MODELLING LONG-TERM MONTHLY RAINFALL VARIABILITY IN SELECTED PROVINCES OF SOUTH AFRICA USING EXTREME VALUE DISTRIBUTIONS

by

VUSI NTIYISO MASINGI

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF 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

2021

Declaration

I, **Vusi Ntiyiso Masingi**, declare that this research titled “Modelling long-term monthly rainfall variability in selected provinces of South Africa using extreme value distributions” is my original work and has not been submitted for any degree at any other university or institution. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references.

Signature:......................Date:...............................
MASINGI VN

Abstract

Several studies indicated a growing trend in terms of frequency and severity of extreme events. Extreme rainfall could cause disasters that lead to loss of property and life. The aim of the study was to model the monthly rainfall variability in selected provinces of South Africa using extreme value distributions. This study investigated the best-fit probability distributions in the five provinces of South Africa. Five probability distributions: gamma, Gumbel, log-normal, Pareto and Weibull, were fitted and the best was selected from the five distributions for each province. Parameters of these distributions were estimated by the method of maximum likelihood estimators. Based on the Akaike information criteria (AIC) and Bayesian information criteria (BIC), the Weibull distribution was found to be the best-fit probability distribution for Eastern Cape, KwaZulu-Natal, Limpopo and Mpumalanga, while in Gauteng the best-fit probability distribution was found to be the gamma distribution. Monthly rainfall trends detected using the Mann–Kendall test revealed significant monotonic decreasing long-term trend for Eastern Cape, Gauteng and KwaZulu-Natal, and insignificant monotonic decreasing long-term trends for Limpopo and Mpumalanga. Non-stationary generalised extreme value distribution (GEVD) and non-stationary generalized Pareto distribution (GPD) were applied to model monthly rainfall data. The deviance statistic and likelihood ratio test (LRT) were used to select the most appropriate model. Model fitting supported stationary GEVD model for Eastern Cape, Gauteng and KwaZulu-Natal. On the other hand, model fitting supported non-stationary GEVD models for maximum monthly rainfall with nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter for Limpopo, while in Mpumalanga the non-stationary GEVD model, which has a nonlinear quadratic trend in the scale parameter and no variation in the location parameter fitted well to the maximum monthly rainfall data. Results from the non-stationary GPD models showed that inclusion of the time covariate in our models was not significant for Eastern Cape, hence the best-fit model was the stationary GPD model. Furthermore, the non-stationary GPD model with a linear trend in the scale parameter provided the best-fit for KwaZulu-Natal and Mpumalanga, while in Gauteng and Limpopo the non-stationary GPD model with a nonlinear quadratic trend in the scale parameter fitted well to the monthly rainfall data. Lastly, GPD with time-varying thresholds was applied to model monthly rainfall excesses, where a penalised regression cubic smoothing spline was used as a time-varying threshold and the GPD model was fitted to cluster maxima. The estimate of the shape parameter showed that the Weibull family of distributions is appropriate in modelling the upper tail of the distribution for Limpopo and Mpumalanga, while for Eastern Cape, Gauteng and KwaZulu-Natal, the exponential family of distributions was found to be appropriate in modelling the upper tail of the distribution. The dissertation contributes positively to the body of knowledge in extreme value theory application to rainfall data and makes recommendations to the government agencies on the long-term rainfall variability and their negative impact on the economy.

Dedication

This dissertation is dedicated to my lovely mother, Mrs Ndaheni Nyanisi Masingi and brother, Abel Masingi.

Acknowledgments

I would like to thank the Almighty God who gave me the strength to pursue this study. I also thank my supervisor Dr. Daniel Maposa and co-supervisor Professor ’Maseka Lesaoana for their inspiration, competent guidance, patience and encouragement through all the different stages of this research.

Another thank you goes to all the postgraduate students in the Department, who have made my time as a student very enjoyable. Special thanks go to Thembhani Hlayisani Chavalala and Xidzuniso Baloyi who did the proofreading of parts of this dissertation. I also thank my best friend Smangele Maluleke for her unflinching support and encouragement through the study period. I would also like to acknowledge the South African Weather Service (SAWS) for providing me with the dataset.

