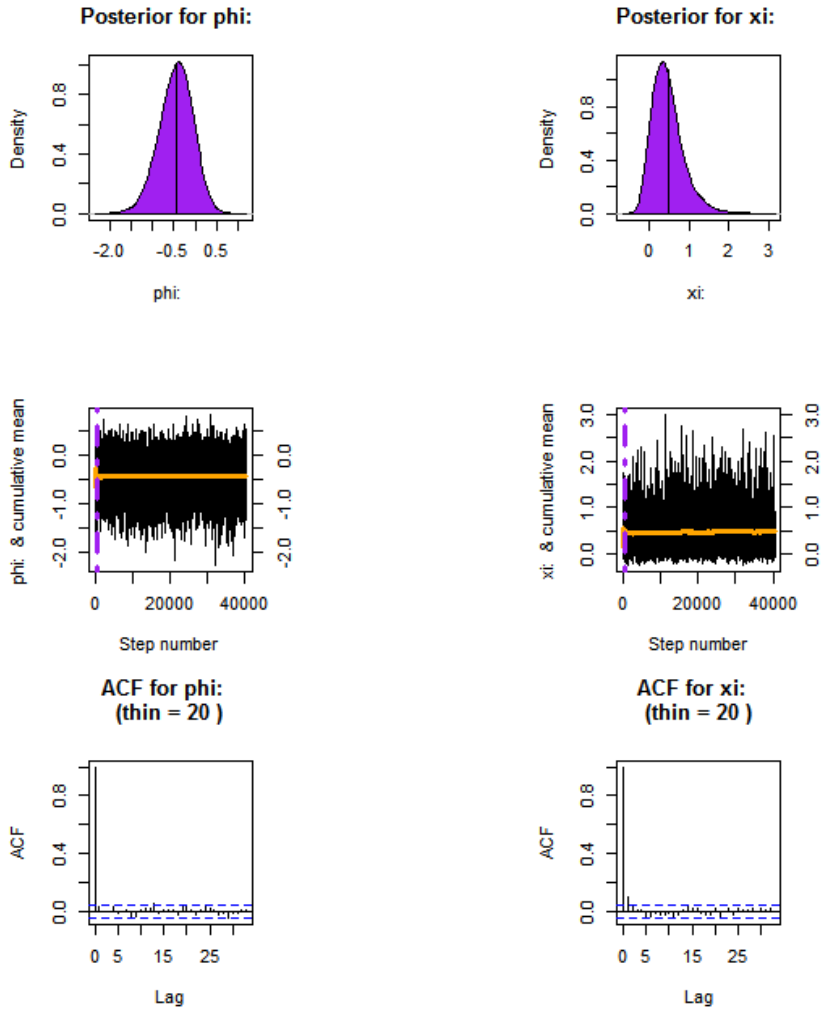

Figure 4.17: MCMC realizations for generalised Pareto distribution parameters in a Bayesian analysis.

The value for ξ in this case is 0.15 and for ϕ is 0.11.

4.6 Return levels after fitting the GEV and GPD using the Maximum likelihood and Bayesian approaches

Table(4.10) summarises the return levels associated with various return periods. The fifty year return level corresponds to 5.6 metres and finally the 100 year return level correspond to 6.759 metres. Prior to information depicted on the table, one would expect per example, that the extreme residuals generated from the best fitting model will exceed about 6.76 metres once in every 100years and will exceed about 5.6 metres every 50 years using the generalised extreme value distribution. As for the generalised Pareto distribution the fifty year return level corresponds to 2.45 metres and the 100 year return level corresponds to 3.181 metres. The confidence 95% confidence intervals were [1.48,9.72] and [0.799,12.72] for 50 and 100 years using the generalised extreme distribution and using the Generalised Pareto distribution they correspond to [1.88;3.02] and [2.21;4.15] for 50 and 100 year return period. From table (4.10), we have the return levels and the confidence interval bands for both the Generalised extreme value distribution and the Generalised Pareto distribution using both maximum likelihood estimation and the Bayesian approach for estimating the estimates and confidence intervals. In each case the first value is the lower bound, the second value is the return level and lastly the last value corresponds to the upper bound for confidence intervals, for example the 20 year return level for the GPD using the maximum likelihood estimation corresponds to (1.460;1.651;1.842) of which 1.46 is the lower bound, 1.651 is the return level and finally 1.842 is the upper bound. In a nutshell this shows that the Generalised extreme value distribution is the best fitting model as shown by the return levels with the corresponding return periods. Therefore for our prediction, for return levels we would rather use the Generalised extreme value

