MODELLING AVERAGE MONTHLY RAINFALL FOR SOUTH AFRICA USING EXTREME VALUE THEORY

by

DANIEL MASHISHI

Submitted in fulfillment of the requirements for the degree of

MASTER SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Dr D Maposa

CO-SUPERVISOR: Prof M Lesaoana

2020

Declaration

I, **Daniel Mashishi**, the undersigned, hereby declare that the work contained in this dissertation submitted to the University of Limpopo, for the degree of Master of Science in Statistics is my original work, and that any work done by others or by myself previously has been, acknowledged and referenced accordingly.

Signature:.....Date:.....

Abstract

The main purpose of modelling rare events such as heavy rainfall, heat waves, wind speed, interest rate and many other rare events is to try and mitigate the risk that might arise from these events. Heavy rainfall and floods are still troubling many countries. Almost every incident of heavy rainfall or floods might result in loss of lives, damages to infrastructure and roads, and also financial losses. In this dissertation, the interest was in modelling average monthly rainfall for South Africa using extreme value theory (EVT). EVT is made up mainly of two approaches: the block maxima and peaks-over threshold (POT). This leads to the generalised extreme value and the generalised Pareto distributions, respectively. The unknown parameters of these distributions were estimated using the method of maximum likelihood estimators in this dissertation. According to goodness-of-fit test, the distribution in the Weibull domain of attraction, Gumbel domain and generalised Pareto distributions were appropriate distributions to model the average monthly rainfall for South Africa. When modelling using the POT approach, the point process model suggested that some areas within South Africa might experience high rainfall in the coming years, whereas the GPD model suggested otherwise. The block maxima approach using the GEVD and GEVD for r-largest order statistics also revealed similar findings to that of the GPD. The study recommend that for future research on average monthly rainfall for South Africa the findings might be improved if we can invite the Bayesian approach and multivariate extremes. Furthermore, on the POT approach, time-varying covariates and thresholds are also recommended.

Dedication

I would like to dedicate this work to my father Ephraim Mashishi and my mother Emily Kokobele and also my family.

Acknowledgments

First of all, I would like to thank God for giving me strength and wisdom to complete this dissertation. I would like to express my sincere appreciation and thanks to my supervisor Dr Daniel Maposa and my co-supervisor Prof Maseka Lesaoana for their guidance throughout this dissertation. I would also like to thank my fellow MSc students at University of Limpopo and my colleague Ms Makwelantle Sehlabana for their moral support towards the completion of this dissertation. Also, I would like to extend my gratitude to National Research Foundation (NRF) and South African Weather Service (SAWS) for their contribution in this dissertation in terms of funding and data supply, respectively. Lastly, special thanks goes to my family, more especially my father.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

A-D	Anderson-Darling
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CDF	Cummulative Density Function
DF	Dickey-Fuller
EVT	Extreme value theory
ICU	Intensive Care Unit
GEV	Generalised Extreme Value
GEVD	Generalised Extreme Value Theory
GPD	Generalised Pareto distribution
K-S	Kolmogorov-Smirnov
MM	Milimeters
MLE	Maximum Likelihood Estimator
POT	Peaks-Over-Threshold
SADC	Southern African Development Community
SAWS	South African Weather Service

Chapter 1

Introduction and background

1.1 Introduction

Civil engineering was the early discipline to apply extreme value models to design structures that can withstand forces exerted upon them (Coles, 2001). There are several studies of extreme events that have produced accurate and useful results around the world (Coles, 2001). For example: extreme value theory is used as a risk management tool in finance (Embrecht et al., 1999); drought in the Western Cape province, South Africa (Khuluse, 2010); heavy rainfall of Tanzania (Ngailo et al., 2016); maximum temperature in South Africa (Nemukula and Sigauke, 2018); and many more.

Extreme value theory (EVT) is defined as one of the superior approaches to measure the stochastic behaviour of a process at unusually high or low levels (Chikobvu and Chifurira, 2015). In particular, EVT provides the statistical framework to make inferences about the probability of very rare and extreme events. Fernandez (2003) indicated how, over the past 50 years, EVT has been

considered as one of the most important statistical disciplines in the applied sciences. EVT has also been applied extensively in many areas spanning: environmental phenomena such as floods and droughts (Masereka et al., 2018; Chikobvu and Chifurira, 2005); finance and insurance (Kratz, 2017; Adesina et al., 2016; Henry III and Hsieh, 2014); survival analysis (Alvarez-Igiesias, 2015); food and telecommunications (Ngailo et al., 2016; Bali, 2003).

In environmental phenomena, EVT can reveal useful information about river-level, wind speed, etc. The information can be essential in the design of structures such as bridges, buildings and roads (Smith, 2005). In finance, EVT can be useful in assessing the risk of large insurance claims and in measuring risks that arise from certain financial institutions. For rainfall, the information can be helpful in predicting the rare events like tornadoes, floods and thunderstorms. The application of EVT in rainfall data is becoming more visible around the globe (De Waal et al., 2017; Chu et al., 2008; Smith, 2005). According to Smith (2005), this area of statistics is not restricted to rainfall data only but also other disciplines. In medicine, EVT is used to build a device that can determine which patients need more care (Hugueny et al., 2010).

1.2 Background

Worldwide, climate change has become a disturbing issue for the past few years. South Africa has approximately 55 million people, with 20% depending on rain-fed agriculture (Khuluse, 2010). As a result, many South Africans face challenges regarding the availability of staple food such as maize and rice. Martin et al. (2011) investigated how to tackle climate change in South Africa and concluded that agriculture, bio-diversity, water and health, are the most affected sectors.

In agriculture, the livelihoods of many individuals are in danger. Their jobs are at stake because crops are failing and livestock are dying. Heavy rainfall, hot spell and drought can contribute negatively towards health. In addition, it is believed that these factors might cause or transport diseases like malaria from one community to another and might also cause food insecurity, hunger and malnutrition. Water is essential to humans and other creatures alike. Currently some areas within South Africa are expirencing severe drought because the demand for water is greater than supply. Kulshreshtha (1998) projected that by the year 2025, in some regions of the world, especially in the developing countries, the demand for water will surpass the supply.

In the recent past decades, climate-related disasters have been the main concern worldwide (Meehl et al., 2006). According to Shongwe et al. (2009) temperature and precipitation extremes have showed significant increase in the Southern African region. It was reported that, in the affected regions floods and drought-related incidents have increased recently, leaving the livelihoods of approximately six million people in danger (Reason et al., 2005). Shongwe et al. (2009) investigated the extreme precipitation in Africa under global warming. Their study recommended that countries in the Southern African Development Community (SADC) region must give more attention to extreme precipitation due to its great impact on human activities. Our study aims at filling this gap by modelling the tail behaviour of the underlying distribution of the average monthly rainfall for South Africa using extreme value theory.

Taiwan received massive rainfall during the rainy seasons of 2009, reaching 2235 milimeters (mm) in a period of two days (fen Chu et al., 2012). Tanzania depends on rain-fed agriculture. In almost every rainy season, the country experiences floods, leaving their agricultural vegetation at risk. Another extreme event in SADC, is drought, which caused many Zimbabwean farmers to lose

their crops and livestock during 1991-1992 (Chikobvu and Chifurira, 2015).

Limpopo province in South Africa, consists of relatively large and poverty-striken households and also has a wide range of national parks. The region experiences severe drought and heavy rainfall in almost every rainy season (October to January). Reason et al. (2005) revealed how the region was affected by drought and severe flooding during the years 2001-2004. Moreover, these aforementioned incidents pose a threat to biodersity, residents and infrastructure.

The Western Cape province has its unique climate in South Africa. The weather variables in the province can be categorised as follows: rainfall, temperature and wind (Bhagwandin, 2017). During the years 2001-2008, the province was hit by severe floods which resulted in a loss of 2.5 billion rand in damages to property (de Waal, 2012). This observation was also supported by de Waal et al. (2017), who investigated extreme 1-day rainfall distributions during the 2001-2014 period. In South Africa, there are numerous extreme events related to heavy rainfall and drought. However, it was cautioned that the country must expect a decrease in precipitation level, especially in the Western Cape and Northern Cape provinces (Khuluse, 2010).

Although the classical statistical techniques have been applied widely to other studies, the use of EVT in modelling environmental phenomena is gaining more attention worldwide due to its capacity to handle rare events. According to Nemukula and Sigauke (2018) modelling with classical statistical techniques might lead to inaccurate estimations because their results are based on the mean and not on the tails of the distribution. It was also revealed that EVT is the most appropriate method that can deal with such situations. In the present study, EVT is the preferred methodology since the study deals with

extreme rainfall events.

1.3 Problem statement

Globally, there are several upheavals caused by extreme rainfall events that resulted in major damages to public infrastructure and agriculture, and loss of lives. For example, Chu et al. (2008) investigated extreme rainfall in Hawaiian Islands. The scholars stated that the Islands received massive rainfall of 940 mm within 24 hours, which destroyed many households and roads in the year 2004. In the same study, it was further revealed that the Islands lost about 88 million US dollars in damages.

There are several extreme rainfall events around South Africa that left many people homeless and caused huge damages in a number of areas in the provinces of Limpopo, Gauteng, Kwa-Zulu Natal, Eastern Cape and Western Cape (Kruger and Nxumalo, 2017; Phakula, 2016; Diriba et al., 2014; de Waal, 2012; Khuluse, 2010; luc Melice and Reason, 2007). For example, a study by Khuluse (2010) highlighted that the Western Cape province was hit by extreme rainfall between the years 2003 and 2008 that caused damages on property and infrastructure worth R5 billion. The same study also revealed that South Africa should expect a decline in average precipitation by the end of the 21st century, which may result in less than 1 000 mm per year by 2025. Other extreme events occuring in South Africa includes the KwaZulu-Natal heavy rainfall and floods in which about 51 people were confirmed dead and some were forced to relocate to neighbouring places (EWN, 2019).

However, the problems that extreme rainfall events present to the government of South Africa, the private sector and other communities, such as loss of livestock, damages to property and infrastructure, cannot be over emphasised.

Most studies on extreme rainfall events in South Africa have been mainly at a local site level or regional level, but not the entire country. Of paramount importance is the prediction of the return periods in reducing the predictive uncertainty of these extreme rainfall events at the national level, and hence reducing their disastrous effects on human life and property. It is, therefore, our intention in this study to model average monthly extreme rainfall for South Africa using the two realisations of EVT, i.e. block maxima and peaks-over threshold (POT).

1.4 Rationale

A study by luc Melice and Reason (2007) showed how the business of Garden Route in South Africa was affected by the 2006 extreme rainfall for about two days. On average, every strike of flood results in either financial loss or loss of lives in the country. Another evidence by Dyson and van Heerden (2001) revealed how the February 2000 extreme rainfall affected the Northern Cape and Mpumalanga provinces, which both recorded an annual average rainfall of 1 000 mm. In their study, nearly 600 people lost their lives and many others were relocated to neighbouring villages for safety.

Reason et al. (2005) also revealed how the extreme conditions during the year 2000 affected the Kruger National Park and other forms of life in the Limpopo province. In a separate study in Cape Town, Diriba et al. (2017) presented some evidence on how extreme wind speed influenced the wildfire that destroy 4000 hectares of land vegetation and some homes. Their study also established how the extreme rainfall affected the traffic in the Gauteng province in 2017. Therefore, it is important to study the patterns of these extreme events in order to develop methods that will produce accurate predictions. This will, in turn help to initiate measures, forestall the negative effects of these occurrences

caused by the extreme rainfalls.

1.5 Aim and objectives of the study

1.5.1 Aim

The aim of the study is to model the average monthly rainfall for South Africa using extreme value theory.

1.5.2 Objectives

The objectives of the study are to:

(i) Test for randomness and stationarity using average monthly rainfall data for South Africa.

(ii) Find a suitable candidate parent distribution(s) for average monthly rainfall for South Africa.

(iii) Predict the return levels and their corresponding return periods using the fitted parent distribution(s).

(iv) Use the block maxima approach to model extreme average monthly rainfall for South Africa.

(v) Use the peaks-over threshold approach to model extreme average monthly rainfall for South Africa.

1.6 Structure of the dissertation

This section describes the structure of the dissertation. The research study consists of six chapters, including references.

Chapter 1 gives the introduction on impacts of extreme rainfall to society. It also provides information about the climate variability in South Africa. Literature review in Chapter 2 presents diverse studies of extreme rainfall in South Africa and other countries around the world. Chapter 3 presents the methodology adopted in the dissertation. It also describes, in detail, what extreme value theory entails. In addition, the approaches of extreme value theory are explained. Discussion and results are presented in Chapter 4, including a detailed extreme value analysis of average monthly rainfall data for South Africa. In addressing the objectives set in Chapter 1, Chapter 5 presents the concluding remarks and recommendations based on the results of the analysis in Chapter 4.

Chapter 2

Literature review

2.1 Introduction

This chapter addresses the theory of extreme events and modelling. Also, an overview of the probability and statistical tools underlying the extreme value theory (EVT), is provided. In addition, the chapter presents the history behind the rainfall patterns in South Africa and some other parts of the world.

Several studies on the applications of EVT in various disciplines have been conducted. In modelling financial risk measurement using the generalised extreme value distribution (GEVD), Bali (2007) concluded that the loss of financial institutions can be accurately estimated using generalised extreme value distribution. In a separate study on patients in intensive care unit (ICU) in the United Kingdom, Hugueny et al. (2010) used EVT to build a probabilistic detector to identify patients who are in a deterioration state. Their study revealed that about 20 000 unforeseen patients admitted to ICU could have been avoided if they had this detector. Another study by Bali (2003) used EVT in

finance to prove that the structure of interest rate volatility does not follow a normal distribution, and concluded that the interest rate swift occurs more frequently than predicted by the normal model. In the study of risk management concerning enviromental phenomena, Nemukula and Sigauke (2018) used EVT to model daily maximum temperature and the results of their study predicted an increase in daily temperature for the forthcoming years. This shows how useful EVT is and how it has emerged in various disciplines.

2.2 World Rainfall

Some studies have highlighted factors responsible for climate change in Africa. Deforestation and atmospheric gas are considered to be the contributing factors towards climate change in Africa (Martin et al., 2011; Smith, 2005).

The climate of South Africa and other sub-Saharan countries consists of the rainy summer with cold and dry winter. These seasonalities are brought by anticyclonic high pressure system that happens during winter; and intermittent thermal trough during summer (Tadross and Johnston, 2012). Their study also established that the summer trough was responsible for producing greater rain over the eastern side and less rain towards the western part of the continent. In winter, the anticyclones over the Atlantic and Indian Oceans shift and unite over land, which creates the dry conditions in Africa (Tyson and Preston-Whyte, 2011).

According to Tadross and Johnston (2012), South Africa consists of 11 climatic regions. For these 11 climatic regions, 35 percent have a precipitation margin of 500 mm; 44 percent have a precipitation of 200-500 mm; and the remaining 21 percent have a precipitation of less than 200 mm. The two phenomena that pose threat to human life are rainy weather and drought. For a certain country

to have a rainy weather, it is believed that, that country must first deal with drought or vice versa. According to Tyson and Preston-Whyte (2011), dry spells are related to rainy weather in many African countries. This implies that, if a certain country experiences drought for some years, then that particular country will have much better rainy weather in the years to follow.

Cumulus convection clouds is the rain generating system responsible for producing greater rain (Tyson and Preston-Whyte, 2011). However, most areas experiencing summer rainfall within South Africa receive much of their rain in the afternoon and early evening. Regions experiencing winter rainfall, receive much of their rain at night and in the morning. Drought is brought about by living habits of the residents within a region or country. Evidence presented by Tyson and Preston-Whyte (2011) stated that drought is prevalent in those regions that depend more on natural resources of food, water and energy. Botswana, Namibia and South Africa are at the brink of becoming potential desserts (Tyson and Preston-Whyte, 2011).

2.3 Rainfall in the world

The rainy season is an exciting season for farmers and all living creatures and living things in the world. Farmers are assured that their agricultural vegetation will receive sufficient amount of water for generating quality product. However, in some countries or regions, instead of having normal rainfall, there is abnormal rainfall, which turns into floods (Goudenhoofdt et al., 2017; Chu et al., 2008).

Several extreme events around the globe often leave many people homeless and also cause huge damages to property and agricultural land, including deaths

(Ngailo et al., 2016; Khuluse, 2010; Chu et al., 2008). The study by Reason et al. (2005) presented evidence on how heavy rain can be harmful to agricultural vegetation. Their study reported that floods can flush away the seeds during the early stages of ploughing. Dyson and van Heerden (2001) presented evidence of heavy rainfall that affected the people of Mozambique, South Africa and Zimbabwe. Their report showed that the Limpopo province suffered a severe loss of R1.3 billion in infrastructure and roads, and around 200 bridges were also destroyed in South Africa. In Mozambique and Zimbabwe, nearly 600 people lost their lives and several others were forced to leave their homes. In a separate study, Chu et al. (2008) showed how the Hawaiian Islands were affected by heavy rainfall. They reported that this heavy rainfall had devastating effects on infrastructure and caused some disruptions at the University of Hawai. Another extreme rainfall in China affected thousands of lives, and some went missing (Ender and Ma, 2014).

Nadarajah (2005) investigated the extremes of daily rainfall in west central Florida. The data was extracted from 14 rainfall stations. It was concluded that the Frechet distribution was the best distribution that can represent the data. The quantile estimates of the return period showed an increase in daily rainfall in west central Florida. In China, Ender and Ma (2014) presented some evidence on extreme precipitation for four cities. Their results showed that GPD was more preferred to GEV when it came to goodness-of-fit. The quantile estimates suggested that China was expected to have abundant rainfall every five years. In another study, Ngailo et al. (2016) modelled extreme rainfall of Tanzania using GEV and GPD. The return levels of both GEV and GPD showed an increase in rainfall in Tanzania and also revealed that the country was expected to experience extreme maximum rainfall every ten years.