Finally, I would like to thank my mother and siblings for their endless support and encouragement during my study period.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and background

1.1 Introduction and background

Rainfall is a principal element of water cycles and the variability of it is important from both the scientific as well as the socio-economic view. Hanum et al. (2015) stated that rainfall is an essential element of weather and normal rainfall is useful for life on the earth. Excessive rainfall is classified as extreme rainfall, which can be disastrous for life and infrastructure. According to Masereka et al. (2018), flood risks caused by extreme rainfall events have resulted in flood disasters that accounted for about 47% of all weather-related calamities, affecting 2.3 billion people worldwide. In the past decades, extreme precipitation occasions have made significant effect to properties, public infrastructure, agriculture, financial misfortunes just as financial issues and tourism in the Hawaiian Islands (Chu et al., 2009).

Muchuru et al. (2014) stated that Southern Africa is a region of significant rainfall variability and is disposed to serious events such as floods and droughts.

Recent increases in the frequency and intensity of extreme rainfall events have raised concern that human activities might have resulted in a change of the climate system (Syafrina et al., 2015). According to Mazvimavi (2010), there is a growing concern in Southern Africa about the declining rainfall patterns as a result of global warming. Extreme rainfall events are one of the primary natural causes of flooding (Alam et al., 2018). In February 2000, about 1,000 people were killed from severe floods caused by cyclone Eline that hit Mozambique and Zimbabwe (Rapolaki et al., 2019). Manhique et al. (2015) reported that the flood that occurred in January 2013 left almost 20,000 people homeless and about 100 dead in central and southern parts of Mozambique. According to Rapolaki and Reason (2018), in January 2015, tropical storm Chedza developed severe rainfalls that left more than 20,000 people homeless and 75 dead in Mozambique, southern Malawi and southern Madagascar.

South Africa is classified as a predominantly semi-arid country. This may be due to its variable geology, characterised by its atmosphere which ranges from desert and semi-desert in the dry northwestern region to sub-humid and wet along the eastern coastal area (du Plessis and Schloms, 2017). South Africa has nine provinces, namely: Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North-West and Western Cape. The present study will be carried out in the provinces of Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo and Mpumalanga.

According to Alexander (2018), South Africa's average annual rainfall is about 464 mm, which is below the world's average of 860 mm per year. Rainfall in South Africa exhibits seasonal variability, with most of the rainfall occurring mainly during summer months (South Africa Weather Service (SAWS), 2019). Botai et al. (2018) stated that annual rainfall in the northwestern region often remains below 200 mm, whereas much of the eastern highveld receives between 500 mm and 900 mm, occasionally exceeding 200 mm of rainfall per annum. The central parts of the country receive about 400 mm of rainfall per annum, with wide variations occurring closer to the coast.

According to Nash et al. (2016), KwaZulu-Natal is the wettest province of South Africa, with rainfall along the northeast coast exceeding 1,300 mm per annum, but declining to 800 mm per annum inland. Dyson (2009) stated that Gauteng province receives most of its rainfall in summer months, with the northwestern part of the province obtaining rainfall more frequently as compared to the south and south-east part of the province. A study conducted by Nel (2009) showed no significant trend, but increases in summer rainfall and decreases in autumn and winter rainfall in KwaZulu-Natal. Thomas et al. (2011) observed an increase in early-season rainfall and a decrease in late-season rainfall in north-west KwaZulu-Natal for the period 1950-2000. In the same study, Thomas et al. (2011) showed a tendency for a later seasonal rainfall onset accompanied by increased dry spells and fewer rain days in the Limpopo province. Rainfall variability in the Eastern Cape province causes water reduction in reservoirs (Pindura, 2016). Oduniyi (2013) highlighted that over the past decade in Mpumalanga province, there has been occurrence of climate change such as excessive temperature, fire outbreaks, rainfall and floods which caused a damage to agricultural productions. The Western Cape has been impacted by severe storms occurring almost annually over the past two decades, resulting in damages to homes, agricultural produces and infrastructure (Holloway et al., 2010).

1.2 Problem statement

According to Masereka et al. (2018), extreme high annual maximum daily rainfall events are among environmental events that have caused the most disastrous consequences for society. For example, tropical cyclone Idai made landfall on 14 March 2019 in the district of Dondo, Sofala province in Mozambique (UN Office for the Coordination of Humanitarian Affairs (UNOCHA, 2019)). The UNOCHA in Mozambique reported that at least 48 people died and in Zimbabwe 30 deaths and 100 missing people were reported (UNOCHA, 2019).