distribution using maximum likelihood approach instead of the Generalised Pareto distribution in order to predict the return levels and the return periods. The hundred year return levels using the Generalised extreme value distribution is comparable with the flood height for 2013 when the flow reached 6.707 metres using the maximum likelihood approach.

Table 4.10: Calculation of prediction Intervals .

year	GPD(MLE)	GPD(Bayes)	GEVD(MLE)	GEVD(Bayes)
20	(1.460;1.651;1.842)	(1.485;1.651;1.896)	(1.919;4.313;6.708)	(1.488;1.652;1.890)
30	(1.629;1.986;2.343)	(1.688;2.019;2.533)	(1.783;4.853;7.924)	(1.682;2.023;2.533)
40	(1.769;2.243;2.717)	(1.859;2.330;3.124)	(1.634;5.265;8.896)	(1.854;2.342;3.109)
50	(1.884;2.454;3.024)	(2.008;2.609;3.731)	(1.484;5.602;9.719)	(1.998;2.633;3.722)
60	(1.978;2.634;3.291)	(2.149;2.867;4.374)	(1.338;5.889;10.442)	(2.132;2.908;4.389)
70	(2.054;2.793;3.532)	(2.279;3.112;5.075)	(1.197;6.142;11.088)	(2.230;3.175;4.982)
80	(2.117;2.935;3.752)	(2.383;3.345;5.694)	(1.059;6.368;11.676)	(2.340;3.438;5.761)
90	(2.169;3.063;3.957)	(2.466;3.572;6.379)	(0.927;6.572;12.217)	(2.442;3.702;6.517)
100	(2.212;3.181;4.150)	(2.542;3.792;7.205)	(0.799;6.759;12.719)	(2.528;3.969;7.381)

4.7 Parametric bootstrap

The flood heights of Limpopo river at Beitbridge border post for the period under investigation is symmetric as shown by Figure (4.18).

4.7.1 Return level plot for the Generalised extreme value distribution

The return level plot in Figure (4.19) and (4.20) summarises various return levels with their corresponding return periods in a diagramatic form.

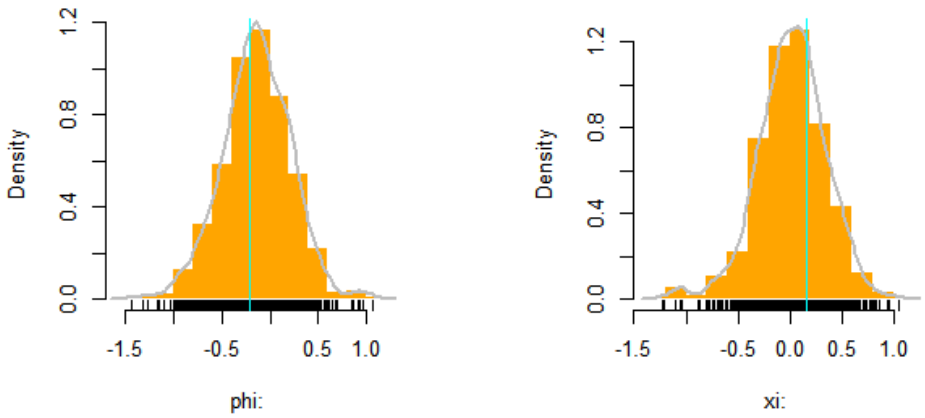

Figure 4.18: parametric bootstrap densities.