Another extreme event that has tormented many people is called the minimum

rainfall or drought. Many studies have pointed out to the importance of understanding the behaviour of minimum rainfall (Chikobvu and Chifurira, 2015; Chifurira and Chikobvu, 2014; de Waal, 2012; Khuluse, 2010 and more). These studies have shown that in Southern Africa, drought tends to occur between the December to March rainy season.

The study by Aguilar et al. (2009) showed that central African countries must expect an increase in extreme temperature and a decrease in precipitation in the coming years. This decrease in precipitation was also highlighted by Jury (2012). Nkrumah (2017) indicated that Ghana is affected by extreme temperature. According to Nkrumah (2017), some areas within the Ghana would experience extreme temperature once every five years. In the same study, the index parameter suggested that the underling distributionof the tail index lies in the Weibull domain of attraction.

There have been some other developments in modelling of extreme value statistics. Recently, some researchers used point processes and EVT to make inference about extreme rainfall (Khuluse, 2010). According to Coles (2001), the use of point processes approach is related to peaks-over threshold approach. In their study, it was stated that the results obtained from point processes approach are most likely to be similar to those adopting the peaks-over threshold approach. However, the Poisson point process model is used to check if our estimation in POT approach will agree with that of the Poisson point process. There are several studies which used point process and EVT approaches to model extreme rainfall data (Cowpertwait et al., 2001 Coles, 2001).

The shift in climate change has affected many countries around the world (Nemukula and Sigauke, 2018; Masereka et al., 2018; Chikobvu and Chifurira, 2015). It is clear that this shift has affected the behaviour of the temperature,

rainfall, wind, etc. However, further research still needs to be conducted on climate change in order to prevent loss of lives and unnecessary expenditure.

2.4 Rainfall in South Africa

South Africa has a wide range of varying climatic conditions than any other country in sub-Saharan Africa. Again, in comparison to other countries within the same range of latitude, South Africa has the most freezing temperatures. However, according to Phakula (2016) these variabilities have impacted negatively towards agriculture, economy and water resources. Several studies in South Africa have shown that most parts of the country experience a massive decrease in precipitation and an increase in warm temperatures (Nemukula and Sigauke, 2018; Khuluse, 2010).

De Waal et al. (2017) used the generalised Pareto distribution to model 1-day rainfall distributions of the Western Cape Province from 76 rainfall stations. Their results predicted an increase in the 50-year return period of 1-day rainfall patterns for 48 stations, while for the remaining stations, the converse was the case. The findings from the Western Cape is a source of concern to residents because it may give rise to extreme events like floods and thunderstorm since the province receives more rainfall during these two extreme events.

Masereka et al. (2018) used empirical continuous probability distribution functions and theoretical continuous probability functions to model annual maximum rainfall in Nelspruit, South Africa. Their findings suggested that the region must wait for about 10 years before it can receive another massive rainfall. Du Plessis and Burger (2015) investigated the short-duration rainfall intensities in South Africa using EVT. Their study concluded with no evidence supporting the increase in rainfall intensities. The information about extreme

events is important for South Africa. It can unveil the usefulness of information to manage floods and droughts. It can also assist in reducing the amount of money spent by the government and insurance companies on disaster relief operations, property recovery and loss of lives (Maposa, 2016).

Nemukula and Sigauke (2018) investigated the use of r-largest order statistics in modelling average daily temperature in South Africa. The results of their study showed an increase in average daily temperature in the coming years, and also that the negative Weibull distribution was a good fit for the data. The study by Debusho and Diriba (2016) presented another version of information on extreme temperature in the Eastern Cape province in South Africa. Their study supported the generalised Pareto distribution as being suitable for modelling the data; and the quantile estimates suggested that the province should expect an increase in extreme temperature. This implies that there is a growing concern on extreme temperature which poses a threat to agricultural vegetation, health and power outages.

According to the above literature, some studies revealed that South Africa has experienced a huge shift in climate change. This shift has changed the characteristics of precipitation and temperature. However, one can say that the above researches pointed out that the country will experience an increase in precipitation and temperature, especially in magnitude and frequency of the occurrences.

2.5 Extreme value theory: an overview

Extreme value theory (EVT) is a branch of statistics interested in the tail behaviour of a probability distribution. In the early days of development, EVT was designed to study the flood levels. However, recent studies showed how this

theory can also be applicable to many disciplines such as insurance, finance, meteorology phenomena and environmental sciences (Kratz, 2017; Adesina et al., 2016; Henry and Hsieh, 2014). For example, the following can be classified as rare events: financial crises arising from counterparties; large claims in insurance; high wind speed in meteorology; and the high concentration of ozone in environmental studies. Fundamentally, the main aim of EVT is to know or predict the occurrence of extreme or rare events using historical data (Charras-Garrido and Lezaud, 2013). There are two fundamental approaches in EVT, that is: the peaks-over-threshold and the block maxima. We explain in detail what is meant by the two approaches.

2.5.1 The block maxima approach

The block maxima is an extreme value approach that uses generalised extreme value distribution (GEVD) and GEVD for r-largest order statistics when analysing the data. The model development for these two distributions is based on the statistical behaviour of the maximum, $\mathbf{M}_n$,

$$\mathbf{M}_n = \max(X_1, \cdots, X_n) \tag{2.1}$$

where $X_1, \cdots, X_n$ is a sequence of independent and identically distributed (iid) random variables with distribution function F. Then, theoretically, the distribution of $\mathbf{M}_n$ can be derived as (Coles, 2001):

$$\begin{aligned} \Pr\{\mathbf{M}_n \leq z\} &= \Pr\{X_1 \leq z, \cdots, X_n \leq z\} \\ &= \Pr\{X_1 \leq z\} \times \cdots \times \Pr\{X_n \leq z\} \\ &= \{F(n)\}^n, \quad \text{for all } n. \end{aligned} \tag{2.2}$$

Since the distribution function F is unknown, then (2.2) is not useful in deter-

mining the distribution of $\mathbf{M}_n$. Another alternative approach is to approximate families of models for F^n using only the extreme data. The fact that the limiting distribution of $\mathbf{M}_n$ degenerates, implies that the behaviour of F^n as n approaches infinity is not sufficient (Smith, 2005; Coles, 2001). Then, to avoid this problem, we consider the linear renormalisation of $\mathbf{M}_n$:

$$\mathbf{M}_n^* = \frac{\mathbf{M}_n - b_n}{a_n} \tag{2.3}$$

for sequences of constants $\{a_n > 0\}$ and $\{b_n\}$. Then (2.3) gives rise to extremal types theorem, which deals with the limit distribution of $\mathbf{M}_n$ (Coles, 2001).

Theorem 2.1 (Extremal Types Theorem)**.** *(Coles, 2001)*
If there exist sequences of constants $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\boldsymbol{Pr}\{(\boldsymbol{M}_n - b_n)/a_n \leq z\} \longrightarrow G(z) \textbf{ as } n \longrightarrow \infty$$

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

$$\boldsymbol{I}: G(z) = \exp\{-\exp\big[-\big(\tfrac{z-b}{a}\big)\big]\}, \quad -\infty < z < \infty;$$

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

$$\boldsymbol{III}: G(z) = \begin{cases} \exp\{-[-(\tfrac{z-b}{a})^{\alpha}]\}, & z < b; \\ 1, & z \geq b; \end{cases}$$

for parameters $a > 0$ *and* b *for the case of families II and III such that* $\alpha > 0$.

These three families of distributions are called the Gumbel, Frechet and Weibull families, respectively; and they are also called the extreme value distributions

(Coles, 2001). Theorem 3.1, in simple terms, implies that if M_n can be normalised for suitable sequences $\{a_n > 0\}$ and $\{b_n\}$, then M_n^* will have a limiting distribution from the three types of extreme value distribution called the generalised extreme value distribution (GEVD) (see Chapter 3). In addition, the above results can be extended to other extreme order statistics, that is:

$$\mathrm{M}_n^{(k)} = \mathrm{k}^{th} \text{ largest of } (X_1, \cdots, X_n) \tag{2.4}$$

for fixed values of k. The limiting distribution of (2.4) as $n \to \infty$ is called the generalised extreme value distribution for r-largest order statistic (see Chapter 3).

According to Ferreira and de Haan (2015) there are three reasons for using the block maxima approach, namely:

- The block maxima may be preferable when the observations are not iid.
- The only available in one of few blocks information may be block maxima.
- The block maxima may be easier to apply since the block periods appear naturally in many situations.

The use of block maxima for both GEVD and GEVD for r-largest order statistics are sometimes criticised for wasting data if more data on extremes are available (Nemukula and Sigauke, 2018; Smith, 2005). For this reason, we next introduce the peaks-over threshold approach.

2.5.2 The peaks-over threshold approach

The peaks-over threshold approach is concerned with those observations that exceed a specified high threshold (Smith, 2005). The model development of this

approach is based on the following:

Let $X_1, X_2, \cdots$ be a sequence of iid random variables with common distribution function F. Then for X_i exceeding a high threshold u, X_i can be considered as an extreme event. Suppose X is an arbitrary term of the sequence, then the conditional probability is given by:

$$\mathbf{Pr}\{X > u + y \mid X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0,$$

describes the stochastic behaviour of extreme events (Coles, 2001). The distribution of exceedances is obtained using the results in Theorem 2.1.

Chapter 3

Methodology

Introduction

In this chapter we give the statistical approach used in this study. We present the analysis of extreme value theory. Furthermore, we define what is meant by stationarity and describe the tests of stationarity. Lastly, the Chapter discusses the goodness-of-fit techniques to be applied in this study.

3.1 Extreme value theory

In Chapter 2, we demonstrated that when modelling using EVT, the model development is based on the statistical behaviour of

$$\mathbf{M}_n = \max\{X_1, \cdots, X_n\}$$

where $X_1, \cdots, X_n$ is a sequence of independent and identically distributed (iid) random variables with the distribution function F. We further introduced the

three types of extreme value distributions and their properties. The next section is constructed based on the last section of Chapter 2.

3.1.1 The generalised extreme value distribution

Theorem 2.1 presented three types of limiting distributions with distinct forms of behaviour. These three types match the different forms of tail behaviour for the distribution function F. As a result, the three models can be unified into one family of models called generalised extreme value distribution (GEVD):

$$G(z) = \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}\right\}, \tag{3.1}$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, such that $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. The equation (3.1) is called the generalised extreme value distribution. The equation has three parameters, namely: location (μ), scale (σ) and shape (ξ). The shape parameter plays an important role in distinguishing the three classes of extreme value distributions (Smith, 2005). When $\xi > 0$ and $\xi < 0$, the equation leads to the types II and III in Theorem 2.1. For the case $\xi = 0$ or $\xi \longrightarrow 0$, the equation leads to the Gumbel-type distribution, that is:

$$G(z) = \exp\left[-\exp\left\{ -\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \quad -\infty < z < \infty \tag{3.2}$$

Thus, Theorem 2.1 can be written in the following form for large values of n:

$$\mathbf{Pr}\{(\mathbf{M}_n - b_n)/a_n \leq z\} \approx G(z).$$

Equivalently,

$$\begin{aligned} \Pr\{\mathbf{M}_n\} &\approx G\{(z-b_n)/a_n\} \\ &= G^*(z), \end{aligned} \tag{3.3}$$

where G^* is a member of the GEVD family. Since a member of the GEVD family managed to approximate a distribution of $\mathbf{M}_n^*$ for large n, then the GEVD family can be fitted directly to a series of observations of $\mathbf{M}_n$.
Therefore, the procedure to model the extreme events of independent observations, $X_1, X_2, \cdots$ for the GEVD is as follows: the data must be blocked into m sequences of length n, where n represents the number of years or periods. Taking maxima of each block (or year) generates a series, $\mathbf{M}_{n1}, \cdots, \mathbf{M}_{nm}$ to fit GEVD. An important aspect about this procedure is that the choice of block size m is crucial. A small value of m can result in poor approximation which can lead to bias. A large value of m can also result in large estimation of variances. As a result, when using the block maxima approach, there is a need to find a balance between the bias and the sizes of variances (Smith, 2005).

Another consideration is the estimation of extreme quantiles of the annual maxima. The quantile estimations play a vital role when modelling extreme events (Smith, 2005). They provide useful information on the behaviour of extreme observations in the successive years (Smith, 2005; Coles, 2001). According to Smith (2005), the quantile estimations are of particular interest, especially in environmental extremes since they also give an estimate of the level the process is expected to exceed once, on average, in a given number of years. The mathematical representation of the quantile function is obtained by inverting the generalised extreme value distribution as follows:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \left\{ -\log(1-p)\right\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma\log\left[-\log(1-p)\right], & \xi = 0 \end{cases} \tag{3.4}$$

where $G(z_p) = 1 - p$ and the quantity z_p is called the return level associated with the return period $\frac{1}{p}$. The quantity z_p is also defined as the level which is expected to be exceeded on average, once every $\frac{1}{p}$ years.

The use of GEVD arising from the block maxima approach is sometimes criticised for wasting data if more data on extremes are available (Nemukula and Sigauke, 2018; Smith, 2005). The GEVD for r-largest order statistics method was developed to overcome this problem. Using the results in (3.3) for fixed values of r, gives:

$$\mathbf{M}_n^k = \text{ k}^{th} \text{ largest of } (X_1, \cdots, X_2)$$

and the limiting behaviour of $\mathbf{M}_n^k$, for fixed k, as $n \longrightarrow \infty$, is given by:

$$\mathbf{Pr}\left[\frac{\mathbf{M}_n^k - b_n}{a_n} \leq z\right] \longrightarrow G_k(k)$$

such that $z : 1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0$ where

$$G_k(z) = \exp\left(-\tau(z)\right) \sum_{k-1}^{s=0} \frac{\tau(z)^s}{s!}$$

with

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

Therefore, the joint probability density function of GEVD for r-largest order statistics,

$$\mathbf{M}_n = \left(\frac{\mathbf{M}_n^{(1)} - b_n}{a_n}, \cdots, \frac{\mathbf{M}_n^{r} - b_n}{a_n} \right),$$

is given by:

$$f(z^{(1)}, \cdots, z^{(r)}) = \exp\left(-\left[1 + \xi\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \right) \times \prod_{k=1}^{r} \sigma^{-1}\left[1 + \xi\left(\frac{z^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}, \tag{3.5}$$

where $-\infty < \mu < \infty, \quad \sigma > 0$ and $-\infty < \xi < \infty, \quad z^{(r)} \leq z^{(r-1)} \leq \cdots \leq z^{(1)}$ and $z^{(k)} : 1 + \xi\left(\frac{z^{k}-\mu}{\sigma}\right) > 0$, for $k = 1, \cdots, r$. Equation (3.5) reduces to the density of the Gumbel family when $r = 1$.

Therefore, the procedure to model r-largest order statistics uses the idea of block maxima approach. The series of iid variables data are blocked into m blocks. Recording the largest r_i observations in the block i, leads to the series: $\mathbf{M}_i^{(r_i)} = \left(z_i^{(1)}, \cdots, z_i^{(r_i)} \right)$, for $i = 1, \cdots, m$. Guedes and Scotto (2004) presented evidence on how the choice of r can lead to high variance and bias if it is not carefully handled.

3.1.2 Peaks-Over Threshold model

Peaks-Over Threshold (POT) approach considers those of the initial observations that exceed a predetermined threshold regardless of the block (Ferreira and de Haan, 2015). The development of r-largest order statistics method was designed to overcome the limitations of the block maxima approach. However, recent studies have noted that the r-largest order statistics approach also has some limitations, for example if one block happens to contain more extreme observations than another (Soares and Scotto, 2015). The POT approach is more preferred in this situation than the two former approaches.

3.1.3 The generalised Pareto distribution

Suppose X is an arbitrary term of the $X_1, X_2, \cdots$ and let F satisfy Theorem 2.1. Then for large n, we have:

$$\mathbf{Pr}\{\mathbf{M}_n \leq z\} \approx G(z),$$

where

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

for some parameters μ and $\sigma >$ belongs to real numbers and ξ. Then for large u, the distribution function of $(X-u)$, conditional on $X > u$, is approximately:

$$\mathbf{H}(y) = 1 - \left(1 + \frac{\xi y}{\bar{\sigma}}\right)^{-\frac{1}{\xi}}, \tag{3.6}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\bar{\sigma}) > 0\}$, where $\bar{\sigma} = \sigma + \xi(u - \mu)$. Equation (3.6) reduces to an exponential distribution with parameter $\frac{1}{\bar{\sigma}}$ if $\xi = 0$, that is:

$$\mathbf{H}(y) = 1 - \exp(-\frac{y}{\bar{\sigma}}), \quad y > 0. \tag{3.7}$$

Thus, (3.6) is called the generalised Pareto distribution (GPD).

The return levels of the GPD can be determined as follows: let GPD be an appropriate model for exceedances over a threshold, u, with parameters σ and ξ. Then for $\xi = 0$ and $x > u$, we have:

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

It follows that,

$$\mathbf{Pr}\{X > x\} = \zeta_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}},$$

where $\zeta_u = \mathbf{Pr}\{X > u\}$ and x_m is called the level that is exceeded on average, once every m observations. The mathematical representation of x_m is given by the solution to:

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

After rearranging, we have

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right], \tag{3.8}$$

provided that m is sufficiently large to ensure that $x_m > u$. When $\xi = 0$, we have:

$$x_m = u + \sigma \mathbf{log}(m\zeta_u),$$

for sufficiently large m. Furthermore, suppose that there are n_y observations per year, then the N-year return level will be given by:

$$z_N = \begin{cases} u + \frac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right], & \xi \neq 0, \\ u + \sigma \mathbf{log}(Nn_y\zeta_u), & \xi = 0. \end{cases}$$

3.1.4 The Choice of Threshold

There are many techniques used to select appropriate threshold for a data set before the modelling of GPD can commence (Coles, 2001). This study presents commonly used techniques of threshold selection, namely: the mean residual life plot, the dispersion index plot and the parameter stability plot (Nemukula and Sigauke, 2018; Maposa, 2016). Next, we give details of these three threshold choices.