Research undertaken in South Africa's Kruger National Park has shown that some of the world's most sensitive and valuable riverine habitat are being destroyed due to increased frequency of cyclone-driven extreme floods (Floodlist, 2018). de Waal et al. (2017) stated that severe floods in the Western Cape of South Africa have caused significant damage to property and infrastructure over the past decade 2003–2014. According to Slabbert and Slatter (2019), tropical cyclone Idai in Mozambique plunged South Africa into phase 4 electricity load shedding. The economies of many African countries and the livelihoods of many of their people are exposed to risks associated with the impact of climate variability on agricultural yields since agriculture is heavily dependent on climate (Connaughton et al., 2017; Alam et al., 2011). An observed increase in extreme events includes increases in drought and it is expected that more people globally will be water stressed in the coming decades (Guilbert, 2016). According to Connaughton et al. (2017), extreme drought in particular can affect large numbers of people over an extended geographical area. Significant loss of life due to drought remains a real threat for millions of people. On the contrary, heavy rainfall at the end of the crop cycle causes damages to crops and financial losses to the farmers (Alam et al., 2011).

According to Singo et al. (2012), Luvuvhu River catchment is one of the regions in Limpopo province that has been negatively impacted by floods, resulting in loss of life and damage to public and private properties. A study conducted by Sauka (2016) revealed that the Crocodile River in the eastern part of Mpumalanga province has experienced three floods in a period of two years, which resulted in loss of life, damage of agricultural land and public properties. In Eastern Cape province, the Port Alfred floods in October 2012 left eight dead and caused the damage estimated at R500 million (Pyle and Jacobs, 2016). Dyson (2009) stated that rainfall resulting in flooding occurs from time to time over the Gauteng province, resulting in widespread flooding and disruption of infrastructure and even loss of life.

1.3 Rationale

Rainfall variability as a result of climate change and global warming has recently become an active area for studies in extreme value theory (EVT). Recent findings by de Waal et al. (2017) and de Waal (2012) suggest a change in the frequency of occurrence and intensity of extreme weather events, particularly rainfall, over the last two decades in the southern part of the Western Cape which resulted in marked damage to infrastructure, agriculture and human life. Extreme rainfall has become a common disaster in Southern Africa (Maposa et al., 2016). Due to its geographical position, South Africa is one of the countries that face challenges in terms of extreme rainfall (Kajambeu, 2016).

De Waal (2012) conducted a study on extreme rainfall distribution using generalised Pareto distribution (GPD) approach to assess changes in the frequency and intensity of extreme rainfall events across the Western Cape province over the historical records of 137 stations. The study revealed that of the 137 stations which were investigated, 62% showed an increase in 50-year return level, 22% showed a decrease in 50-year return level, while only 16% of the stations displayed little change in rainfall intensity over time. The results also indicated an increase in frequency of intense rainfall in the latter half of the 20^{th} century and early 21^{st} century. In a separate study, Hanum et al. (2015) modelled extreme rainfall using the Gamma-Pareto distribution to the monthly rainfall data from Jatiwangi station in Jakarta, Indonesia. The results showed that the Gamma-Pareto distribution was very appropriate for extreme monthly rainfall.

Kajambeu (2016) conducted a study on flood heights of the Limpopo River at Beitbridge Border Post using the generalised extreme value distribution (GEVD) in the presence of a trend covariate and Southern Oscillation Index (SOI). The study revealed the importance of considering non-stationary models when using statistics of extremes in a changing climate as these models provide an improvement in fit over the time-homogeneous models.

Mélice and Reason (2007) studied the return period of extreme rainfall at Saint George, South Africa, using EVT to assess the likely return period of such extreme rainfall. The study found that according to the Gumbel distribution family of EVT, the greatest annual maximum daily rainfall of 230 mm observed at the town in August 2006 had a return period of 1,222 years. The authors concluded that the August 2006 extreme rainfall at Saint George can be considered as a particularly rare event.

Singo et al. (2012) applied GEVD, Gumbel, log-normal and log-Pearson type III distributions in their study. The study used annual maximum flow data from 8 stations with 50 years hydrological data in Luvuvhu River catchments in Limpopo province of South Africa. The aim of the study was to analyse flood frequencies in the catchments. The extreme value analysis revealed that the

Gumbel and Log-Pearson type III were the best fit distributions. Kruger (2006) investigated the trends in daily extreme precipitation indices. The study employed data from SAWS for 138 rainfall stations in South Africa for the period 1910-2004. The author noted that there was an increasing trend in the number of extreme rainfall days in the Eastern Cape province, Southern Free State and parts of KwaZulu-Natal.