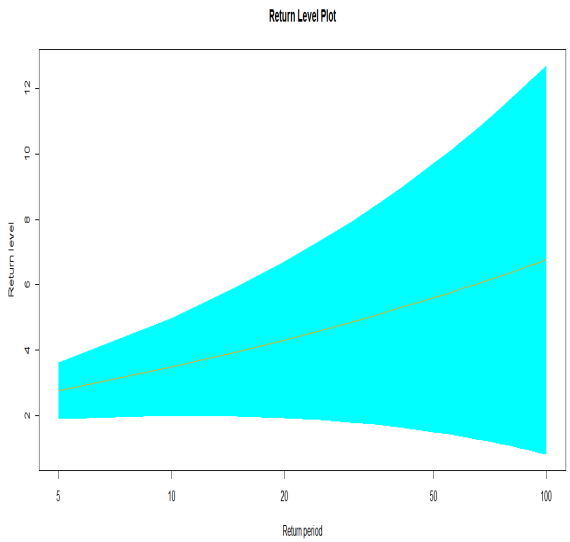

Figure 4.19: Return level plot for the generalised extreme value distribution using the Maximum likelihood approach to estimate parameters.

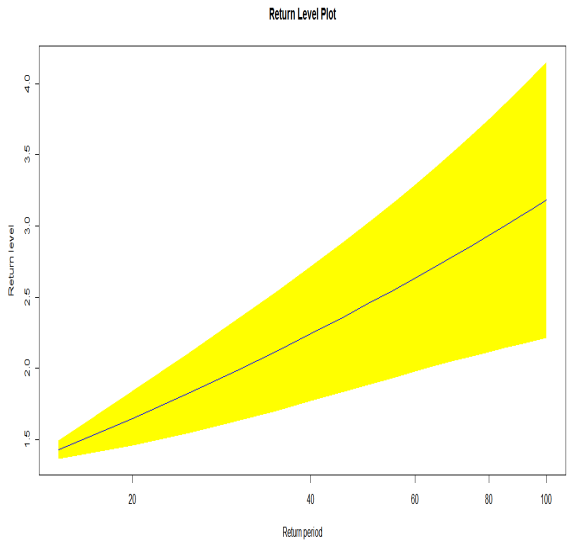

Figure 4.20: Return level plot for the Generalised Pareto distribution using the Maximum likelihood approach to estimate parameters.