The mean residual plot

The mean residual plot is based on the mean of the GPD (Beirlant et al., 2004; Coles, 2001). Let Y be a GPD with parameters σ and ξ. Then:

$$\mathbf{E}(Y) = \frac{\sigma}{1-\xi}, \tag{3.9}$$

provided $\xi < 1$. For the case $\xi \geq 1$, we then have $\mathbf{E}(Y) = \infty$. Suppose that Y is valid as a model for the excess of a threshold u_0 generated by a series $X_1, \cdots, X_n$, of which an arbitrary term is denoted by X. Then, from (3.9), we have:

$$\mathbf{E}(X - u_0 \mid X > u_0) = \frac{\sigma_{u_0}}{1-\xi},$$

provided $\xi < 1$, where σ_{u_0} denotes the scale parameter corresponding to u_0 and also hold for all thresholds. That is, if $u > u_0$, then we have:

$$\begin{aligned} \mathbf{E}(X - u \mid X > u) = & \frac{\sigma_u}{1-\xi} \\ = & \frac{\sigma_{u_0} + \xi u}{1-\xi}. \end{aligned} \tag{3.10}$$

Equation (3.10) shows that the $\mathbf{E}(X - u \mid X > u)$ is linear for $u > u_0$. Therefore, $\mathbf{E}(X - u \mid X > u)$ is called the mean of the excesses of the threshold u. Using the results in (2.2), we have:

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\}, \tag{3.11}$$

where $x_{(1)}, \cdots, x_m$ consist of the n_u observations such that $u > u_0$. Which X_i is called the mean residual life plot. This procedure will be used to determine appropriate threshold u (Nkrumah, 2017).

The dispersion index plot

The dispersion index plot is another technique used to determine an appropriate threshold u. This plot assumes that the data is generated by a Poisson process. According to Coles (2001), all the observations above the specified high threshold must be Poisson distributed. Thus, suppose that X has a Poisson distribution with parameter λ, then

$$\mathrm{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}$$

and $\mathrm{E}(X) = \mathrm{Var}(X)$. Therefore, the dispersion index (DI) is given by:

$$\mathrm{DI} = \frac{\sigma^2}{\mu},$$

where σ^2 is the intensity of Poisson process and λ is the mean number of events in a year or block. The appropriate threshold is selected after testing if the ratio DI differs from 1. That is, if DI is close to 1, the corresponding threshold is not reject (Khuluse, 2010).

The parameter stability plot

The idea of parameter stability plot is that exceedances of specified threshold u_0 follow a GPD with parameters ξ and σ_{u_0} (Nkrumah, 2017). Hence the exceedances of the threshold u_0 such that $u > u_0$ will also follow a GPD with $\xi_u = \xi$ and $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$ being the shape and scale parameters, respectively. Suppose

$$\sigma^* = \sigma_u - \xi_u u. \tag{3.12}$$

Then, if u_0 is a suitable high threshold, (3.12) does not depend on u. The parameter stability plot is then defined by the following locus points:

$$\{(u, \sigma^*); u < x_{\max}\} \quad \text{and} \quad \{(u, \xi_u); u < x_{\max}\}$$

where $x_{\max}$ is the maximum of the observations, σ^* and ξ_u are constants for all $u > u_0$. Then, the correct threshold will be chosen at the value where the shape and scale parameters remain constant (Maposa, 2016; Coles, 2001).

3.2 Maximum likelihood estimation

There are several techniques for parameter estimation when modelling extreme observations (Nkrumah, 2017; Maposa, 2016; Smith, 2005). According to Coles (2001), the method of maximum likelihood estimation (MLE) is considered to be the best when dealing with large samples, but performs badly when the sample size is small. Before we present the MLEs of GEVD and GPD, we first need to define the following terminologies: the likelihood function, maximum likelihood estimator, delta method and profile likelihood.

Definition 3.1. *(Likelihood function) (Hogg et al., 2015).*
Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from a distribution that depends on one or more unknown parameters $\alpha_1, \alpha_2, \cdots, \alpha_m$, with pdf that is denoted by $g(y; \alpha_1, \alpha_2, \cdots, \alpha_m)$. Suppose that $(\alpha_1, \alpha_2, \cdots, \alpha_m)$ is restricted to a given parameter space Ω. Then the joint pdf of $Y_1, Y_2, \cdots, Y_n$ is given by:

$$\boldsymbol{L}(\alpha_1, \alpha_2, \cdots, \alpha_m) = g(y_1; \alpha_1, \cdots, \alpha_m) g(y_2; \alpha_1, \cdots, \alpha_m) \cdots g(y_n; \alpha_1, \cdots, \alpha_m)$$

such that $(\alpha_1, \alpha_2, \cdots, \alpha_m) \in \Omega$. Then, the function $\boldsymbol{L}$ is called the likelihood function.

Definition 3.2. *(Maximum likelihood estimator) (Hogg et al., 2015).*
Suppose $u_1(y_1, \cdots, y_n), u_2(y_1, \cdots, y_n), \cdots, u_m(y_1, \cdots, y_n)$ are m-tuple in Ω that maximises $\boldsymbol{L}(\alpha_1, \alpha_2, \cdots, \alpha_m)$. Then

$$\hat{\alpha}_1 = u_1(Y_1, \cdots, Y_n); \;\; \hat{\alpha}_2 = u_2(Y_1, \cdots, Y_n); \;\; \cdots; \;\; \hat{\alpha_m} = u_m(Y_1, \cdots, Y_n)$$

are called the maximum likelihood estimators of $\alpha_1, \alpha_2, \cdots, \alpha_m$*, respectively, and the corresponding observed values of these statistics, that is:*

$$u_1(Y_1, \cdots, Y_n), u_2(Y_1, \cdots, Y_n), \cdots, u_m(Y_1, \cdots, Y_n$$

are called maximum likelihood estimates.

Theorem 3.3. *(Delta method) (Coles, 2001).*
Let $\hat{\alpha}_0$ *be the large-sample maximum likelihood estimator of the d-dimensional parameter* α_0 *with approximate variance matrix* V_α*. Then if* $\phi = g(\alpha)$ *is a scalar function, the maximum likelihood estimator of* $\phi_0 = g(\alpha_0)$ *satisfies:*

$$\hat{\phi}_0 \sim \boldsymbol{N}(\phi_0, V_\phi),$$

where

$$V_\phi = \Delta\phi^\top V_\alpha \Delta\phi,$$

with

$$\Delta_\phi = \left[\frac{\partial\phi}{\partial\alpha_1}, \frac{\partial\phi}{\partial\alpha_2}, \cdots, \frac{\partial\phi}{\alpha_d}\right]$$

evaluated at $\hat{\alpha}_0$*.*

Definition 3.4. *(Profile likelihood function) (Coles, 2001).*
According to Coles (2001) the profile log-likelihood function is a more accurate method which is based on profile likelihood. For example, the profile log-likelihood for θ_i *will be defined as:*

$$\ell_p(\theta_i) = \boldsymbol{max}_{\theta_{-i}} \ell(\theta_i, \theta_{-1})$$

where $\ell_p(\theta_i)$ *is called the profile of the log-likelihood surface viewed from the* θ_i *axis.*

From the above definitions and theorems, we can now present the log-likelihood functions for GEVD, GPD and their corresponding quantile functions.

3.2.1 The log-likelihood function of the GEVD

Assume that $z_1, z_2, \cdots, z_m$ are independent variables having the GEVD. Then, the log-likelihood for the GEVD when $\xi \neq 0$ is given by:

$$\ell(\mu, \sigma, \xi) = -m\log\sigma - (1 + \frac{1}{\xi}) \sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{3.13}$$

such that $1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0, \quad$ for $i = 1, \cdots, m$. Then for the case $\xi = 0$, the log-likelihood in (3.13) changes to:

$$\ell(\mu, \sigma) = -m\log\sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\} \tag{3.14}$$

Since there is no analytical solution for both (3.13) and (3.14), then there is a need for numerical solutions to obtain maximum likelihood estimates.

3.2.2 The likelihood function of GEVD for r-largest order statistics

The likelihood function of GEVD for r-largest order statistics is given by:

$$\mathrm{L}(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp\left\{-\left[1 + \xi\left(\frac{z_i^{r_i} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \times \prod_{k=1}^{r_i} \sigma^{-1}\left[1 + \xi\left(\frac{z_i^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right), \tag{3.15}$$

such that $1+\xi\left(\frac{z^{(k)}-\mu}{\sigma}\right) > 0, \;\; k = 1, \cdots, r_i, \;\; i = 1, \cdots, m$; otherwise the likelihood is zero. Then for the case $\xi = 0$, we have:

$$\mathrm{L}(\mu,\sigma,\xi)=\prod_{i=1}^{m}\left(\exp\left\{-\exp\left[-\left(\frac{z^{(r_i)}-\mu}{\sigma}\right)\right]\right\}\times\prod_{k=1}^{r_i}\sigma^{-1}\exp\left[-\left(\frac{z_i^{(k)}-\mu}{\sigma}\right)\right]\right). \tag{3.16}$$

In the case $r=1$, the likelihood in (3.16) reduces to the GEVD. Since there is no analytical solution for (3.15) and (3.16), the numerical solution will be applied to obtain maximum likelihood estimators.

3.2.3 The log-likelihood function for the GPD

Let $x_1, x_2, \cdots, x_k$ represent k excesses of a threshold u. Then, for $\xi \neq 0$ the log-likelihood is given by:

$$\ell(\sigma,\xi)=-k\log\sigma-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\log\left(1+\frac{\xi x_i}{\sigma}\right) \tag{3.17}$$

such that $\left(1+\frac{\xi x_i}{\sigma}\right)>0,\quad$ for $i=1,\cdots,k$; otherwise the log-likelihood is negative infinity, that is $\ell(\sigma,\mu)=-\infty$. Then for the case $\xi=0$, the log-likelihood in (3.17) reduces to:

$$\ell(\sigma)=-k\log\sigma-\frac{1}{\sigma}\sum_{i=1}^{k}x_i. \tag{3.18}$$

An analytical solution is not possible for both (3.17) and (3.18). The use of algorithm will be adopted in order to obtain maximum likelihood estimates. Furthermore, R programming will be used to obtain the maximum likelihood estimates of unknown parameters.

3.2.4 Inference on return levels of GEVD

In order to derive the maximum likelihood estimate of z_p, we need to use MLEs of GEVD parameters in (3.12). Then, for $0<p<1$, the $\frac{1}{p}$ return level is given by:

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right], & \text{for } \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log y_p, & \text{for } \hat{\xi} = 0 \end{cases}$$

such that $y_p = -\log(1-p)$. Then by Theorem 3.3 the variance of $\hat{z}_p$ is given by:

$$\text{Var}(\hat{z}_p) \approx \Delta z_p^T V \Delta z_p \ ,$$

where V is the variance-covariance matrix of the estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and

$$\Delta z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1, -\xi^{-1}(1 - y_p^{-\xi}), \sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1} y_p^{-\xi}\log y_p\right]$$

evaluated at $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$.

3.2.5 Inference on return levels of GPD

Suppose that $m = N \times n_y$, then the N-year return levels is given by:

$$z_N = \begin{cases} u + \frac{\sigma}{\xi}\left[(N n_y \zeta_u)^{\xi} - 1\right], & \xi \neq 0 \\ u + \sigma\log(N n_y \zeta_u), & \xi = 0. \end{cases}$$

Since the estimation of return levels requires unknown parameters to be estimated, we then start by estimating the probability of an individual observation exceeding the threshold, $\hat{\zeta}_u = \frac{k}{n}$. Now, in order to determine the estimation of return levels, estimation of unknown parameters is required. Since $\hat{\zeta}_u$ follows a binomial distribution and it is estimated by $\frac{k}{n}$, then by standard properties of the binomial distribution the variance of ζ_u, is given by $\text{Var}(\hat{\zeta}_u) \approx \hat{\zeta}_u(1 - \hat{\zeta}_u)/n$. This implies that the variance-covariance matrix for $\hat{\zeta}_u, \hat{\sigma}, \hat{\xi}$ is approximated by:

$$V = \begin{bmatrix} \hat{\zeta}_u(1-\hat{\zeta}_u)/n & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix}$$

where $v_{i,j}$ denotes the (i,j) term of variance-covariance matrix of $\hat{\sigma}$ and $\hat{\xi}$. By Theorem 3.3:

$$\mathbf{Var}(\hat{x}_m) \approx \Delta x_m^T V \Delta x_m$$

where,

$$\begin{aligned} \Delta x_m^T &= \left[\frac{\partial x_m}{\partial \zeta_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi} \right] \\ &= \left[\sigma m^{\xi} \zeta^{\xi-1}, \xi^{-1}\{(m\zeta_u)^{\xi} - 1\}, -\sigma\xi^{2}\{(m\zeta_u)^{\xi} - 1\} + \sigma\xi^{-1}(m\zeta_u)^{\xi}\mathbf{log}(m\zeta_u) \right] \end{aligned}$$

which is evaluated at $\left(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi}\right)$.

3.3 Candidate distributions

This section presents the investigation of goodness-of-fit of candidate distributions, namely; 2-parameter Weibull, 3-parameter Weibull, Gumbel, Gamma, 2-parameter log-normal, 3-parameter log-normal, 2-parameter Pareto and 3-parameter Pareto. In this section we are interested in looking at how the candidate distributions fit the tails compared with the EVT approach. Next, we define these distributions.

Weibull distribution

We have two types of Weibull distributions: the cumulative distribution function (CDF) of the two-parameter Weibull distribution is given by:

$$F(x) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right),$$

while the CDF of the three-parameter Weibull distribution is given by:

$$F(x) = 1 - \exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right),$$

where γ is a continuous location parameter, β is the continuous scale parameter and α is a continuous shape parameter. This distribution is commonly used in hydrology and reliability studies (Alam et al., 2018; Maposa, 2016).

Gumbel distribution

The Gumbel distribution is an extreme value type I which is commonly used in flood frequency analysis and engineering (Alam et al., 2018; Maposa, 2016). The CDF of the Gumbel distribution is given by:

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right),$$

where $\mu, \sigma > 0$ are called the continuous location and scale parameters, respectively.

Gamma distribution

The CDF of the gamma distribution is defined as:

$$F(x) = \frac{\Gamma_{(x-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)},$$

where $\gamma, \beta > 0$ and α are the continuous location, scale and shape parameters respectively. Γ is called the gamma function (Alam et al., 2018; Maposa, 2016; Beirlant, 2004).

Pareto distribution

There are two types of Pareto distributions, the two and three parameter distribution. According to Arnold (2003) Pareto distribution was designed to present the distribution of income. Suppose X is a random variable that follows a two-parameter Pareto distribution, then the CDF of X is

$$F(x) = \left(\frac{x}{\sigma}\right)^{-\xi}, \quad x > \sigma,$$

where σ and ξ are called the scale and shape parameter respectively. The CDF of the three-parameter Pareto distribution is given by:

$$F(x) = \left[1 + \left(\frac{x-\mu}{\sigma}\right)\right]^{-\xi}, \quad x > \mu,$$

where σ, μ and ξ are called scale, location and shape parameter respectively.

Log-normal distribution

There are two log-normal distributions (Alam et al., 2018; Maposa, 2016; Beirlant, 2004): the two-parameter, whose CDF is given by:

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\mu}\right),$$

and the three-parameter Log-normal defined as:

$$F(x) = \Phi\left(\frac{\ln(x-\gamma) - \mu}{\sigma}\right),$$

where $\gamma, \sigma > 0$ and μ are the continuous location, scale and shape parameters, respectively. Φ is called the Laplace integral.

The parameters of all candidate distributions discussed in this section will be estimated using the method of maximum likelihood.

3.4 Stationarity

There are two main purposes for modelling stationary time series. Firstly, it maintains model stability and secondly, it provides a framework in which averaging can be properly used to describe the time series (Arltova and Fedorova, 2016). However, many researchers define stationarity as a statistical structure of series which is independent of time (Khuluse, 2010; Nason, 2006; Smith, 2005). Before we can define the concept of stationarity, we shall begin with the simple building blocks and then proceed to complex structures.

Definition 3.5. *(Purely random process) (Nason, 2006).*
A purely random process is a stochastic process, $\{\xi_t\}_{t=\infty}^{\infty}$*, where each element* ξ_t *is statistically independent of every other element,* ξ_s *for* $s \neq t$*, and each element has an identical distribution.*

Next, we define the concept of stationarity using Definition 3.5.

Definition 3.6. *(Stationary process) (Coles, 2001).*
A random process $x_1, x_2, \cdots$ *is said to be stationary if given any set of integers* $\{j_i, \cdots, j_k\}$ *and for any integer* m*, the joint distribution of* $(X_{j_i}, \cdots, X_{j_k})$ *and* $(X_{j_i+m}, \cdots, X_{j_k+m})$ *are identical.*

Definition 3.5 and 3.6 enable us to provide details on the techniques used to test for stationarity tests.

3.4.1 Test for unit root

According to Arltova and Fedorova (2016) the most essential task in modelling is to keep the order of analysed time series fixed through the use of unit root tests. There are many techniques used to check whether or not a series contains a unit root (Oliver and Mung'atu, 2018; Hasan et al., 2012). The study focuses on the augmented Dickey-Fuller (ADF) test for stationarity.