The present study will explore the use of non-stationary GEVD and GPD in a changing climate to model monthly rainfall of the five selected provinces of South Africa. This study will adopt the use of peaks-over-threshold distribution with time-varying covariates and thresholds to model monthly rainfall time series data since literature in these techniques is scarce.

1.3.1 Aim

The aim of this study is to model long-term monthly rainfall variability in the selected provinces of South Africa using extreme value distributions.

1.3.2 Objectives

The objectives of the study are to:

1. Model monthly rainfall using the parent distributions.
2. Investigate the long-term trends of the monthly rainfall and their variability across the selected provinces.
3. Use the non-stationary GEVD and non-stationary GPD with a fixed threshold to model monthly rainfall.
4. Model monthly rainfall time series data using a GPD with a time-varying threshold estimated by the non-parametric extremal mixture model.

1.4 Significance of the study

The outcome of this study will assist the country, particularly the government agencies of South Africa and outside South Africa, with extreme rainfall risk assessment using extreme value analysis techniques. In addition, the study will contribute to the body of knowledge in extreme value theory application to rainfall data and make recommendations to the government agencies on the long-term rainfall variability. The citizens of South Africa who live in the five provinces will also benefit from this study through taking the necessary measures to save their livestock and property during the rainy season. The need for modelling variability of rainfall in the selected provinces of South Africa is directed towards helping the agricultural and economic sectors, reduce the number of rainfall-related deaths and damages to infrastructure in these provinces.

1.5 Structure of the dissertation

The rest of the dissertation is organised as follows: **Chapter 2** presents a literature review with relevant methods and previous studies on rainfall. The methods used in the modelling of monthly rainfall across the selected provinces of South Africa in the study are presented in **Chapter 3**. The data analysis and research findings are presented and summarised in **Chapter 4**. This is followed by the conclusion of the study in **Chapter 5**. RStudio is used for data analysis and some RStudio codes used in this dissertation are presented in the appendix.

Chapter 2

Literature review

2.1 Introduction

This chapter reviews relevant studies in the modelling of rainfall that have been previously conducted in other studies as well as models applied by different researchers.

2.2 Rainfall modelled worldwide

Ender and Ma (2014) employed extreme value theory (EVT) including both generalised extreme value distribution (GEVD) and generalised Pareto distribution (GPD) to model extreme rainfall events using 60 years of daily data for four cities in China. The main finding was that GPD has a better fitting performance than GEVD. Zin et al. (2009) used annual series of maximum daily rainfall from 1975 to 2004 for 50 rain gauge stations in Peninsular Malaysia to investigate the best fitting distribution. They fitted five EVT distributions namely: GEVD, GPD, generalized logistic (GLD), log-normal (LN3) and Pearson (P3) distributions (Zin et al., 2009). The main finding of the study was that most stations in Peninsular Malaysia follow a GLD. The second most frequently selected distribution was the GEVD.

In India, Sharma and Singh (2010) conducted a study entitled: "Use of probability distribution in rainfall analysis". The aim of the study was to identify the best fit distribution. The study revealed that GEVD was the best fitting distribution. Recent studies by Namitha and Ravikumar (2018), and Namitha and Vinothkumar (2019) also found a similar result. However, the results obtained by Amin et al. (2016) in Pakistan found the LP3 distribution to be the best-fit distribution, contradicting the results obtained by Sharma and Singh (2010), Namitha and Ravikumar (2018), and Namitha and Vinothkumar (2019).

Gao et al. (2016) modelled annual maximum rainfall in China using both non-stationary and stationary GEVD. The authors found that the GEVD fits well to the data. The results were also confirmed by Syafrina et al. (2019) in Malaysia. Chu et al. (2013) applied non-stationary and stationary GEVD to model the annual maximum daily rainfall data for 18 stations in Taiwan for the period 1961-2010. Findings from the study showed that out of 18 stations, four stations were well fitted by the non-stationary model and the remaining stations were well fitted by the stationary model with the Gumbel distribution.

Roghani et al. (2016) conducted a study on the influence of Southern Oscillation on autumn rainfall in Iran. The study examined the relationship between Southern Oscillation Index (SOI) and autumn (October-December) rainfall covering the period of 1951-2011. The results showed that average SOI and SOI phase during July to September were related with October to December rainfall in some regions located in the west and northwest of Iran. Hanum et al.