4.7.2 Predictive densities for the Generalised extreme value distribution

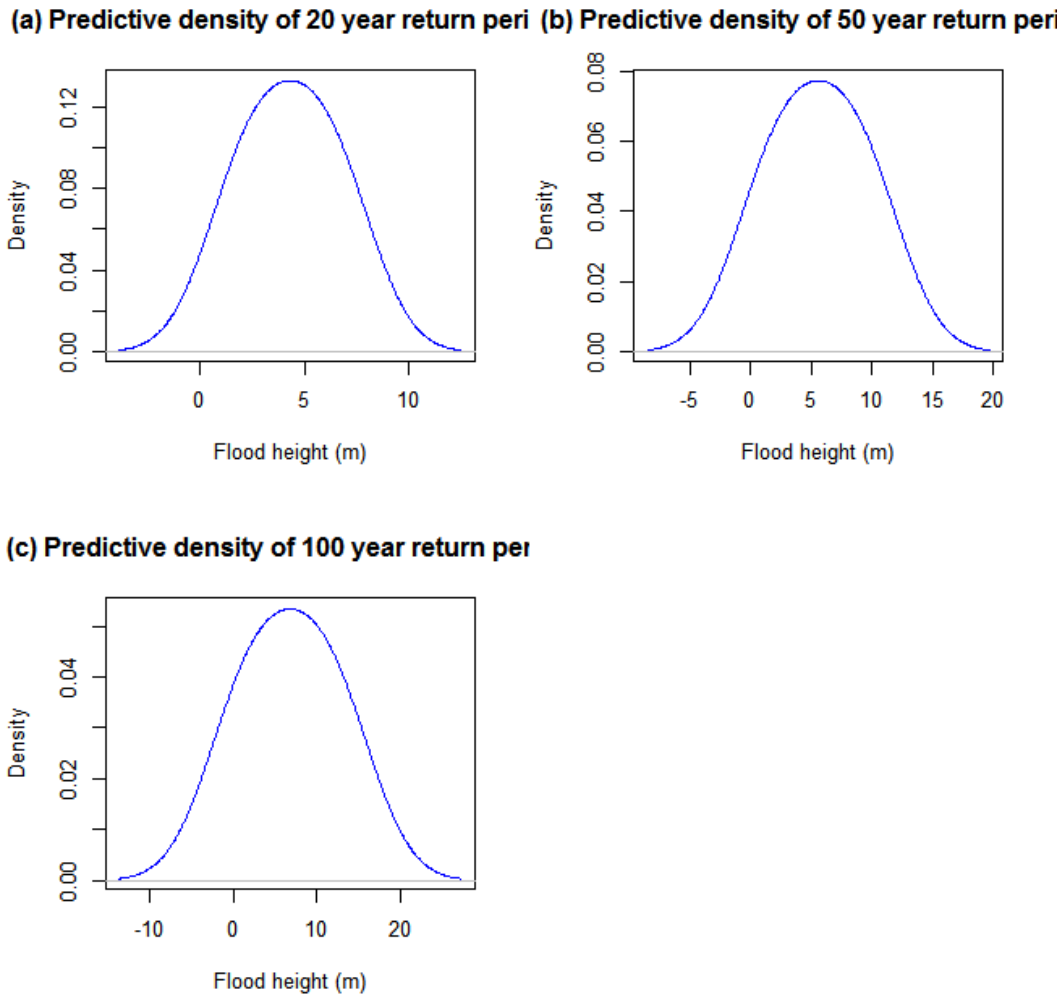

Figure 4.21: Top panel from left (a) predictive density for 20 year return period (b) predictive density for 50 year return period: Bottom panel from left(c) predictive density for 100 year return period extracted from the maximum likelihood estimation.

Figure (4.21) shows the predictive densities of return levels, and from this we can simply visualise the confidence interval bands.

4.8 Concluding Remarks

In this Chapter we have managed to show the suitable model that fits the Limpopo river at Beitbridge Border Post. Calculation of confidence intervals have also been dwelled upon in this chapter. Return periods and return levels calculation of Limpopo river at Beitbridge Border Post have also been exploited in this chapter. Even the descriptive statistics that includes the time series analysis, box plot, density plot and normal quantile- quantile plot for Limpopo river flood height data has also been given in this chapter. Of much importance is the inclusion of Bayesian EVT and r largest order statistics in this chapter.

Chapter 5

Conclusions

5.1 Introduction

This dissertation presents and analyses a number of statistical models for modelling flood heights of Limpopo river at Beitbridge border post. Modelling of flood heights is quite important in the field of hydrology for decision making. The stakeholders will be informed of return levels and return periods through modelling of the flood heights. This helps in decision making and to give an alarm to the community living around Limpopo river as to when they are likely to experience extreme floods.

5.2 Conclusion

The study considered the use of statistics of extremes in changing climate for Limpopo river at Beitbridge border post. The maximum likelihood estimation method and Bayesian approach were used to estimate the parameters of the Generalised extreme value distribution in the presence of a trend covariate and the Southern Oscillation

Index(SOI). The study has also revealed the importance of considering non stationary models when using statistics of extremes in a changing climate as these models provide an improvement in fit over the time homogeneous models.

The developed models established in this dissertation are consistent with cumulative(or moving sums) annual maximum series flood height and therefore appear reliable to use for flood frequency analysis. The use of the identified time dependent GEV models with a trend in the scale parameter at Beitbridge border post would also reduce sensitivity of the frequency of floods which is known to vary with changes in the scale parameter estimate, and therefore lead to more reliable estimates in the frequency of floods.

5.3 Limitations of the dissertation

(i) The raw data for Limpopo river at Beitbridge border post were originally recorded as monthly and yearly flood heights (or water levels). The data records at Beitbridge border post stretch back from 1951 up to 2014. However, due to missing values, the records used in this dissertation are for the period 1992 up to 2014.

(ii) This research focused on a single site analysis. It would be better to consider a multi-site model to reduce the large uncertainty caused by single site analysis Renard *et al.* (2006).