Dickey-Fuller and Augmented Dickey-Fuller Tests

We start by introducing the concept of Dickey-Fuller (DF) test and then use this concept to build the augmented Dickey-Fuller (ADF) test.

The DF test is most widely used to test whether a certain series has a unit root. The test is based on the model of the first-order autoregressive process (Arltova and Fedorova, 2016). Thus, we have:

$$y_t = \beta_1 y_{t-1} + \epsilon_t \ , \ \ t = 1, \cdots, T, \tag{3.19}$$

where β_1 is the autoregression parameter and ϵ_t is a white noise process. The null hypothesis to be tested is given by: $\mathrm{H}_0 : \beta_1 = 1$ (the process contains a unit root and hence it is non-stationary) and the alternative hypothesis is: $\mathrm{H}_1 : |\beta_1| < 1$ (the process does not contain a unit root and is stationary). Using the following equation:

$$\Delta y_t = \alpha y_{t-1} + \epsilon_t, \quad \text{where} \quad \alpha = \beta_1 - 1,$$

The DF test statistic is given by:

$$t_{\mathrm{DF}} = \frac{\hat{\beta}_1 - 1}{s_{\hat{\beta}_1}}, \tag{3.20}$$

where $\hat{\beta}_1$ is the estimate of β and $s_{\hat{\beta}_1}$ is the standard error of the estimator, $\hat{\beta}$. This follows the DF distribution, and the critical values can be obtained from the Dickey and Fuller table.

For ADF test, we extend (3.20) by a constant or a linear trend, that is:

$$y_t = \alpha_0 + \beta_1 y_{t-1} + \epsilon_t,$$

$$y_t = \alpha_0 + \alpha_1 t + \beta_1 y_{t-1} + \epsilon_t. \tag{3.21}$$

The ADF test is constructed by transforming the following equation:

$$y_t = \beta_1 y_{t-1} + \sum_{i=1}^{\rho-1} \gamma_i \Delta y_{t-1} + \epsilon_t.$$

The test statistic of the ADF test is derived from:

$$\Delta y_t = (\beta_1 - 1) y_{t-1} + \sum_{i=1}^{\rho-1} \gamma_i \Delta y_{t-i} + \epsilon_t.$$

An important part when using the ADF test is the choice of lags ρ. The number of lags should be carefully chosen. A very small value of ρ will affect the autocorrelation of the test and a very large ρ will substantially reduce the power of the test (Arltova and Fedorova, 2016). To avoid this problem, we expand (3.21) with a linear trend, and the new ADF test is based on the model:

$$y_t = d_t + \beta_1 y_{t-1} + \sum_{i=1}^{\rho-1} \gamma_i \Delta y_{t-i} + \epsilon_t, \tag{3.22}$$

where $d_t = \sum_{i=1}^{\rho} \phi_i t^i, \quad$ for $\rho = 0, 1$, contains the analytical parts of the models mentioned earlier in the section. When $T \to \infty$, the limit distribution of the ADF test statistic is identical to the distribution of the DF test.

3.5 Goodness of fit tests

The goodness-of-fit tests are widely used to assess how a given data follow a specified distribution. Suppose that $x_1, x_2, \cdots, x_n$ is a sample of n average monthly rainfall observed. Let F be the cumulative distribution function (CDF) of the random variable X. Next, we present the following two tests: the Anderson Darling (A-D) test and the Kolmogorov-Smirnov (K-S) test.

3.5.1 Anderson Darling test

According to Maposa (2016) the mechanism behind this test is that it compares the fitted observed CDF to a theoretical CDF. Furthermore, the test statistic of A-D test is given by:

$$\mathrm{A}^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(x) + \ln(1-F(x))], \tag{3.23}$$

where $F(x)$ is the theoretical CDF and $F_n(x)$ represents the empirical CDF. The null hypothesis to be tested says that the data follow a specified distribution, while the alternative hypothesis states that the data does not follow the specified distribution. The rejection rule states that we reject the null hypothesis at $\alpha\%$ level of significance when A-D test is greater than the tabulated value, or we reject the null hypothesis if the p-value is less than the specified level of significance.

3.5.2 Kolmogorov-Smirnov test

In the case of K-S, a comparison is made between the largest vertical distance $\mathrm{D}_{\max}$, of the empirical CDF $F_n(x)$ and the theoretical CDF $F(x)$ (Maposa, 2016). The test statistic of the K-S test is given by:

$$\mathrm{D}_{\max} = \mathbf{Max}_x \mid F_n(x) - F(x) \mid . \tag{3.24}$$

The null hypothesis to be tested is given by $\mathrm{H}_0 : F(x) = F(x;\beta)$, while the alternative hypothesis is: $\mathrm{H}_1 : F(x) \neq F_0(x;\beta)$, where F_0 is a specified distribution and β is a vector of unknown parameters. The rejection rule states that we reject the null hypothesis at $\alpha\%$ level of significance if $\mathrm{D}_{\max}$ is greater than the tabulated value of D_α, or we reject the null hypothesis if the p-value is less than the specified level of significance.

According to Maposa (2016) the K-S test is more sensitive to the centre of the distribution. In a separate study, Nemukula and Sigauke (2018) presented some evidence about the A-D test in which they stated that the A-D test is more sensitive to the tail of the distribution. Using the two findings, and given that the modelling of extreme observations is interested in the tails of the distribution, this implies that the A-D test will be appropriate for testing the tails of the distribution while the K-S will be appropriate for testing the centre of the distribution.

3.5.3 Deviance statistic

The deviance statisitc is a statistical procedure used to assess the goodness-of-fit of models. The idea behind this approach is that, it uses the maximum likelihood function of i and j to obtain deviation statistic to be compared to a chi-square of one degrees of freedom. The deviance statistic is given by:

$$\mathrm{D}_{(i,j)} = 2\left[\ln \lambda(r_i) - \ln \lambda(r_j)\right] \sim \chi^2_1, \text{ for } i, j = 2, 3, \cdots, 6 (i \neq j), \tag{3.25}$$

where $\lambda(r_i)$ and $\lambda(r_j)$ are the maximum likelihood functions of r_i and r_j respectively. The test validates the model based on r_i relative to r_j (Nemukula and Sigauke, 2018; Coles, 2001).

3.5.4 Diagnostic plots

Another procedure of assessing the goodness-of-fit is by using the diagnostic plots. There are many diagnostic plots used to assess the goodness-of-fit, but in this study we focus mainly on the probability-probability (P-P) plots and the quantile-quantile (Q-Q) plots, the definitions of which have been quoted from Maposa (2016).

Definition 3.7. *(P-P plot) (Maposa, 2016; Beirlant et al., 2004).*
A P-P plot is used to graphically assess the goodness-of-fit of a specified distribution. The P-P plot is plotted based on the empirical CDF values against the theoretical CDF values.

Definition 3.8. *(Q-Q plot) (Maposa, 2016; Beirlant et al., 2004).*
A Q-Q plot is used to visualise and assess the goodness-of-fit of a distribution graphically. The plot uses this information to form its structure, $(Q(i/(n+1)); x_{i,n})$, *for* $i = 1, 2, \cdots, n$ *and the structure must be linear if* $x_1, \cdots, x_n$ *are from a distribution with quantile function* Q.

The best model for both P-P and Q-Q plots will be chosen when the specified distribution fits the observed data. That is, the P-P and Q-Q plots must be approximately linear (Maposa, 2016).

Nemukula and Sigauke (2018) criticised the use of P-P plots in favour of Q-Q plots. It was argued that Q-Q plots are not affected by the symmetry of the distribution and also the shifts in location and scale parameters.

3.6 Point process approach

This section presents the point process models about the statistics of extremes. There are two fundamental Poisson processes, namely: the homogeneous and

non-homogeneous Poisson point processes. This study is interested in modelling extremes using non-homogeneous Poisson process in time. Next, we define the point process and Poisson point process.

3.6.1 Point process

Point processes are defined as the stochastic or random processes composed of time series of point events that occur in continuous time (Daley and Vere-Jones, 2003). For example, the point processes in time is the occurrence of tornado or heat wave at a certain location in time (Khuluse, 2010). According to Coles (2001) there are two purposes of using point processes. Firstly, point processes provide an interpretation of extreme value behaviour that unifies all the models of GPD and GEVD. Secondly, the point process models lead directly to a likelihood that is non-stationary in threshold excess than that obtained from the GPD. Next, we define the concept of Poisson point process.

3.6.2 The Poisson point process

For statistical purposes a point process need to be characterised. We present some definitions that will assist in characterising a point process.

Definition 3.9. *(Statistical properties of a point process) (Coles, 2001).*
Suppose that $N(A)$ is a set of non-negative integer-valued random variables for each $A \subset \mathcal{A}$, such that $N(A)$ is the number of points in the set A. Then, the probability distribution of each of the $N(A)$ determines the characteristics of the point process, that is N.

In summary, we can define the features of a point process as,

$$\Lambda(A) = E\{N(A)\},$$

which is the expected number of points in any subset $A \subset \mathcal{A}$. Thus, Λ is sometimes called the intensity measure of the process. In order to present the mathematical representation of a Poisson point process, we first need to define the intensity function of the process and the canonical point process.

Definition 3.10. *(The intensity function) (Coles, 2001).*
Suppose that $A = [a_1, x_1] \times [a_2, x_2] \times \cdots \times [a_k, x_k] \subset \mathbb{R}^k$ exists. Then, the intensity function of the process is given by the derivative function as follows:

$$\lambda(x) = \frac{\partial \Lambda(A)}{\partial x_1 \partial x_2 \cdots \partial x_k}.$$

Definition 3.11. *(The canonical point process) (Coles, 2001).*
The canonical point process is the one-dimensional homogeneous Poisson process with a parameter $\lambda > 0$ such that $\mathcal{A} \subset \mathbb{R}$ satisfies:

1. *for all $A = [t_1, t_2] \subset \mathcal{A}$,*

$$N(A) \sim \textit{Poi}(\lambda(t_2 - t_1))$$

2. *for all non-overlapping subset A and $\mathcal{B}$ of $\mathcal{A}$, $N(A)$ and $N(\mathcal{B})$ are independent random variables.*

Summarising the above definitions, it is clear that given an interval with a certain number of points, it is believed that the interval will follow the Poisson distribution having the mean proportional to the interval length and the occurrence of number of points in separate intervals are mutually independent.

Coles (2001) states that the Poisson process with parameter λ can be shown to be an appropriate stochastic model for points that occur randomly in time (at a uniform) of λ per unit time interval, with its intensity measure given by:

$$\Lambda([t_1, t_2]) = \lambda(t_2 - t_1),$$

and $\lambda(t) = \lambda$ is called the intensity density function.

Therefore, Definition 3.11 can be generalised to a model of points that occur randomly in time at a variable rate of $\lambda(t)$. Thus,

$$N(A) \sim \text{Poi}(\Lambda(A)) \tag{3.26}$$

is called the one-dimensional homogeneous Poisson process having the properties as in Definition 3.11, but with the modified property $A = [t_1, t_2] \subset \mathcal{A}$. The unknown parameters in (3.26) are given by the following;

$$\Lambda(A) = \int_{t_1}^{t_2} \lambda(t)dt.$$

with $\Lambda(\cdot)$ being the intensity measure and $\lambda(\cdot)$ being the density function. Therefore, a k-dimensional non-homogeneous Poisson process with intensity density function $\lambda(\cdot)$ such that $\mathcal{A} \subset \mathbb{R}^k$ is given by;

$$N(A) \sim \text{Poi}(\Lambda(A)),$$

where

$$\Lambda(A) = \int_A \lambda(x)dx,$$

provided that it satifies the property of independent counts on non-overlapping subsets and for all $A \subset \mathcal{A}$.

3.6.3 Maximum likelihood of the Poisson point process

The statistical approach of point process is similar to that of the POT. The estimation of the process requires a set of observed points, that is, $x_1, \cdots, x_n$ in a region or interval. Then from those observed points, we choose the appropriate class of points to estimate the process models. Since in this study our inter-

est is on non-homogeneous Poisson process, $\lambda(\cdot)$ is said to belong to the family of parametric models of $\lambda(\cdot;\theta)$. In line with that, we now know that the only problem which is related to model verification is the estimation of the vector θ. Therefore, we are now ready to present the likelihood function of the Poisson process.

Now using the likelihood approach, we let $I_i = [x_i, x_i + \delta_i]$, for $i = 1, 2, \cdots, n$ be small intervals based around the observations and $\mathcal{I} = \mathcal{A} \bigcup_{i=1}^{n} I$. Then using the properties of Poisson process,

$$\mathbf{Pr}\{N(\mathcal{I}_i) = 1\} = \exp\{-\Lambda(I_i;\theta)\}\Lambda(I_i;\theta), \tag{3.27}$$

where

$$\Lambda(I_i;\theta) = \int_{x_i}^{x_i+\delta_i} \lambda(u)du \approx \lambda(x_i)\delta_i. \tag{3.28}$$

Substituting (3.28) into (3.27), we have

$$\mathbf{Pr}\{N(I_i) = 1\} \approx \exp\{-\lambda(x_i)\delta_i\}\lambda(x_i)\delta_i \approx \lambda(x_i)\delta_i, \tag{3.29}$$

such that $\exp\{-\lambda(x_i)\delta_i\} \approx 1$ and also

$$\mathbf{Pr}\{N(I) = 0\} = \exp\{\Lambda(I)\} \approx \exp\{-\Lambda(\mathcal{A})\}, \tag{3.30}$$

for small δ_i. Therefore, the likelihood function of Poisson point process is given by:

$$\begin{aligned}\mathbf{L}(\theta; x_1, \cdots, x_n) &= \mathbf{Pr}\{N(\mathcal{I}) = 0, N(I_1) = 1, N(I_2) = 1, \cdots N(I_n) = 1\} \\ &= \mathbf{Pr}\{N(\mathcal{I}) = 0\} \prod_{i=1}^{n} \mathbf{Pr}\{N(I_i) = 1\} \\ &\approx \exp\{-\Lambda(\mathcal{A}; \theta)\} \prod_{i=1}^{n} \lambda(x_i; \theta)\delta_i. \end{aligned} \tag{3.31}$$

After dividing (3.31) by δ_i, we have

$$\mathbf{L}(\theta; x_i, \cdots, x_n) = \exp\{-\Lambda(\mathcal{A}; \theta)\} \prod_{i=1}^{n} \lambda(x_i; \theta), \tag{3.32}$$

where,

$$\Lambda(\mathcal{A}; \theta) = \int_{\mathcal{A}} \lambda(x; \theta) dx.$$

3.6.4 Applying the Poisson process to EVT

In the beginning of this section, it was stated that the point process framework is similar to that of POT. It was also stated that the inference made by the point process model could be similar to the one from the threshold exceedance approach.

Therefore, suppose that $X_1, X_2, \cdots$ are iid random variables with the same distribution function F as stated in section 3.1.1. Again, we suppose that the X_i are extreme values such that $\mathbf{M}_n = \max\{X_1, \cdots, X_n\}$.

Now, if there are sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

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

where

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

with z_- and z_+ being the lower and upper endpoints of G, respectively. Therefore, we have the sequence of point processes

$$N_n = (i/(n+1)), (X_i - b_n)/a_n : i = 1, \cdots, n)$$

that will converge on this region $(0,1) \times [u, \infty)$, for any $u > z_-$, to a Poisson process with intensity measure on $A = [t_1, t_2] \times [z, z_+]$ which is given by:

$$\Lambda(A) = (t_2 - t_1)\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \tag{3.33}$$

3.6.5 Connections between the Poisson process and POT models

Let X_i, for $i = 1, \cdots, n$, be iid random variables. Assume that the distribution of the exceedance follows the GPD. Let $\zeta = \mathbf{Pr}\{X_i > u\}$, so that by (3.7)

$$\zeta = \mathbf{Pr}\{X_i > u\} \approx \frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \tag{3.34}$$

where (μ, σ, ξ) are the parameters corresponding to that of the GEVD and $\bar{\sigma} = \sigma + \xi(u - \mu)$ (Coles, 2001).

Since in the peaks-over threshold approach we only focus on the model distribution of the observations above the threshold, then it is believed that the likelihood of the observations below u, is

$$\mathbf{Pr}\{X_i < u\} = 1 - \zeta. \tag{3.35}$$

This implies that the likelihood contribution for all the observations exceeding u, is

$$\mathbf{Pr}\{X_i = x\} = \mathbf{Pr}\{X_i > u\}\mathbf{Pr}\{X_i = x \mid X_i > u\} = \zeta f(x - u; \hat{\sigma}, \xi), \tag{3.36}$$

where $f(\cdot; \bar{\sigma}, \xi)$ is the density function of GPD having the following parameters, $\bar{\sigma}$ and ξ. Thus, the product of the independent observations gives the likelihood of

$$\mathbf{L}(\zeta, \hat{\sigma}, \xi; x_1, \cdots, x_n) = (1 - \zeta)^{n - n_u} \prod_{i=1}^{n_u} \zeta \hat{\sigma}^{-1} \left[1 + \xi \left(\frac{x_i - u}{\hat{\sigma}}\right)\right]^{-\frac{1}{\xi} - 1}, \tag{3.37}$$

where n_u is the number of exceedance over the threshold, u. If u is very high, then n_u is said to be smaller than n, that is

$$(1 - \zeta)^{n - n_u} \approx (1 - \zeta)^n \approx \exp\{-n\zeta\}. \tag{3.38}$$

This implies that by using $\bar{\sigma} = \sigma + \xi(u - \mu)$ and (3.34), we have:

$$\begin{aligned} \zeta \hat{\sigma}^{-1} \left[1 + \xi \left(\frac{x_i - u}{\hat{\sigma}}\right)\right]^{-\frac{1}{\xi} - 1} &= (n\hat{\sigma})^{-1} \left[1 + \xi \left(\frac{x_i - \mu}{\hat{\sigma}}\right)\right]^{-\frac{1}{\xi} - 1} \times \left[1 + \xi \left(\frac{u - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \\ &= (n\sigma)^{-1} \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi} - 1}. \end{aligned} \tag{3.39}$$

3.6.6 The maximum likelihood of the point process model

Suppose that $\mathcal{A} = (0, 1) \times [u, \infty)$, where u represents the threshold frequency (Khuluse, 2010; Coles, 2001). Then, the observations

$$\{(t_1, x_1), (t_2, x_2), (t_3, x_3), \cdots, (t_{N(A)}, x_{N(A)})\}$$

can be treated as observed exceedances. Therefore, the term n_y is multiplied by $\Lambda(A)$ in order to represent the extreme value limits in annual forms. We now assume that the Poisson process is valid and determine the likelihood function:

$$\begin{aligned} \mathrm{L}(A;\mu,\sigma,\xi) &= \exp\{-\Lambda(A)\}\prod_{i=1}^{N(A)}\lambda(t_i,\lambda_i) \\ &= \exp\left\{-n_y\left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}\right\}\times\prod_{i=1}^{N(A)}\frac{1}{\sigma}\left(1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}-1}. \end{aligned} \tag{3.40}$$

The parameter estimates are determined by taking the logarithm of 3.40 and minimise it. The parameters correspond to that of the GEVD. The GPD parameter, which is the scale, is determined in this manner $\hat{\sigma}_* = \hat{\sigma} + \hat{\xi}(u - \hat{\mu})$. But the shape parameter and threshold are the same as that of the GPD model.