(2017) assessed the application of modelling gamma-Pareto (G-P) distributed data using Generalised linear model (GLM) gamma for monthly rainfall observed in Sukadana station in Indonesia. The study sought to analyse whether Tropical Rainfall Measuring Mission (TRMM) satellite data is a good estimator for unobserved station's data. Transformed station's data was considered as dependent variable in GLM gamma. The findings of the study revealed that the station's data are G-P distributed and the transformed data are gamma distributed. The findings obtained by Aksoy (2000) and Husak et al. (2007) are not different from those obtained by Hanum et al. (2017).

2.3 Rainfall modelled in other countries in the African continent

Maposa (2019) utilised GEVD to annual flood heights time series models in examining suitable annual maximum moving sums that can be used to model extreme flood heights in the lower Limpopo River basin of Mozambique. Time series models were split into four parts, namely: annual daily maxima (AM1), annual maxima 2 days (AM2), annual maxima 5 days (AM5) and annual maxima 10 days (AM10). The study showed that models AM5 and AM10 were appropriate annual maxima time series models for Chokwe and Sicacate stations, respectively. In another study Maposa et al. (2014) did the comparative analysis of the maximum likelihood (ML) and Bayesian parameter estimates of the GEVD in the lower Limpopo River basin of Mozambique. The authors used Markov chain Monte Carlo (MCMC) Bayesian method to estimate the parameters of the GEVD in order to estimate extreme flood heights and their return periods. In their study, the Bayesian approach showed an improvement over the MLE approach.

Boudrissa et al. (2017) fitted the GEVD to model annual maximum daily rainfall for selected stations in the north of Algeria. The empirical results revealed that the Gumbel distribution fits well for Algiers and Miliana stations while the Fréchet distribution was found to be more suitable for the Oran station. Chikobvu and Chifurira (2015) modelled extreme minimum annual rainfall in Zimbabwe using the GEVD. Annual rainfall data from 1901 to 2009 were fitted to the GEVD. Results from model diagnostics showed that the minimum annual rainfall for Zimbabwe follows a Weibull class of distributions.

The study conducted by Olofintoye et al. (2009) examined the peak daily rainfall distribution characteristic in Nigeria. The study used annual rainfall data from 20 stations. Five probability distributions namely: Gumbel, log-Gumbel, LN3, P3 and log-Pearson (LP3) were fitted. The authors found that the LP3 distribution performed the best by occupying 50% of the total stations, followed by P3 with 40% of the total stations and lastly by log-Gumbel occupying 10% of the total stations. Another study on extreme rainfall was conducted in Tanzania by Ngailo et al. (2016) who found that the Gumbel distribution was the most suitable distribution to the extreme daily rainfall, while for the data above 99% the exponential distribution was found to be more appropriate.

2.4 Rainfall modelled in South Africa

Singo et al. (2016) studied evaluation of flood risks using flood frequency models in Luvuvhu River Catchment in Limpopo province, South Africa. The goal of the study was to estimate flood risks through rainfall distribution. The Gumbel and LP3 distributions were chosen to perform flood frequency analysis. The study revealed a general increase in the frequency of extreme events, accompanied by floods of higher magnitude. The study by Sauka (2016) used historic-climatic data for the Crocodile River to determine the critical threshold for past flood events and to predict the extent of future flood events in the Crocodile

River in the eastern part of Mpumalanga province. The statistics showed that when discharge reaches 241.75 m^3/s, both locations (Riverside and Tekwane) are at risk to flooding.

De Waal et al. (2017) employed a GPD and peaks-over-threshold (POT) sampling approach to the 76 rainfall stations across the Western Cape province in South Africa. The study sought to determine the changes in extreme 1-day rainfall high percentiles (95th and 98th) and both the 20- and 50- year return period rainfall, comparing the period 1950-1979 against that of 1980-2009. Of these stations, 48 (63%) showed an increase in the 50-year return period of extreme 1-day rainfall and 28 (37%) showed a decrease in the 1980-2009 period at the 95th percentiles POT, while at the 98th percentiles POT, 49 (64%) showed an increase and 27 (36%) a decrease for the later period.

Dyson (2009) modelled daily rainfall over the Gauteng province in South Africa for the summer months of October to March using 32 years (1977-2009) daily rainfall data from about 70 South African Weather Service stations. The results revealed that the month with the highest monthly average rainfall as well as the highest number of heavy and very heavy rainfall days was January, followed by February, March and lastly October. The conclusion of the study was that significantly high seasonal rainfall is associated with the above-normal rainfall in late summer.