5.4 Findings and Contributions

(i) Modelling of flood heights at Beitbridge Limpopo river has not been done to the best of our knowledge.

(ii) Accurate estimation of return levels and periods of extreme flood heights helps in risk mitigation, for example, designing of bridges by civil engineers.

(iii) One of the main objectives of this study is risk analysis. It is possible now to quantify the damage caused by the flood height for this given period.

(iv) The Fréchet class of distributions is found to be the best fit to the data in all the modelling frameworks of this dissertation. This implies that the distributions of extreme flood heights for Limpopo river at Beitbridge border post are thin tailed.

5.5 Future research

The results of this dissertation provide possible areas for future research. The following possible future research directions are proposed:

(i) One area which needs future research is a probabilistic description and modelling of flood heights using Poison point process. This modelling approach helps in estimating the frequency of occurrences of flood heights for Limpopo river at Beitbridge border post.

(ii) Future research will consider multi-site regional analysis and compare the return periods and return levels for Limpopo river.

(iii) Inclusion in the EVT models (GEVD and GPD) more covariates in the form of cycles and or a physical variable such as dummy variable indicating the occurrence of cyclones in the region will also be considered in future research.

(iv) Future research will also explore the use of scoring rules such as the continuous probability score(CPRS)in assessing the predictive performance of the (GEVD) and (GPD) models.

References

[1] Castillo, E.(2012). *Extreme value theory in engineering.* Academic Press, London.

[2] Coles, S.(2001). *An introduction to statistical modelling of extreme values.* Springer-Verlag, London.

[3] Cyet, R.M and DeGroot, M.H. (1970). Multiperiod decision models with alternating choice as a solution to the duopoly problem. *The quarterly journal of Economics*, **84(3)**, pp: 410-429.

[4] Danielsson, J., Peng, L., and de Vries, C.G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of multivariate analysis*, **76(2)**, pp: 226-248.

[5] de Waal, D.(2009). Posterior predictions on river discharges.*Risk and decision analysis in maintenance optimization and flood management.*

[6] De Waal, J.H. (2012).Extreme rainfall distributions: analysing change in Western Cape. PhD thesis, Stellenbosch: Stellenbosch University, URL http://hdl.handle.net/123456789/187: accessed 2012-02-25.

[7] Ferreira, A., de Haan, L., (2015). On the block maxima method in extreme value theory: Pwm estimators. *The annals of statistics*, 43(1), pp: 276-298.

[8] Ferro, C. A. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal statistical Society: Series B (Statistical Methodology)*, 65(2), pp: 545-556.

[9] Fisher, R. A. and Tippet, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 24, pp:180-190. Cambridge Univ Press.

[10] Gilleland, E. and Katz, R. W.(2006).Analyzing seasonal to interannual extreme weather and climate variability with the extremes toolkit.*In 18th Conference on Climate Variability and Change, 86th American Meteorological Society (AMS) Annual Meeting,* Volume 29. Citeseer.

[11] Goegebeur, Y., Beirlant, J., de Wet, T., (2008). Linking pareto-tail kernel goodness of fit statistics with tail index at optimal threshold and second order estimation. *Revstat*, 6(1), pp: 51-69.

[12] Haile, A.T (2011). Regional flood frequency analysis for Southern Africa.

[13] Hall, P. (1990). Perfomance of balanced bootstrap resampling in distribution function and quantile problems. *Probability theory and related fields*, 85(2), pp: 239-260.

[14] Hefferson, J. E. and Southworth, H. (2012). Extreme value modelling of dependent series using r.

[15] Kagoda, P. A. (2006).*A comprehensive analysis of extreme rainfall*, Msc Thesis, University of Witswatersrand, Johannesburg, URL http://wiredspace.wits.ac.za: accessed October 2006.

[16] Kamwi, I. S. (2005)*Fitting extreme value distributions to the Zambezi river flood water levels recorded at Katima Mulilo in Namibia*, Msc thesis, University of the Western Cape, URL http://hdl.handle.net/11394/1834: accessed 2005.