Quantile estimation of the point process

Before presenting the quantile function of the point process model, it is important to estimate the threshold exceedance proportion given by:

$$\hat{\tau} = \frac{n_u}{n}. \tag{3.41}$$

Thus, the quantile function of the point process is:

$$x_N = u + \frac{\sigma_*}{\xi}[(\tau n_y N)^{\xi} - 1,] \tag{3.42}$$

which corresponds to that of the GPD.

3.7 Model diagnostics

The main purpose of this section is to obtain models that are adequate representations of the observed data. However, there are cases where all models fit the observed data to a similar degree, making it difficult to determine which model is the best. Among several statistical methods developed to search for the best model, are the following: stepwise regression, likelihood ratio tests, Alkaike information criterion (AIC) and Bayesian information criterion (BIC). This study concentrates on the AIC and BIC because the first two methods have some limitations when comparing at least two models (Takane and Hamparsum, 1987).

3.7.1 Akaike information criterion

Takane and Hamparsum (1987) define AIC as a useful statistic for statistical model selection and evaluation. The procedure was developed by Akaike in 1973 and is defined as follows:

$$\mathrm{AIC} = -2\log(\mathrm{L}) + 2\mathrm{K}, \tag{3.43}$$

where K is the number of parameters in the model and L is the value of the likelihood function. One important advantage of AIC is that, it is simple and easy to use. Furthermore, another important aspect of AIC is that, the best model chosen does not imply the true model, but it means that the model is best among competing models. The selection rule states that, the best model will be the one with the lowest value of AIC (Takane and Hamparsum, 1987).

3.7.2 Bayesian information criterion

The development of BIC uses the concept of AIC. By the early 1978, Glideon Schwarz added a penalty term to the AIC equation, which resulted in the pro-

cedure called the Bayesian information criterion (BIC), defined by:

$$\text{BIC} = -2\text{log}(L) + \text{K}\text{log}(n), \tag{3.44}$$

where L is the maximised value of the likelihood function, n is the number of observations and K is the number of parameters in the model. The selection rule states that the best model will be the one with the lowest value of BIC.

Chapter 4

Results and discussion

4.1 Introduction

The literature in Section 2.4 explained how climate change has affected the characteristics of rainfall patterns in South Africa. It also revealed that rainy seasons differ from one location to another. This chapter presents the analysis of average monthly rainfall data in South Africa and it is organised into two parts. The first part consists of stationarity test, summary statistics and fitting of candidate distributions, while the second part presents an extreme value analysis of the given data.

4.2 Data description

Secondary data on average monthly rainfall (in millimeters) for the period 1940-2017 obtained from the South Africa Weather Service (SAWS) is used in this study.

The statistical analysis analysis was performed using the statistical package and particular packages such ismev, evd, extRemes and fitdistr were utilised.

4.3 Descriptive Statistics

In this section, we present the summary statistics and the test for stationarity of the data.

In Table 4.1, the test for stationarity was conducted using augmented Dickey-Fuller (ADF) test.

Table 4.1: Stationarity test using ADF.

Name	t-stat
ADF	-9.41
P-value	0.01

The level of significance used in this study is 5%. Since p-value is 0.01 in Table 4.1, this implies that the null hypothesis stating that the time series is not stationary was rejected. According to the ADF test in Table 4.1, the time series data is stationary.

Table 4.2: Summary statistics of the rainfall data.

min	mean	median	Q1	Q3	max	kurtosis	skewness
3.20	51.15	44.80	19.57	75.03	175.00	3.05	0.80

Table 4.2 presents the summary statistics for average monthly rainfall data. The results in Table 4.2 reveal that the average monthly rainfall readings for

South Africa range from 3.20 mm to 175.00 mm, with a median of 44.80 mm. The data is positively skewed (mean $>$ median) and the kurtosis value (which is greater than 3) suggests that the data follows a heavy-tailed distribution. However, the kurtosis value in Table 4.2 is not far from 3, hence one can conclude that the data might follow a normal distribution. The above findings give rise to the next section.

4.4 Candidate distributions

This section presents an assessment on the goodness-of-fit of candidate distributions. The study focuses on the following candidate distributions: log-normal, Pareto, gamma and Weibull. Figure 4.1 presents the diagnostic plots of the log-normal and Pareto distributions.

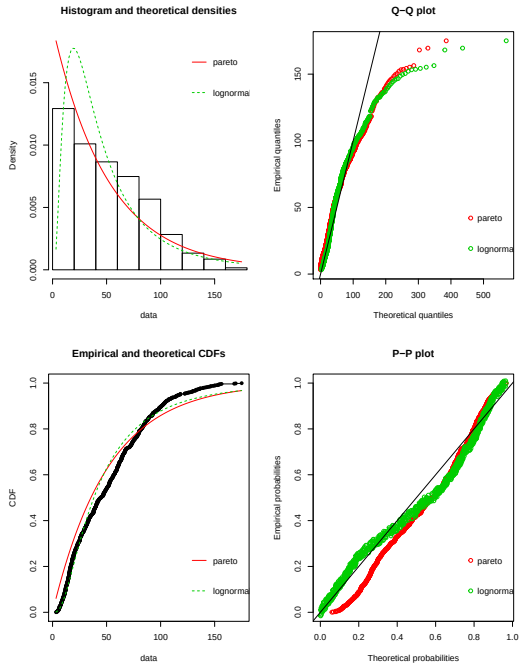

Figure 4.1: Diagnostic plots for log-normal and Pareto distributions (Key: The red line represents the Pareto distribution and the green line represents the log-normal distribution).

From Figure 4.1, the quantile-quantile (Q-Q) plot suggests a lack of fit for both the Pareto and log-normal distributions at the tails. Furthermore, the probability-probability (P-P) plot shows a lack of fit at the centre of both the Pareto and the log-normal distributions.

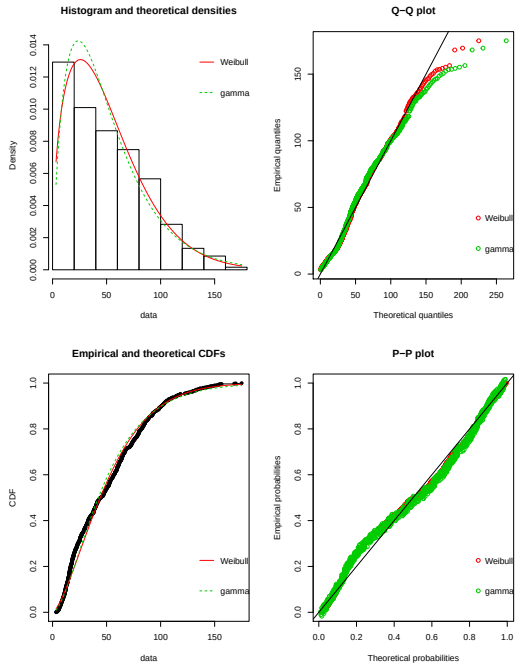

Figure 4.2: Diagnostic plots for Weibull and gamma distributions (Key: The red line represents the gamma distribution and the green line represents the Weibull distribution).

Figure 4.2 presents the diagnostic plots for the Weibull and gamma distributions. The diagnostic plots for the Q-Q plot in Figure 4.2, reveal that there is lack of fit at the tails of both the Weibull and gamma distributions. However, the P-P plot suggests a reasonably good fit for both the Weibull and gamma distributions. We next make use of the goodness-of-fit selection criterion to choose the best distribution(s) to represent the average monthly rainfall data.

Table 4.3: Selection of the most appropriate parent distribution.

Name	Pareto	Weibull	gamma	log-norm
Akaike's Information Criterion (AIC)	9241.879	9059.184	9061.132	9122.403
Bayesian Information Criterion (BIC)	9251.563	9068.867	9070.815	9132.086

Table 4.3 presents results for the goodness of fit tests AIC and BIC. The results in Table 4.3 suggest that the Weibull distribution is the best distribution to represent the average monthly rainfall in South Africa based on both the AIC and BIC. In the next section, we determine the return levels and return period for the Weibull distribution.

4.4.1 Estimation of return levels and return periods using the Weibull distribution

In this section we present the return levels and their corresponding return periods for the Weibull distribution.

Table 4.4: Quantile estimates and the number of exceedances based on the Weibull model.

Quantiles	Rainfall (mm)	T (year)	Number of exceedances
90th	100.23	10	98
95th	120.02	20	44
97.5th	138.41	40	21
98th	144.10	50	16
99th	161.13	100	3

Table 4.4 presents the results of the return levels and their corresponding return periods for the Weibull distribution. Thus, the average monthly rainfall that is expected, to be exceeded at least once every 20 years (0.95 quantile) is 120.02 mm. This implies that some areas in South Africa will have a greater

chance of receiving average monthly rainfall above 120.02 mm at least once every 20 years.

Mohamed and Ibrahim (2016) fitted probability distributions to an annual rainfall data in Sudan. Their results revealed that the normal and gamma distributions were the best distributions. In another study Kang and Yusof (2013) fitted candidate distributions to the rainfall data in Malaysia. Their study showed that Wakeby, generalised extreme value distribution (GEVD) and Weibull distributions turn out to perform well in the estimation which support the findings in this study.

4.5 Extreme value analysis

This section presents the analysis of two building block approaches of extreme value theory (EVT). The section is divided into four subsections. The first subsection presents the fitting of GEVD model. The second one is the GEVD for r-largest order statistics. The third subsection presents the fitting of the generalised Pareto distribution (GPD) model, while the forth and last subsection presents the point process approach.

4.5.1 GEVD model

Table 4.5: Parameter estimates and standard errors (in parentheses) of the GEVD.

Location (μ)	Scale (σ)	Shape (ξ)	95% CI of ξ	Neg. log-likelihood (λ)
107.41 (3.15)	24.53 (2.30)	-0.25 (0.09)	(-0.43, -0.07)	362.31

Table 4.5 represents the parameter estimates of the GEVD with standard errors (in brackets). The sign of the shape parameter in Table 4.5 is negative,

which suggests that the data can be modelled by a distribution that falls in the Weibull domain of attraction. The normal confidence interval of the shape parameter does not include zero, which further confirms that the data can be modelled by a distribution that falls in the Weibull domain attraction.

Chikobvu and Chifurira (2015) modelled minimum rainfall for Zimbabwe using GEVD. Their results revealed that the distribution that can best fit the data fall in the Weibull domain of attraction which is in line with our findings in this study.

The results for the profile likelihood are presented in Tables A1 and A2. The main purpose of profile likelihood is to produce accurate confidence limits about the parameter estimates (Coles, 2001). The 95% normal confidence intervals of location, μ, and scale, σ, from Table A2, are (101.24, 113.58) and (20.03, 29.02), respectively. The 95% confidence intervals obtained from the profile likelihood of parameters μ and σ are (101.24, 113.57) and (20.03, 29.02), respectively (Table A1), which are the same for μ and σ to those in Table A2. The 95% normal confidence interval of the shape parameter, ξ, obtained from the profile likelihood in Table A1 is (-0.41, 0.05), which is slightly different to the one in Table 4.5.

Figure 4.3 presents the diagnostic plots for the GEVD. The results in Figure 4.3 reveal that both the Q-Q and P-P plots appear linear, which implies a good fit for the GEVD model. The density plot also appears to follow a normal distribution, which indicates that the GEVD model fits well to the rainfall data.

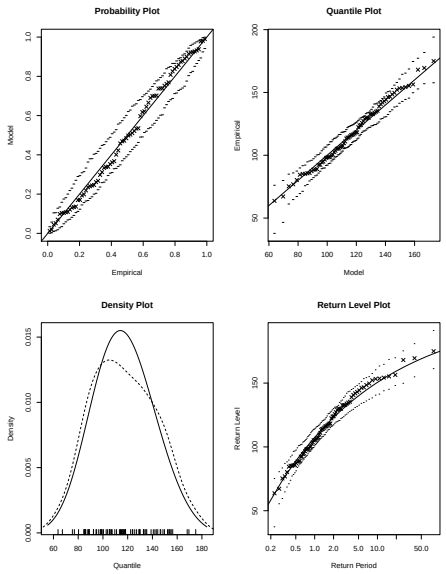

Figure 4.3: Diagnostic plots for the GEVD.

Table 4.6: GEVD goodness-of-fit test.

Goodness-of-fit tests	Statistic	P-values
Anderson-darling test	0.39	0.38
Kolmogrov-Smirnov test	0.07	0.54

Table 4.6 presents an assessment of the goodness-of-fit tests for the GEVD using the Anderson Darling (A-D) and Kolmogorov-Smirnov (K-S) tests. The results from Table 4.6 reveal that the null hypotheses of both A-D and K-S tests were not rejected since the p-values (0.38 and 0.54, respectively) were greater than 5% level of significance. This implies that indeed the GEVD model is appropriate for the average monthly rainfall for South Africa.

The EVT results in this section are in agreement with those based on the candidate distributions in the previous section. Both results revealed that the Weibull distribution is the best fitting model for the time series data in this study.

Quantile estimation of the GEVD model

In this section, we determine the average monthly rainfall expected to be exceeded at least once every T years.

Table 4.7: Quantile estimates and number of exceedances based on the GEVD model.

Quantiles	Rainfall (mm)	T (years)	Number of exceedances
90th	149.80	10	9
95th	159.11	20	3
97.5th	166.77	40	3
98th	168.96	50	2
99th	174.99	100	0

Equation (3.4) was used to determine the return levels for the GEVD. Column 4 of Table 4.7 presents the return levels, corresponding return periods and the number of observed average monthly rainfall that is greater than the estimated tail quantiles. Results from Table 4.7 reveal that the 0.95 quantile corresponds to $z_{0.05} = 159.11$ mm which is the 20-year return period in Table 4.7. This implies that an average monthly rainfall of 159.11 mm is expected to be exceeded, at least once every 20 years. However, this magnitude is lower than the maximum value of 175 mm in our observed actual data which in turn is equal to the 100-year return level. Therefore, based on these GEVD findings, it cannot be concluded that South Africa will expect extreme average monthly rainfall in the near future. Another consideration was the 100-year return level. Table 4.7 present the 100-year return level of the GEVD (174.99) which is slightly equivalent to the maximum observed value average monthly rainfall for South Africa.

The 95% normal and profile likelihood confidence intervals for the 20-year, 50-year and 100-year return levels are presented in Tables A3 and A4 respectively.

The 95% confidence intervals obtained from profile likelihood in Table A4 for all the return levels are slightly different from those in Table A3. It can be observed that the 95% profile likelihood confidence intervals in Table A4 are slightly higher than the corresponding 95% confidence intervals in Table A3.

4.5.2 GEVD for r-largest order statistics

In this section, we present the fitted model of the GEVD for r-largest order statistics (GEVD$_r$). Firstly, we start by estimating model parameters using the maximum likelihood estimation method. Table 4.8 shows the estimates of location($\hat{\mu}$), scale($\hat{\sigma}$), shape($\hat{\xi}$), the 95% confidence interval for the shape parameter and the negative log-likelihood (λ_i).

Table 4.8: Parameter estimates and standard errors (in parentheses) of r-largest order statistics models fitted to average monthly rainfall data.

r	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	CI(95%) of ξ	λ_i
1	107.41(3.15)	24.53(2.30)	-0.25(0.09)	(-0.43,-0.07)	362.31
2	95.42(1.98)	21.65(1.44)	-0.09(0.07)	(-0.23,0.05)	718.56
3	87.29(1.61)	21.96(1.13)	-0.10(0.05)	(-0.20,-0.002)	1080.04
4	79.61(1.37)	21.49(0.98)	-0.05(0.04)	(-0.13,0.03)	1441.55
5	72.94(1.23)	21.45(0.89)	-0.02(0.04)	(-0.09,0.06)	1807.80
6	66.87(1.19)	23.10(0.84)	-0.05(0.03)	(-0.11,0.01)	2195.52
7	60.50(1.19)	24.74(0.85)	-0.06(0.03)	(-0.12,-0.0012)	2596.49
8	54.09(1.18)	25.92(0.85)	-0.04(0.03)	(-0.1,0.02)	3002.42
9	48.09(1.16)	26.38(0.87)	-0.12(0.03)	(-0.18,-0.06)	3404.04
10	42.46(1.13)	26.31(0.87)	0.04(0.04)	(-0.04,0.12)	3801.57
11	37.30(1.09)	25.70(0.87)	0.10(0.04)	(0.02,0.18)	4193.00
12	32.28(1.04)	24.67(0.86)	0.18(0.04)	(0.10,0.26)	4576.79

The shape parameter estimates in Table 4.8 are negative for $r \leq 9$, suggesting that the data can be modelled by a distribution that fall in the Weibull domain of attraction. The confidence limits for the shape parameter when $r = 2, 4, 5, 6, 8, 10$ includes zero which indicates that the Gumbel distribution

might also be suitable to model the data. The value of the standard errors for all parameter estimators are stable, especially for the shape parameter when $r \leq 6$. According to Nemukula and Sigauke (2018), Guedes and Scotto (2004) and Coles (2001), this suggests a good fit for $r \leq 6$.

The results obtained from the GEVD$_r$ in Table 4.8 for $r \leq 9$ reveal that the data can be modelled well by a distribution in the Weibull domain of attraction. Figures 4.4, 4.5 and 4.6 display the diagnostic plots for $r = 2, 5$ and 8, respectively. Other plots are included in the Figures A5 to A11.

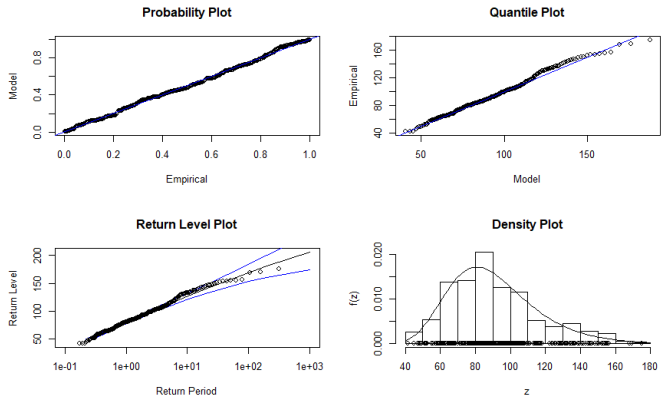

Figure 4.4: Diagnostic plots showing the GEVD$_r$ fit of average monthly rainfall for $r = 2$.

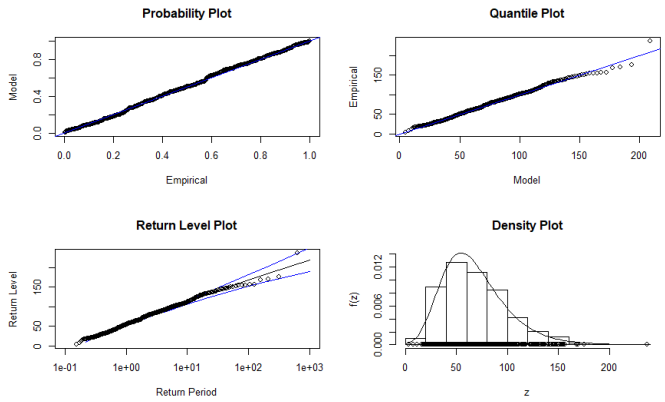

Figure 4.5: Diagnostic plots showing the GEVD$_r$ fit of average monthly rainfall for $r = 5$.

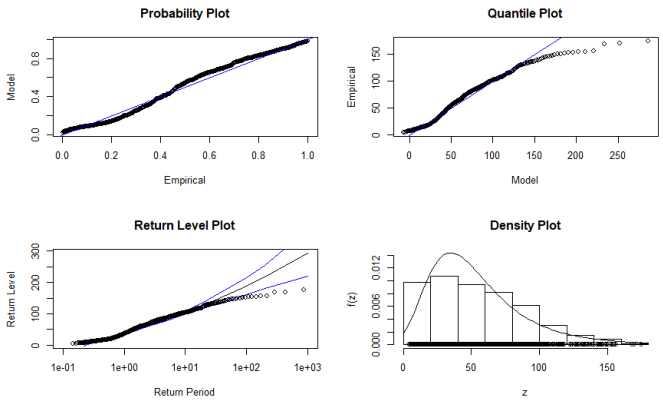

Figure 4.6: Diagnostic plots showing the GEVD$_r$ fit of average monthly rainfall for $r = 8$.

Figure 4.5 suggests that the model fitted with $r = 5$ order statistic is a good fit compared to those fitted with $r = 2$ order statistic (Figure 4.4) and $r = 8$ (Figure 4.6). The diagnostic plots alone are not sufficient to reveal the adequacy of the GEVD$_r$ fits to the data. Another procedure to assess the goodness-of-fit is the deviance statistic. The mechanism behind this procedure is that, it uses the maximum likelihood function for r_i and r_j to obtain deviance statistics to be compared with the chi-square distribution with one degree of freedom. Table 4.9 presents an assessment of goodness-of-fit using the deviance statistics. The test will validate the model based on r_i relative to r_j.

Table 4.9: The deviance statistics.

$D_{(1,2)}$	$D_{(2,3)}$	$D_{(3,4)}$	$D_{(4,5)}$	$D_{(5,6)}$
-712.5	-722.96	-723.02	-732.5	-775.44

The critical value of the χ^2_1 distribution is 3.84. In deviance statistics we compare the log-likelihood estimates of the following statistics: $D_{(1,2)}, D_{(2,3)}, D_{(3,4)}, D_{(4,5)}$ and $D_{(5,6)}$ in Table 4.9. Thus all the statistics are less than 3.84 meaning that at 5% level of significance we fail to reject $\lambda(r_i)$ for $i = 2, 3, \cdots, 6$ (Nemukula and Sigauke, 2018; Guedes and Scotto, 2004). In other words, all the log-likelihood

estimates for $r = 2, 3, \cdots, 6$ order statisitics are valid. Therefore according to deviance statistics test and diagnostic plots we can conclude that $r = 5$ is a reasonable order statistic in this study.

The 95% confidence intervals for the location, μ, scale, σ, and shape, ξ, parameters are presented in Table A6. Table A7 presents the 95% confidence intervals for the location, scale and shape parameters obtained from profile likelihood for $r = 5$. The 95% confidence intervals obtained from the profile likelihood for the location and scale parameters are slightly different from those in Table A6. Moreover, the 95% normal confidence interval for the shape parameter in Table A6 is similar to the one obtained from profile likelihood (Table A7). The return levels and their corresponding return periods for $r = 5$ are presented in Table 4.10.

Table 4.10: Quantile estimates and the number of exceedances based on the GEVD model for r-largest order statistics for $r = 5$.

Quantiles	Rainfall (mm)	T (years)	Number of exceedances
90th	120.14	10	44
95th	134.80	20	21
97.5th	148.97	40	10
98th	153.45	50	6
99th	167.21	100	2

The 95^{th} percentile corresponds to a 20-year return period. Therefore, on average, 134.80 mm of monthly rainfall is expected to be exceeded at least once every 20 years. It was observed that the estimated 100-year return level of 167.21 mm is far less than the largest observation of the actual data of 175 mm. Hence, based on the results in Table 4.10 of GEVD$_r$ it cannot be concluded that heavy rainfall is expected in South Africa in the near future years. The 100-year return level of the GEVD$_r$ (167.21 mm) from 4.10 is lower than that of the GEVD (174.99 mm).

The profile likelihood confidence intervals were determined for $r = 5$. Tables A8 and A9 present the 95% confidence intervals obtained from quantile estimates and profile likelihood, respectively. The confidence limits show that the two sets of confidence intervals are slightly different.

4.5.3 GPD model

In this section, we present the analysis of the results from fitting the generalised Pareto model to the average monthly rainfall in South Africa. We start by using the mean residual life and threshold stability plots to determine an appropriate threshold. We later present the full model of the GPD. Figure 4.7 presents the mean residual life plot for the data.

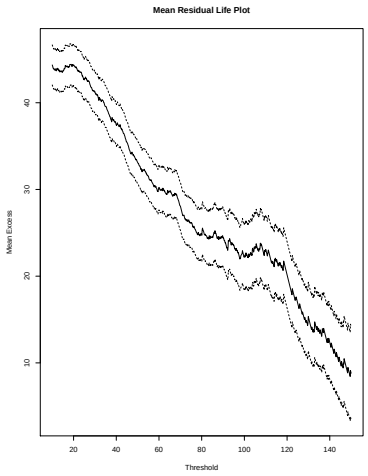

Figure 4.7: Mean residual life plot for average monthly rainfall for GPD model.

Figure 4.7 provided some evidence of linearity above $u = 22$ mm for the average monthly rainfall. Figures 4.8 and 4.9 present the threshold stability plots which are also essential when determining an appropriate threshold.

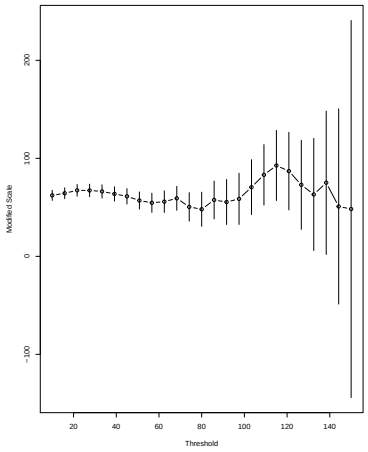

Figure 4.8: Threshold stability plot for the modified scale parameter for GPD model.

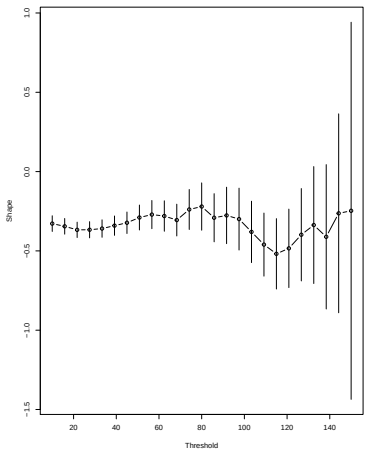

Figure 4.9: Threshold stability plot for the shape parameter for GPD model.

The threshold stability plots in Figures 4.8 and 4.9 suggest that the stable threshold is $u = 24$ mm. Using threshold stability and mean residual life plots, the most appropriate threshold is chosen to be $u = 24$. Therefore, 652 exceedances were extracted with a proportion above 0.6966.

Table 4.11: Parameter estimates and standard errors (in parentheses) of the GPD model.

Scale (σ)	Shape (ξ)	CI(95%) of ξ	CI(95%) of σ	Neg. log-likelihood (λ)
58.39 (2.59)	-0.37 (0.03)	(-0.43, -0.31)	(53.30, 63.45)	3064.99

From Table 4.11, the shape parameter is negative, which suggests that the GPD has a light-tailer than the exponential distribution. The diagnostic plots are also presented in Figure 4.10.

The 95% normal confidence interval of scale, σ, and shape, ξ, from Table 4.13 are (53.30, 63.45) and (-0.42, -0.32), respectively. The 95% confidence intervals obtained from the profile likelihood for the parameters σ and ξ are (53.47, 63.65) and (-0.41, -0.31), respectively and are presented in Table A5, and are slightly different from those in Table 4.11.

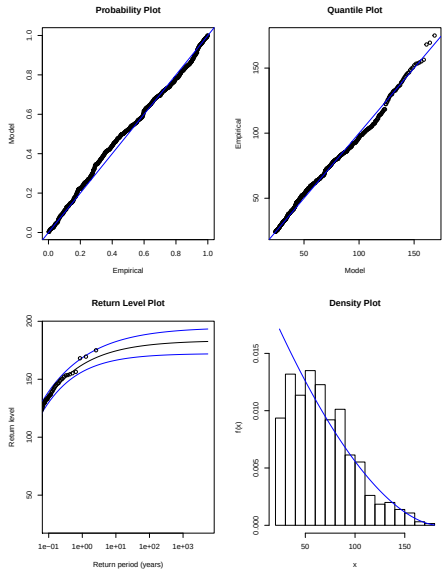

Figure 4.10: Diagnostic plots for GPD.

Figure 4.10 shows linearity and thus the GPD is appropriate and can well

represent the data.

Quantile estimation of the GPD model

Table 4.12 presents the return levels and their corresponding return periods, as well as the number of exceedances based on the GPD model.

Table 4.12: Quantile estimates and the number of exceedances based on the GPD model.

Quantiles	Rainfall (mm)	T (years)	Number of exceedances
90th	151.13	10	9
95th	158.07	20	3
97.5th	163.44	40	3
98th	164.89	50	3
99th	168.72	100	2

Before computing return levels, we first need to estimate the parameter: ($\hat{\zeta}_u = \frac{k}{n} = \frac{652}{936} = 0.6966$), where k represents the number of exceedances and n is the number of observations. The 0.95 quantile corresponds to a 20-year return period, and based on the results from Table 4.12, the average monthly rainfall expected to be exceeded, at least once, every 20 years is 158.07 mm. The 100-year return level based on the GPD results from Table 4.12 is 168.72 mm which is lower than the maximum observed average monthly rainfall for South Africa of 175 mm. Thus, the return levels for the GPD results are quite low compared to those obtained from the GEVD, but higher than those of the GEVD$_r$.

4.5.4 Point process approach

This subsection presents the analysis of average monthly rainfall using the point process approach. Since it was stated in Chapter 3 that this approach is more similar to models of exceedances (Khuluse, 2010; Coles, 2001), this implies that the mean residual life plot and threshold stability plot will be

used to determine a reasonable high threshold, u. It is also advisable to check for clusters when modelling exceedances (Nkrumah, 2017).

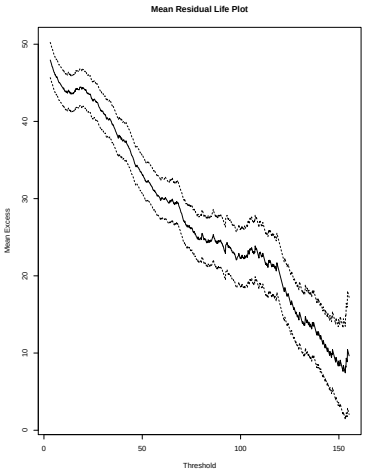

Figure 4.11: Mean residual life plot for average monthly rainfall for point process model.

The results in Figure 4.11 reveal that linearity is observed above $u = 22$ mm for average monthly rainfall data. The threshold stability plots results are presented in Figure 4.12.

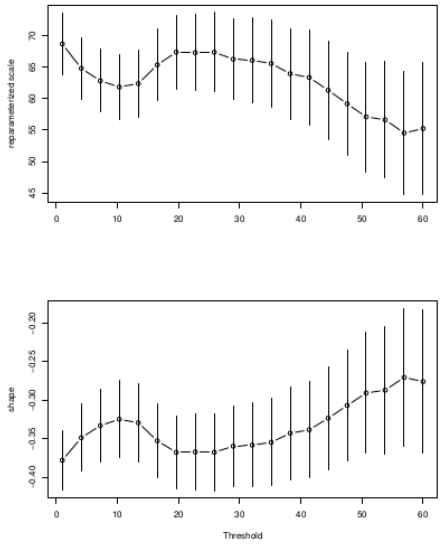

Figure 4.12: Threshold stability plots for the modified scale and shape parameters for point process model.

According to Figure 4.12, the graph is stable when $u = 24$. Therefore, using the results from mean residual and threshold stability plots, the chosen threshold is $u = 24$ for the point process model (Nkrumah, 2016; Khuluse, 2010; Coles, 2001).

Table 4.13: Parameter estimates and standard errors (in parentheses) of the point process model when $u = 24$.

Location (μ)	Scale (σ)	Shape (ξ)	Neg. log-likelihood (λ)	AIC	BIC
62.48 (3.10)	7.68 (0.80)	-0.37 (0.03)	105.55	217.11	230.55

From Table 4.13, we observed that the shape of the point process model is similar to that of the GPD in Table 4.11. This indicates that, the results obtained from the GPD are likely to be the silimar to those of the point process model. Table A10 presents the 95% normal confidence interval of the location, scale and shape parameters.

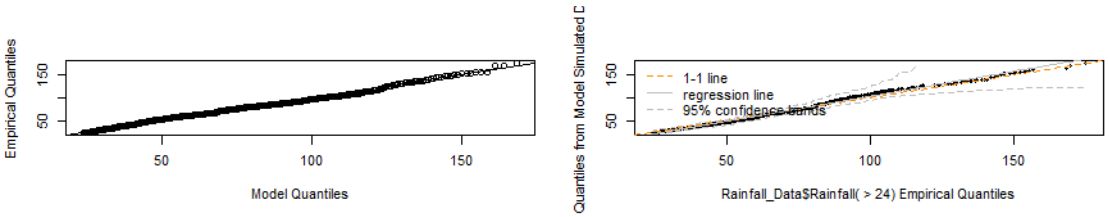

Figure 4.13: Diagnostic plots for the point process model for $u = 24$.

The results from Figure 4.13 reveal that both the P-P and Q-Q plots are linear, suggesting a good fit for the point process model when $u = 24$.

Table 4.14: Quantile estimates and the number of exceedances based on the point process model.

Quantiles	Rainfall (mm)	T (years)	Number of exceedances
90th	174.26	10	1
95th	176.40	20	0
97.5th	178.01	40	0
98th	178.44	50	0
99th	179.58	100	0

The point process return levels and their corresponding return periods for $u = 24$ are presented in Table 4.14. The 0.95 quantile for the point process results from Table 4.14 suggests that 176.40 mm is the average monthly rainfall expected, to be exceeded at least once every 20 years. This estimate of return level is slightly above the maximum observed value with average monthly rainfall of 175.0 mm which suggests that South Africa might see the floods of February 2000 coming back again more frequently. The 100-year return level based on the point process model is estimated to be 179.58 mm which is quite higher than the maximum value in our actual data for South Africa. Table A11 presents 95$\%$ normal confidence limits of quantile estimates of the point process model.