[17] John Crag, C. (2010)*Non Parametric Smoothing in Extreme Value Theory* , Msc thesis, University of Cape Town, URL http://wiredspace.uct.ac.za:/11394/1834: accessed 2010.

[18] Kartz, R. W. (2010). Statistics of extremes in climate change. *Climate Change*, 100(1), pp: 71-76.

[19] Maposa, D., Cochran, J., Lesaoana, M., and Sigauke, C. (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo river basin of Mozambique using a model based Bayesian approach. *Natural Hazards and Earth System Sciences Discussions*, 2(8), pp: 5401-5425.

[20] Marin, J. M., Diez, R. M., and Insua, D. R. (2003). Bayesian methods in plant conservation biology. *Biological Conservation*, 113(3), pp: 379-387.

[21] Matthys, G., Delafosse, E., Guillou, A., and Beirlant, J. (2004). Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance: Mathematics and Economics*, 34(3), pp: 517-537.

[22] Melice, J. L. and Reason, C.J. (2007). Return period of extreme rainfall at George, South Africa. *South African Journal of science*, 103(11-12), pp: 499-501.

[23] Orbe, J., Ferreira, E., and Nunez-Anton, V. (2003). Censored partial regression. *Biostatistics*, 4(1), pp: 109-121.

[24] Pocernich, M. J. (2002).*Application Of Extreme Value theory and threshold models to hydrological event*, Msc thesis, University of Colorado, URL http://ucdenver.edu :accessed 2002.

[25] Pickands 111, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3, pp: 119-131.

[26] Poirier, D. J. (1988). Frequentist and subjectivist perspectives on the problems of model building in economics. *The Journal of Economic Perspective*, 2(1), pp: 121-144.

[27] Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 25, pp: 111-164.

[28] Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modelling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5), pp: 20-25.

[29] Salarpour, M., Yusop, Z., Yusof, F., Shahid, S., and Jajarmizadeh, M. (2012). Flood frequency analysis based on t-copula for johor river, Malaysia, *Journal of Applied Sciences*, 13(7), pp :1021-1023.

[30] Scarrot, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1), pp: 33-60.

[31] Singo, L., Kundu, P., Odiyo, J., Mathivha, F., and Nkuna, T. (2012). Flood frequency analysis of annual maximum stream flows for Luvuvhu river catchment, Limpopo province, South Africa. *Unpublished work*, pp: 1-7.

[32] Sithole, A. and Murewi, C. T.(2009). Climate variability and change over Southern Africa: impacts and challenges. *African Journal of Ecology*,47(1), pp :17-20.

[33] SMITH, R. (1987). Estimating tails of probability distributions. *The annals of Statistics*, pp:1174-1207.

[34] Smith, R. L. (1989) Extreme value analysis of environmental time series:an application to trend detection in ground-level ozone. *Statistical Science*,4(4), pp: 367-377.

[35] Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp: 515-528.

[36] Tadesse, A. H. (2011).*Regional Flood Frequency Analysis in Southern Africa*, Msc thesis, University of Oslo, URL http://duo.uio.no :accessed 2011.

[37] Tawn, J.A (1988). An extreme value theory model for dependent observations. *Journal of hydrology*, 101(1), pp: 227-250.

[38] Van Dongen, S. (2006). Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*, 242(1), pp: 90-100.

[39] Von Mises, R. (1936). La distribution de la plus grande de n valeurs, *reprinted in Selected Papers 11, Amer. Math.Soc., Providence, Rhode Island*,1954, pp: 271-294.

[40] STEPHENSON, A.G. and RIBATET, M.A. (2006). A Users Guide to the evdbayes Package (Version 1.1). http://cran.r-project.org.

[41] COLES, S.G. and POWELL, E.A. (1996).Bayesian methods in extreme value modelling: A review and new developments. Int. Statist. Rev., 64, 119136.

[42] BALKEMA, A. and DE HAAN, L. (1974).Residual life time at great age. Annals of Probability, 2, 792804.

[43] BEIRLANT, J., GOEDGEBEUR, Y., SEGERS, J. and TEUGELS, J. (2004). Statistics of Extremes Theory and Applications: Wiley, England.