4.6 Summary of the chapter

The augmented Dickey-Fuller test revealed that the average monthly rainfall data for South Africa is stationary. The assessment of goodness-of-fit on the candidate distributions was also conducted. The diagnostic plots with the help of K-S and A-D tests revealed that the best candidate parent distribution that can model the average monthly rainfall data for South Africa is the distribution that falls in the Weibull domain of attraction.

The GEVD was fitted to the average monthly rainfall for South Africa for the period 1940-2017 using block maxima approach. The GEVD results revealed that the data can be modelled by a distribution that falls in the Weibull domain of attraction. The quantile estimation of the GEVD revealed that the maximum value of the observed average monthly rainfall in the series of 175 mm is equal to the 100-year return level. That is, the year 2000 average monthly rainfall of 175 mm has a return period of 100 years.

The GEVD$_r$ was fitted to average monthly rainfall data for South Africa. The order statistic, $r = 5$, gave a suitable fit for the data. The deviance statistics and diagnostic plots also played an important role in determining the best model when using the r-largest order statistics approach. The r-largest order statistics approach also revealed the distribution that can model the average monthly rainfall falls in the Weibull domain of attraction. However, the 95% confidence limits results suggest that the Gumbel distribution may also be a suitable distribution to model average monthly rainfall for South Africa.

The GPD and point process models were fitted to average monthly rainfall for South Africa. The mean residual and threshold stability plots were used to determine the thresholds. A threshold of $u = 24$ was obtained for both the GPD and point process model. The quantile estimates of the point process model were found to be higher than those of GEVD, GPD and GEVD$_r$. The 100-year return level of the point process model (179.58 mm) is greater than the maximum observed average monthly rainfall of 175 mm which in turn is equal to the 100-year return level of the GEVD (174.99 mm), whereas the 100-year return level of the GPD (168.72 mm) and GEVD$_r$ (167.21 mm) are slightly lower.

The profile likelihood method was incorporated in determining the 95% confidence intervals for the GEVD, GEVD$_r$ and GPD.

Chapter 5

Conclusion and Recommendations

5.1 Introduction

The chapter presents the conclusion and recommendations based on the findings of the statistical analysis on average monthly rainfall for South Africa. The first part of the chapter presents concluding remarks based on the previous chapter. The second part offers some recommendations and future research directions.

5.2 Conclusion

In Chapter 1, five objectives of this study were stated. In Chapter 4 the time series was first checked for stationarity. The augmented Dickey-Fuller test

revealed that the series is stationary at 5% level of significance. The fitting of candidate parent distributions revealed the distribution that can model the average monthly rainfall for South Africa falls in the Weibull domain of attraction. The return level estimates of the Weibull distribution suggested that 120.02 mm is the average monthly rainfall expected, to be exceeded, once every 20 years. The 100-year return level of 161.13 mm based on the Weibull distribution is less than the maximum value of 175 mm observed in the series which occured in February 2000 and destroyed many households in the Limpopo province.

In the block maxima approach, two models, i.e. GEVD and GEVD$_r$ were fitted. The results of both GEVD and GEVD$_r$ reveals that the underlying distribution of the average monthly rainfall belongs to the Weibull domain of attraction. This means the distribution has a finite right end-point and hence do not increase indefinitely. The GEVD$_r$ results further suggested the Gumbel domain to be a possible model for the average monthly rainfall for South Africa. Furthermore, the GPD and point process models were also fitted using the peaks-over threshold approach.

The return levels of the estimated GEVD and GEVD$_r$ showed that 159.11 mm and 134.80 mm are respectively the average monthly rainfall expected, to be exceeded, at least once every 20 years. In the case of the GPD, 158.07 mm is the average monthly rainfall that is expected, to be exceeded, at least once every 20 years. Thus, this magnitude is greater than that of the GEVD$_r$ but less than that of the GEVD. On the other hand, the point process results showed that 176.40 mm is the average monthly rainfall that is expected to be exceeded at least once every 20 years. This is slightly higher than the maximum observed average monthly rainfall for South Africa, 175 mm.

The 0.95 quantile estimates from GEVD and GPD were close and also greater than the corresponding 0.95 quantile estimate of the GEVD$_r$ model. However, for the point process model, the 0.95 quantile estimate was greater than that of the other three models and also higher than the maximum observed value in the series. Thus, for planning purposes, the estimates from the point process offers a realistic estimate to help in obtaining exceedance probabilities beyond the observed maximum in the rainfall data.

The 100-year quantile return level estimates of the GEVD, GEVD$_r$, GPD and point process models revealed that the 100-year return level of the GEVD and point process model were equal to and greater than, respectively, the maximum observed average monthly rainfall for South Africa. Furthermore, the 100-year return of both the GPD and GEVD$_r$ were lower than the maximum observed average monthly rainfall for South Africa.

The 0.95 quantile estimate obtained from the point process model suggests that, some areas in South Africa are expected to experience heavy rainfall at least once every 20 years whereas the GPD, GEVD and GEVD$_r$ suggested otherwise. The 100-year return level of the point process model is higher as compared to those from the GEVD, GPD, GEVD$_r$ and the parent distribution Weibull.

Therefore, when we model using the block maxima approach, the GEVD and GEVD$_r$ revealed that we cannot conclude that South Africa might experience higher than expected rainfall in the near future years. The findings of the GPD model were similar to those of the GEVD$_r$ and GEVD whereas the findings from the point process model suggest that South Africa might experience higher rainfall in the forthcoming years.

5.3 Contribution

This section presents the major contribution of the study.

The impacts of maximum rainfall around the world has tormented people and animals and tempered with the daily activities of the society. Following the recent disruptions caused by high rainfall in Mozambique, Zimbabwe, Malawi and some parts of South Africa (OCHA, 2019), the findings from this study will act as an awareness tool for these countries. It will help reduce the impact of high rainfall and countries can better prepare for such disasters.

The extreme value models found and recommended in this study will act as a benchmark for future studies on average monthly rainfall for South Africa.

5.4 Future research

The study suggests some future research directions that may help improve the accuracy and realiability of the findings.

Since the study revealed some evidence that South Africa might experience higher than expected rainfall in the coming years based on the point process model, it is now left to the meteorologists and hydrologists to determine the locations of the likely impact or vulnerable areas. The study recommends that the results might be improved by modelling with multivariate extremes and Bayesian approach to include expert knowledge in estimation. The use of non-stationary time series and other parameter estimation methods such as moments may be employed in the future. On the peaks-over threshold approach, further studies might also consider using time-varying covariates and thresholds.

References

ADESINA, O., ADELEKE, I., AND OLADEJI, T. (2016). Using extreme value theory to model insurance risk of Nigeria's motor industrial. *The Journal of Risk Management and Insurance*, **20**, 40–51.

AGUILAR, E., BARRY, A. A., BRUNET, M., EKANG, L., FERNANDES, A., MASSONKINA, M., MBAH, J., MHANDA, A., DO NASCIMENTO, D. J., UMBA, T. C. P. O. T., TOMOU, M., AND ZHANG, X. (2009). Changes in temperature and precipitation extremes in Western Central Africa, Guinea Conakry and Zimbabwe. *Journal of Geophysical Research Atmospheres*, 1–11.

ARLTOVA, M. AND FEDOROVA, D. (2016). Selection of unit root test on the time series and value of AR(1) parameter. *STATISTIKA*, **96**, 47–64.

BALI, T. (2003). An extreme value approach to estimating volatility and Value-At-Risk. *The Journal of Business*, **76**, 83–108.

BALI, T. (2007). A generalised extreme value approach to financial risk measurement. *Journal of Money, Credit and Banking*, **39**, 1613–1649.

BEIRLANT, J., GOEGEBEUR, Y., AND TEUGELS, J. (2004). *Statistics of extremes: Theory and applications*. John Wiley and Sons Ltd.

BHAGWANDIN, L. (2017). *Multivariate exetreme value theory with an application to climate data in the Western Cape province*. MSc dissertation, University of Cape Town.

CHARRAS-GARRIDO, M. AND LEZAUD, P. (2013). Extreme value analysis: An introduction. *Journal de la Societe Francaise de Statistique*, 66–97.

CHIKOBVU, D. AND CHIFURIRA, R. (2015). Modelling of extreme minimum rainfall using GEV distribution for Zimbabwe. *South African Journal of Science*, **111**, 1–8.

CHU, P.-S., ZHAO, X., RUAN, Y., AND GRUBBS, M. (2008). Extreme rainfall events in the Hawaiian Islands. *Journal of Applied Meteorology and Climatology*, **48**, 502–516.

COLES, S. (2001). *An introduction to statistical modelling of extreme values*. 1st edition. Springer-Verlaq, London.

DALEY, D. J. AND VERE-JONES, D. (2003). *An Introduction to the theory of point processes: Volume I: Elementary theory and methods*. 2nd edition. Springer-Verlaq, New York.

DE WAAL, J. (2012). *Extreme rainfall distributions: Analysing change in the Western Cape*. MSc dissertation, Stellenbotch University.

DE WAAL, J., CHAPMAN, A., AND KEMP, J. (2017). Extreme 1-day rainfall distributions: Analysing change in the Western Cape. *South African Journal of Science*, **113**, 1–10.

DEBUSHO, L. AND DIRIBA, T. (2016). Bayesian modelling of summer daily maximum temperature data. *Advance Mathematics and Computational Science*, 126–133.

DIRIBA, T., DEBUSHO, L. K., BOTAI, J., AND HASSEN, A. (2014). Analysis of extreme rainfall at East London, South Africa. *South Africa Statistical Journal*, 25 – 32.

DIRIBA, T. A., DEBUSHO, L. K., BOTAI, J., AND HASSEN, A. (2017). Bayesian modelling of extreme wind speed at Cape Town, South Africa. *Environmental and Ecological Statistics*, **24** (2), 243–267.

DYSON, L. AND VAN HEERDEN, J. (2001). The heavy rainfall and floods over the northeastern interior of South Africa during February 2000. *South African Journal of Science*, **97** (3-4), 80–86.

EMBRECHT, P., RESNICK, S., AND SAMORODNITSKY, G. (1999). Extreme value theory as a risk management tool. *North American Actuary Journal*, **3** (2).

ENDER, M. AND MA, T. (2014). Extreme value modeling of precipitation in case studies for China. *International Journal of Scientific and Innovative Mathematical Research*, **2**, 23–36.

EWN (2019). News24: KwaZulu-Natal floods report.
URL: *https://www.news24.com/SouthAfrica/News/at-least-51-confirmed-dead-in-kzn-floods-reports-20190424*

FEN CHU, L., MCALEER, M., AND CHANG, C. C. (2012). Statistical modelling of extreme rainfall in Taiwan. *Atlantis Press*, 1–20.

FERNANDEZ, V. (2003). Extreme value theory: Value-At-Risk and returns dependence around the World. *Center of Applied Economics*, (161), 1–39.

FERREIRA, A. AND DE HAAN, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statisitics*, 276–298.

GOUDENHOOFDT, E., DELOBBE, L., AND WILLEMS, P. (2017). Regional frequency analysis of extreme rainfall in Belgium based on radar estimates. *Hydrology and Earth System Sciences*, **21**, 5386–5399.

GUEDES, S. C. AND SCOTTO, M. (2004). Application of the r largest-order statistics for long-term predictions of significant wave height. *Coastal Engineering*, **51** (5-6), 387–394.

HASAN, H., RADI, N. F. A., AND KASSIM, S. (2012). Modeling of extreme temperature using generalized extreme value distribution: A case of Penang. *World Congress on Engineering*, 1–6.

HENRY, J. B. AND HSIEH, P.-H. (2014). Extreme value analysis for partitioned insurance losses. *Casualty Actuarial Society*, **3**, 214–238.

HOGG, R., TANIS, E., AND ZIMMERMAN, D. (2015). *Probability and statistical inference*. 9th edition. Pearson, New York.

HUGUENY, S., CLIFTON, D., AND TARASSENKO, L. (2010). Probabilistic patient monitoring using extreme value theory: A multivariate, multimodal methodology for detection patient in deterioration state. *NIHR Biomedical Research Centre, Oxford*, 1–20.

JURY, M. (2012). Climate trents in Southern Africa. *South African Journal*, **109**, 1–11.

KANG, H. M. AND YUSOF, F. (2013). Determination of best-fit distribution and rainfall events in Damansara and Kelantan, Malaysia. *Matematika*, 43–52.

KHULUSE, S. A. (2010). *Modelling heavy rainfall over time and space*. MSc dissertation, University of Witwatersrand.

KRATZ, M. (2017). Extreme value theory and its applications to insurance and finance. *ETH Risk Center Zurich*, 1–61.

KRUGER, A. AND NXUMALO, M. (2017). Historical rainfall trends in South Africa: 1921-2015. *African Journals Online*, **43**, 285–296.

KULSHRESHTHA, S. N. (1998). A global outlook for water resources to the year 2025. *Kluwer Academic Publishers*, **12**, 167–184.

LUC MELICE, J. AND REASON, C. (2007). Return period of extreme rainfall at George, South Africa. *South African Journal of Science*, **103**, 499–501.

MAPOSA, D. (2016). *Statistics of extremes with applications to extreme flood heights in the lower Limpopo river basin of Mozambique*. PhD thesis, University of Limpopo.

MARTIN, R., PAXTON, B., AND VAN WILGEN, B. (2011). Quest: Tackling climate change in South Africa. *Academy of Science of South Africa*, **11**, 1–48.

MASEREKA, E. M., OCHIENG, G. M., AND SNYMAN, J. (2018). Statistical analysis of annual maximum daily rainfall for Nelspruit and its environs. *Journal of Disaster Risk Studies*, **10**, 23–36.

MEEHL, G. A., WASHINGTON, W. M., SANTER, B. D., COLLINS, W. D., ARBLASTER, J. M., HU, A., LAWRENCE, D. M., TENG, H., BUJA, L. E., AND STRAND, W. G. (2006). Climate change projections for the twenty-first century and climate change commitment in the CCSM3. *Journal of Climate*, **19** (11), 2597–2616.

MOHAMED, T. M. AND IBRAHIM, A. A. A. (2016). Fitting probability distribution of annual rainfall in Sudan. *Journal of Engineering and Computer Science*, **17**, 35–39.

NADARAJAH, S. (2005). Extremes of daily rainfall in West Central Florida. *Climatic Change*, **69** (2), 325–342.

NASON, G. P. (2006). *Stationary and non-stationary time series*. doi:10.1144/IAVCEI001.
URL: *https://doi.org/10.1144/IAVCEI001*

NEMUKULA, M. M. AND SIGAUKE, C. (2018). Modelling average maximum daily temperature using r-largest order statistics: An application to South African data. *Journal of Disaster Risk Studies*, **10** (1), 1–11.

NGAILO, T. J., REUDER, J., RUTALEBWA, E., NYIMVUA, S., AND MESQUITA, M. (2016). Modelling of extreme maximum rainfall using extreme value the-

ory for Tanzania. *International Journal of Scientific and Innovative Mathematical Research*, **4**, 34–45.

NKRUMAH, S. (2017). *Extreme value analysis of temperature and rainfall: Case study of some selected regions in Ghana*. MSc dissertation, University of Ghana.

OCHA (2019). Mozambique: The cyclone Idia and floods.
URL: *https://reliefweb.int/report/mozambique/mozambique-cyclone-idai-floods-situation-report-no-18-22-april-2019*

OLIVER, U. AND MUNG'ATU, J. K. (2018). Modelling extreme maximum rainfall using generalised extreme dalue distribution: Case study Kigali City. *International Journal of Science and Research*, 121–125.

PHAKULA, S. (2016). *Modelling seasonal rainfall characteristics over South Africa*. MSc dissertation, University of Pretoria.

REASON, C. J. C., HACHIGONTA, S., AND PHALADI, R. F. (2005). Interannual variability in rainy season characteristics over the Limpopo region of Southern africa. *International Journal of Climatology*, **25** (14), 1835–1853.

SHONGWE, M. E., OLDENBORGH, G. J. V., AND VAN DEN HURK, B. J. J. M. (2009). Projected changes in mean and extreme precipitation in Africa under global warming. Part I: Southern Africa. *Journal of Climate*, **22**, 3819–3836.

SMITH, E. (2005). *Bayesian modelling of extreme rainfall data*. PhD thesis, University of Newcatle.

TADROSS, M. AND JOHNSTON, P. (2012). A five city network to pioneer climate adaptation through participatory research and local action. *ICLEI - Local Governments for Sustainability*, 1–35.

TAKANE, Y. AND HAMPARSUM (1987). Akaike information criterion - Introduction. *Psychometrika*, 1.

TYSON, P. AND PRESTON-WHYTE, R. (2011). *The weather and climate of Southern Africa*. 2nd edition. Oxford University Press Southern Africa, Cape Town.

Appendix A

Plots and tables for Chapter 4

Table A1: GEVD 95% confidence interval from Profile log-likelihood.

	CI(95%)
Location (μ)	(101.24, 113.57)
Scale (σ)	(20.03, 29.02)
Shape (ξ)	(-0.41, 0.05)

Table A2: GEVD 95% normal confidence interval for location (μ) and scale (σ) parameters.

	CI(95%)
Location (μ)	(101.24, 113.57)
Scale (σ)	(20.03, 29.02)

Table A3: GEVD 95% normal confidence interval for the quantile estimation.

	CI(95%)
20-year return level	(149.63, 169.61)
50-year return level	(155.71, 182.35)
100-year return level	(158.47, 191.32)

Table A4: GEVD 95% confidence interval for the quantile estimation from Profile likelihood.

	CI(95%)
20-year return level	(151.45, 173.52)
50-year return level	(159.91, 191.26)
100-year return level	(164.50, 204.22)

Table A5: GPD confidence interval obtained from profile log-likelihood of scale (σ) and shape (ξ) parameters.

	CI(95%)
Scale (σ)	(53.47, 63.65)
Shape (ξ)	(-0.41, -0.31)

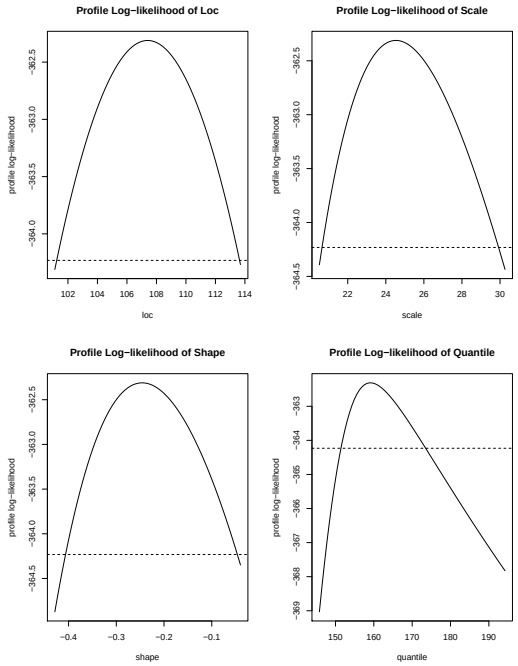

Figure A1: Profile log-likelihood for shape, location, scale and quantile estimation.

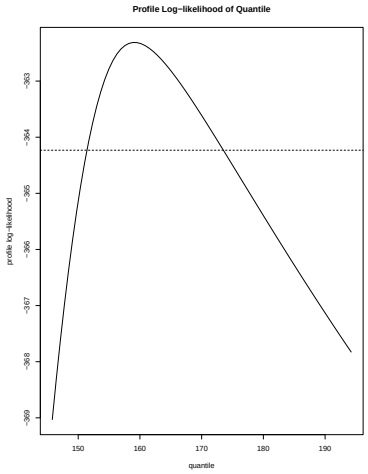

Figure A2: GEVD profile log-likelihood for quantile estimation for a 20-year return level.

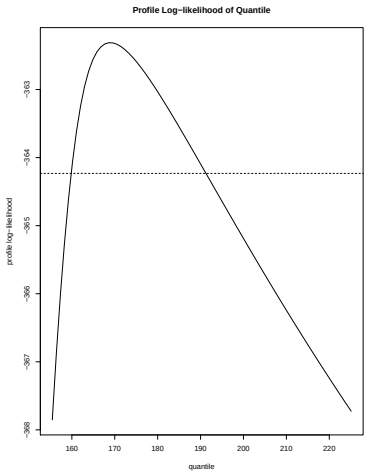

Figure A3: GEVD profile likelihood for quantile estimation for a 50-year return level.

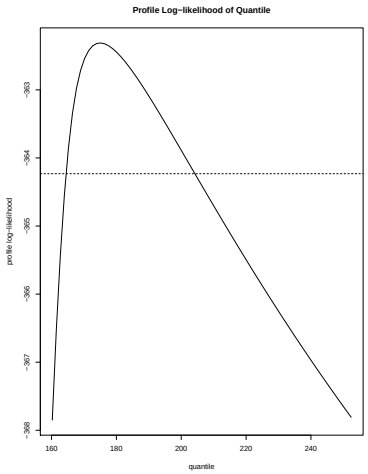

Figure A4: GEVD profile log-likelihood for

quantile estimation for a 100-year return level.

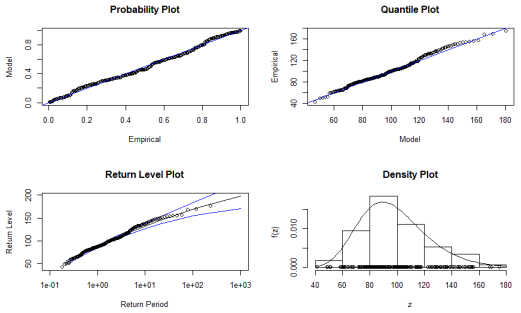

Figure A5: Diagnostic plots showing the fit of average monthly rainfall for $r = 1$.

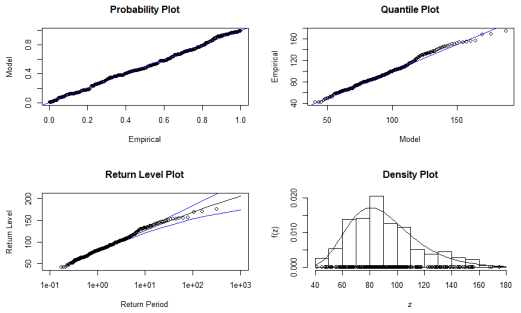

Figure A6: Diagnostic plots showing the fit of average monthly rainfall for $r = 2$.

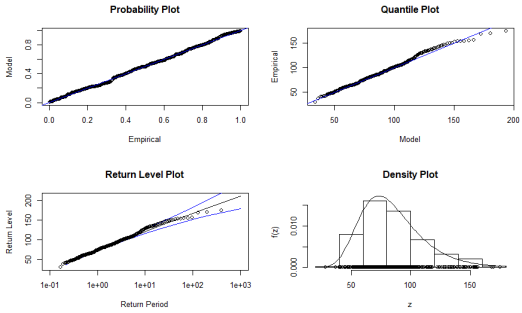

Figure A7: Diagnostic plots showing the fit of average monthly rainfall for $r = 3$.

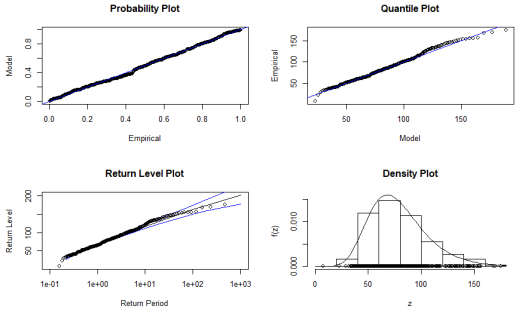

Figure A8: Diagnostic plots showing the fit of average monthly rainfall for $r = 4$.

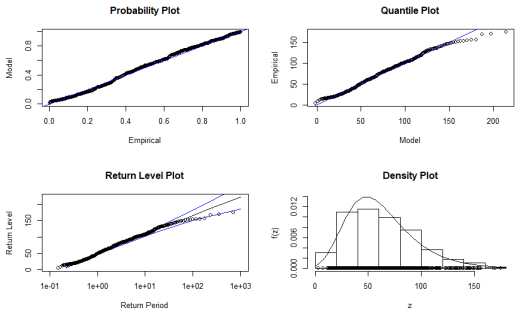

Figure A9: Diagnostic plots showing the fit of average monthly rainfall for $r = 6$.

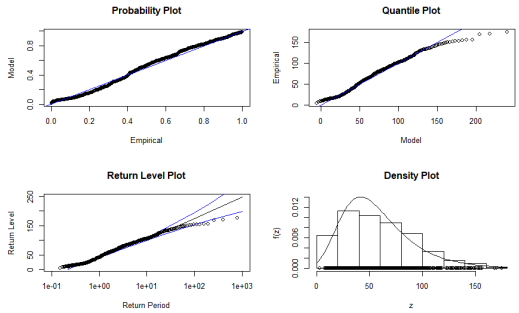

Figure A10: Diagnostic plots showing the fit of average monthly rainfall for $r = 7$.

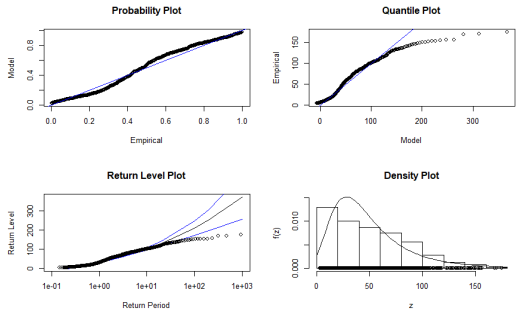

Figure A11: Diagnostic plots showing the fit of average monthy rainfall for $r = 9$.

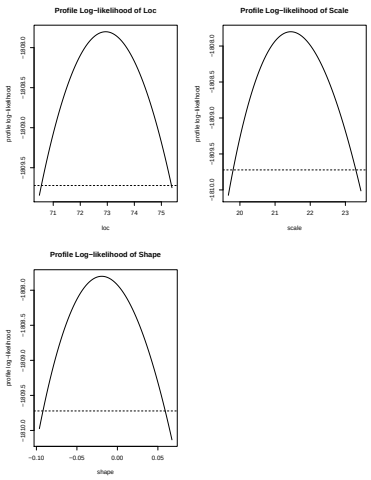

Figure A12: Profile log-likelihood for shape, location and scale parameters when $r = 5$.

Table A6: The 95% normal CI of GEVD for r-largest order statistics for scale, shape and location parameters when $r = 5$.

	95% CI
Location (μ)	(70.54, 75.34)
Scale (σ)	(19.71, 23.18)
Shape (ξ)	(-0.09, 0.06)

Table A7: The profile log-likelihood 95% CI of GEVD for r-largest order statisitcs for scale, shape and location parameters when $r = 5$.

	95% CI
Location (μ)	(70.57, 75.37)
Scale (σ)	(19.81, 23.29)
Shape (ξ)	(-0.09, 0.06)

Table A8: The confidence interval of GEVD for r-largest order statisitcs for the quantile estimation when $r = 5$.

	95%CI
20-year return level	(127.60, 142.15)
50-year return level	(152.17, 182.62)
100-year return level	(142.25, 164.83)

Table A9: The profile log-likelihood confidence interval of GEVD for r-largest order statisitcs for the quantile estimation when $r = 5$.

	95% CI
20-year return level	(128.47, 143.32)
50-year return level	(154.85, 186.06)
100-year return level	(144.07, 167.20)

Table A10: The 95% normal confidence intervals of the point process model for location, scale and shape parameters.

	95% CI
Location	(156.42, 168.55)
Scale	(6.11, 9.26)
Shape	(-0.42, -0.32)

Table A11: The 95% normal confidence intervals for the quantile estimates of the point process model.

	CI
10-year return level	(166.25, 182.27)
20-year return level	(167.85, 184.84)
40-year return level	(169.00, 187.01)
50-year return level	(169.30, 187.58)
100-year return level	(170.06, 189.09)

SOME SELECTED R CODES

The main R codes of the study

```
#Choosing the working directory
getwd()
#installing Packages for Extreme Value Analysis
install.packages(c("VGAM","rmutils"))
library(VGAM)
install.packages("extRemes")
library(extRemes)
install.packages("ismev")
library(ismev)
install.packages("evd")
library("evd")
install.packages("fitdistrplus")
library(fitdistrplus)
install.packages("nortest")
library(nortest)
install.packages("aTSA")
library(aTSA)
install.packages("actuar")
library(actuar)
install.packages("MASS")
library(MASS)
##################End of installation#####################
```

```
#Importing data to R
Rainfall_Data <- read.csv2("Datum.csv")

#Computing summary statistics
install.packages("moments")
library(moments)
kurtosis(Rainfall_Data$Rainfall)
skewness(Rainfall_Data$Rainfall)
summary(Rainfall_Data$Rainfall)

#Testing for Stationarity using Argumented Dickey Fuller
adf.test(Rainfall_Data$Rainfall)

####################Candidate distributions################

P <- fitdist(Rainfall_Data$Rainfall, "pareto",
start=list(shape = 0.4, scale = 1))
W <- fitdist(Rainfall_Data$Rainfall, "weibull")
G <- fitdist(Rainfall_Data$Rainfall, "gamma")
Lon <- fitdist(Rainfall_Data$Rainfall, "lnorm")

#Plotting diagnostic plots for Weibull and Gamma
par(mfrow = c(2, 2))
plot.legend <- c("Weibull", "gamma")
denscomp(list(W, G), legendtext = plot.legend)
qqcomp(list(W, G), legendtext = plot.legend)
cdfcomp(list(W, G), legendtext = plot.legend)
ppcomp(list(W, G), legendtext = plot.legend)
```

```
#Plotting diagnostic plots for Pareto and log-normal
par(mfrow = c(2, 2))
plot.legend <- c("pareto", "lognormal")
denscomp(list(P, Lon), legendtext = plot.legend)
qqcomp(list(P, Lon), legendtext = plot.legend)
cdfcomp(list(P, Lon), legendtext = plot.legend)
ppcomp(list(P, Lon), legendtext = plot.legend)

#Goodness of fit test using KPSSS and AD tests
ss <- gofstat(list(P, W, G, Lon), fitnames =
c("Pareto", "Weibull", "Gamma", "lnorm"))

######################THE BLOCK MAXIMA APPROACH#########################
# The GEVD model
#Importing data to R
DataGEVD <- read.csv2("dsbase_r1.csv")
DataGEVD.fit <- fgev(DataGEVD$Rainfall)
#Diagnostic plots
par(mfrow = c(2, 2))
plot(DataGEVD.fit)

# profile log-likelihood for parameters
DataGEVD.prof <- profile(DataGEVD.fit,conf = 0.95)
par(mfrow = c(2, 2))
plot(DataGEVD.prof)
confint(DataGEVD.fit)
```

```
# Profile log-likelihood for quantile
DataGEVD.qfit <- fgev(DataGEVD$Rainfall,prob = 0.05)
DataGEVD.qprof <- profile(DataGEVD.qfit, which = "quantile")
plot(DataGEVD.qprof)
confint(DataGEVD.qprof)

#Fitting GEVD for r-largest order statistics model

#Gev when r=1
R1 <- read.csv2("dsbase_r1.csv")
fit1 <- fevd(R1$Rainfall, type = "GEV", method = "MLE")
ci(fit1, type = "parameter")
plot(fit1)
return.level(fit1, return.period = c(10, 20, 40, 50, 100))

#Gev when r=2
R2 <- read.csv2("dsbase_r2.csv")
fit2 <- fevd(R2$Rainfall, type = "GEV", method = "MLE")
ci(fit2, type = "parameter")
plot(fit2)
return.level(fit1, return.period = c(10, 20, 40, 50, 100))

#Gev when r=3
R3 <- read.csv2("dsbase_r3.csv")
fit3 <- fevd(R3$Rainfall, type = "GEV", method = "MLE")
fit3
ci(fit4, type = "parameter")
plot(fit3)
```

```
return.level(fit3, return.period = c(10, 20, 40, 50, 100))

#Gev when r=4
R4 <- read.csv2("dsbase_r4.csv")
fit4 <- fevd(R4$Rainfall, type = "GEV", method = "MLE")
ci(fit4, type = "parameter")
plot(fit4)
return.level(fit4, return.period = c(10, 20, 40, 50, 100))

#Gev when r=5
R5 <- read.csv2("dsbase_r5.csv")
fit5 <- fevd(R5$Rainfall, type = "GEV", method = "MLE")
ci(fit5, type = "parameter")
plot(fit5)
return.level(fit5, return.period = c(10, 20, 40, 50, 100))

#Gev when r=6
R6 <- read.csv2("dsbase_r6.csv")
fit6 <- fevd(R6$Rainfall, type = "GEV", method = "MLE")
ci(fit6, type = "parameter")
plot(fit6)
return.level(fit6, return.period = c(10, 20, 40, 50, 100))

#Gev when r = 7
R7 <- read.csv2("dsbase_r7.csv")
fit7 <- fevd(R7$Rainfall, type = "GEV", method = "MLE")
ci(fit7, type = "parameter")
plot(fit7)
```

```
return.level(fit7, return.period = c(10, 20, 40, 50, 100))

#Gev when r=8
R8 <- read.csv2("dsbase_r8.csv")
fit4 <- fevd(R8$Rainfall, type = "GEV", method = "MLE")
ci(fit4, type = "parameter")
plot(fit8)
return.level(fit8, return.period = c(10, 20, 40, 50, 100))

#Gev when r =9
R9 <- read.csv2("dsbase_r9.csv")
fit9 <- fevd(R9$Rainfall, type = "GEV", method = "MLE")
ci(fit9, type = "parameter")
plot(fit9)
return.level(fit9, return.period = c(10, 20, 40, 50, 100))

#Gev when r=10
R10 <- read.csv2("dsbase_r10.csv")
fit4 <- fevd(R10$Rainfall, type = "GEV", method = "MLE")
ci(fit10, type = "parameter")
plot(fit10)
return.level(fit10, return.period = c(10, 20, 40, 50, 100))

#Gev when r=11
R11 <- read.csv2("dsbase_r11.csv")
fit11 <- fevd(R11$Rainfall, type = "GEV", method = "MLE")
ci(fit11, type = "parameter")
plot(fit11)
```

```
return.level(fit11, return.period = c(10, 20, 40, 50, 100))

#Gev when r=12
R12 <- read.csv2("dsbase_r12.csv")
fit12 <- fevd(R12$Rainfall, type = "GEV", method = "MLE")
ci(fit12, type = "parameter")
plot(fit12)
return.level(fit12, return.period = c(10, 20, 40, 50, 100))

####################THE PEAKS-OVER-THRESHOLD APPROACH#################
#GPD model
mrlplot(Rainfall_Data$Rainfall,tlim = c(10,150))
tcplot(Rainfall_Data$Rainfall,tlim = c(10,150))
fitGPD <- fevd(Rainfall_Data$Rainfall, threshold = 24, type = "GP", meth
plot(fitGPD)
return.level(fitGPD, return.period = c(10, 20, 40, 50, 100))

#Point Process Model
threshrange.plot(Rainfall_Data$Rainfall, r = c(1,60), nint =20)
mrlplot(Rainfall_Data$Rainfall)
fitpm <- fevd(Rainfall_Data$Rainfall, threshold = 24, type = "PP")
plot(fitpm)
ci(fitpm, type = "parameter")
ci(fitpm, return.period = c(10, 20, 40, 50, 100))
```