[44] SMITH, R. (1985).Estimating tails of probability distributions. The An- nals of Statistics, 15(3), 11741207.

[45] SMITH, R. (1987).Estimating tails of probability distributions. The Annals of Statistics, 15(3), pp 1174-1207.

[46] McNeil, A. J. (1999).Extreme value theory for risk managers. Departement Mathematik ETH Zentrum.

Appendix

Some selected R codes

```
attach(mheight)
head(mheight)
tail(mheight)
min(height)
max(height)
###################################
############################
#decluster the sequence by using the automatic declustering method
win.graph()
ei <- extremalIndex(height, threshold = 1.33)
ei
dc <- declust(ei)
par(mfrow=c(1,1))
plot(dc,col="blue", xlab="Observation number", ylab="Positive Detrended SDPED")
dc
dc <- declust(height, threshold = 1.33)
##Fitting the Generalised Pareto distribution to cluster maxima
resid.gpd <- evm(dc) #MLE
resid.gpd
par(mfrow = c(2, 2))
plot(resid.gpd)
#GPD parameter uncertainty: PARAMETRIC BOOTSRAP
boot <- evmBoot(evm(dc), trace = 1001)
summary(boot)
par(mfrow = c(1, 2))
plot(boot)
```

Some selected R codes continue

```
attach(mheight)

head(mheight)

tail(mheight)

library(texmex)

min(height)

max(height)

###############################################################################################

###############################################################################################

#decluster the sequence by using the automatic declustering method

win.graph()

ei <- extremalIndex(height, threshold = 1.33)

ei

dc <- declust(ei)

par(mfrow=c(1,1))

plot(dc,col="blue", xlab="Observation number", ylab="Positive Detrended SDPED")

dc

dc <- declust(height, threshold = 1.33)

##Fitting the Generalised Pareto distribution to cluster maxima

resid.gpd <- evm(dc) #MLE

resid.gpd

par(mfrow = c(2, 2))

plot(resid.gpd)

#GPD parameter uncertainty: PARAMETRIC BOOTSRAP

boot <- evmBoot(evm(dc), trace = 1001)

summary(boot)

par(mfrow = c(1, 2))

plot(boot)
```

Some selected R codes continue

```
win.graph()\\
portRL $<-$ predict(residpos, M = seq(20, 100, by = 20), ci.fit = TRUE)\\
plot(portRL)\\
portRL[c(1, 2, 3)]\\
par(mfrow=c(2,2))\\
y=c(1.48,1.64,1.89)\\
plot(density(y),xlab="Flood height (m)",
    main="(a) Predictive density of M.20")\\
x=c(1.85,2.28,3.14)\\
plot(density(y),xlab="Flood height (m)",\\
    main="(b) Predictive density of M.40")\\
z=c(2.12,2.75,4.46)\\
 plot(density(y),xlab="Flood height (m)",\\
    main="(c) Predictive density of M.60")\\
```

Some selected R codes continue

```
attach(datasoi)

head(datasoi)

tail(datasoi)

win.graph()


library(ismev)

ffit = gev.fit(fheight)

gev.diag(ffit)


plot(fheight)

plot(soi)


plot(datasoi[, c("soi", "fheight")],col="blue",

    xlab="October SOI (1955-2014)",ylab="Flood height")



ti=matrix(ncol=1,nrow=60)

ti[,1]=seq(1,60,1)

covar=matrix(ncol=2,nrow=60)

covar[,1]=datasoi[,3] #Stores the SOI values in column 1 of covar

covar[,2]=ti#Stores the time indicator in column 2 of covar

c = gev.fit(datasoi[,2])#original

gev.diag(c)


plot(x=SOI, y=r1, xlab="October SOI",

    ylab="Annual maximum flood height(m)",

    col="blue", xlim=c(-21,20), ylim=c(0,14))
```

R codes for R largest order statistics

```
attach(j)

head(j)

win.graph()

a=ts(shape)

plot(a,type="b",xlab="r",ylab="Shape parameter of GEV",col="blue",ylim=c(-0.8,2))

lines(lower.bound,col="red",type="c")

lines(upper.bound,col="green",type="c")

#min(lower.bound)

#max(upper.bound)
```

R codes for Correlation plots and correlation coefficients

```
attach(datasoi1)

head(datasoi1)

tail(datasoi1)



#create matrix from vectors

M<- cbind( heighty,Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov,Dec)

# covariance and corelation matrices

cov(M)

cor(M)

win.graph()

par(mfrow=c(2,3))

plot(Nov,heighty)

plot(Dec,heighty)

plot(Jan,heighty)

plot(Feb,heighty)

plot(Mar,heighty)

plot(Oct,heighty)
```

R codes for calculation of return levels using Bayesian approach

```
#Calculation of Return levels using Bayesian approach for the GEVD model.

head(yheight1)

library(texmex)

palette(c("black", "purple", "cyan", "orange"))

set.seed(20120118)

mpedc <- evm(heighty, family = gev)

mpedc

plot(mpedc)

residpos =update(mpedc, method = "simulate", trace = 40001, penalty = "gaussian")

par(mfrow = c(3,3))

plot(residpos)

residpos

residpos <- thinAndBurn(residpos, burn = 500, thin = 20)

dim(residpos$param)

summary(residpos)

par(mfrow=c(3,2))

plot(residpos)

residpos

win.graph()

predict(residpos, type = "lp")

predi <- predict(residpos, M = seq(15, 100, by = 5), ci.fit = TRUE)

plot(predi)

predi[c(1, 2, 3,4,5,6,7,8,9,10,11,12,13,14,15,16)]
```

R codes for calculation of return levels using Maximum likelihood approach

```
#GEVD Maximum Likelihood


attach(yheight1)

head(yheight1)

library(texmex)

palette(c("black", "purple", "cyan", "orange"))

set.seed(20120118)

mpedc <- evm(heighty, family = gev)

mpedc

plot(mpedc)

win.graph()

predict(mpedc, type = "lp")

predi <- predict(mpedc, M = seq(15, 100, by = 5), ci.fit = TRUE)

plot(predi)

predi[c(1, 2, 3,4,5,6,7,8,9,10,11,12,13,14,15,16)]
```

R codes for calculation of return levels using Maximum likelihood approach

```
attach(mheight)

head(mheight)

tail(mheight)

min(height)

max(height)

###################################

############################

#decluster the sequence by using the automatic declustering method

library(texmex)

win.graph()

ei <- extremalIndex(height, threshold = 1.33)

ei


dc <- declust(ei)

par(mfrow=c(1,1))

plot(dc,col="blue", xlab="Observation number", ylab="Positive Detrended SDPED")

dc

dc <- declust(height, threshold = 1.33)

##Fitting the Generalised Pareto distribution to cluster maxima

resid.gpd <- evm(dc) #MLE

resid.gpd

par(mfrow = c(2, 2))

plot(resid.gpd)

#####################################################

#Prediction using MLE

predict(resid.gpd, type = "lp")

portRL <- predict(resid.gpd, M = seq(15, 100, by = 5), ci.fit = TRUE)
```

R codes for calculation of return levels using Bayesian approach

```
attach(mheight)

head(mheight)

tail(mheight)

min(height)

max(height)

###################################

###################################

#decluster the sequence by using the automatic declustering method

library(texmex)

win.graph()

ei <- extremalIndex(height, threshold = 1.33)

ei

dc <- declust(ei)

par(mfrow=c(1,1))

plot(dc,col="blue", xlab="Observation number", ylab="Positive Detrended SDPED")

dc

dc <- declust(height, threshold = 1.33)

#GPD parameter uncertainty: BAYESIAN ESTIMATION

#residpos <- evm(r2pos, data = data, qu = 0.90, "simulate")

m1<- evm(dc) #ML

m1

residpos =update(m1, method = "simulate", trace = 40001, penalty = "gaussian")

par(mfrow = c(3,3))

plot(residpos)

residpos

residpos <- thinAndBurn(residpos, burn = 500, thin = 20)

dim(residpos$param)
```