Masereka et al. (2018) conducted a study on the statistical analysis of annual maximum daily rainfall (ADMR) for Nelspruit in Mpumalanga province of South Africa. Empirical continuous probability distribution functions (ECPDF) and theoretical continuous probability distribution functions (TCPDF) were applied to carry out the statistical analysis of the extreme high ADMR events. Findings of empirical frequency analysis revealed that the return period of flood disasters was 10 years. Mzezewa et al. (2010) modelled rainfall data collected over the period 1983-2005 to study the basic statistical rainfall characteristics at the University of Venda ecotope. Annual and monthly rainfall was fitted to theoretical probability distributions. Furthermore, the Anderson-Darling goodness-of-fit test was used to select the best fit models. The authors found that the distribution of daily rainfall was highly skewed, with high frequency of occurrence of low-rainfall events.

2.5 Rainfall trends worldwide

Acero et al. (2011) applied POT to study the trends in extreme rainfall over the Iberian Peninsula at a daily scale. The study used data from 52 observations regularly distributed over Iberia with no missing data available for the period 1958-2004. The results indicated a high variability of extreme events over the coastline of the Iberian Peninsula. In another study, Acero et al. (2012) used non-parametric Mann–Kendall (M-K) test statistic and parametric test based on the statistical theory of extreme values, involving time-dependent parameters to account for possible temporal changes in the frequency distribution. The scholars found that, in winter, there were significant negative trends for a greater part of the Iberian Peninsula, while significant positive trends were found for the southeast, particularly over areas that shrank as the number of days considered for the precipitation event increased.

Wi et al. (2016) fitted non-stationary GEVD and non-stationary GPD models to annual maximum precipitation (AMP) and POT, respectively, of 65 weather stations scattered across South Korea. The M-K test statistic was used for trend detection and the results showed an increasing trend in AMP to the stations concentrated in the mountainous areas of South Korea. The results from GPD indicated fairly different results, with a significantly reduced number of stations showing an increasing trend, while some stations showed a decreasing trend.

In a slightly different study, Bharti (2015) employed remotely sensed TRMM 3B42 version 7 precipitation data to investigate extreme rainfall events during the monsoon season over the Northwest Himalaya for the period 1998-2013. Three percentiles: 98th, 99th and 99.99th were used to identify the extreme rainfall index. Findings from the study revealed that rainfall intensities associated with these three percentiles for each pixel showed higher rainfall intensity for the regions with less than 3,000 m elevation. The study also revealed an increasing trend of heavy and very heavy rainfall intensity events over the region. Furthermore, the authors showed that the plains and foothills of Northwest Himalaya with elevation less than 500 m receive higher number of extreme events.

2.6 Rainfall trends for some other African countries

Chikodzi et al. (2013) used time series analysis to investigate trends in the southeastern region of Zimbabwe. The study used climate data records from three Zimbabwe Meteorological Services Department run weather stations in the region. The findings from the study revealed a significant decline in rainfall at two of the three stations used. Mazvimavi (2008) fitted non-parametric test to determine possible changes in extreme annual rainfall in Zimbabwe. The M-K test statistic was applied to investigate whether rainfall data from 40 rainfall stations in Zimbabwe differ with time. Findings from the study showed insignificant trend with time for the data. This was also confirmed by Mazvimavi (2010). Mosase and Ahiablame (2018) used daily rainfall, and maximum and minimum temperature gridded data from the Climate Forecast System Reanalysis (CFSR) global weather database to investigate the long-term trend of the Limpopo River basin. Modified non-parametric M-K test statistic was used to check the long-term trend. Trend analysis showed upward trends for both annual and season rainfall in most parts of the basin, except for the winter season which showed a decreasing trend.

2.7 Rainfall trends in South Africa

Gyamfi et al. (2016) used historical rainfall records from 13 stations of Olifants basin in South Africa. The historial rainfall records were obtained from South Africa Weather Service (SAWS) and Department of Water Affairs (DWA) spanning the period 1975-2013. M-K test statistic trend was applied to detect changes in rainfall pattens under a changing climate. Results of the study indicated an insignifiant declining rainfall trend in the Olifants basin with a mean annual rainfall of 664 mm. Kruger and Nxumalo (2017) conducted a study about the historical rainfall trends in South Africa. The study used two interlinked datasets namely: the district rainfall and individual rainfall stations covering the period 1921-2015. The authors found that there was an increase in rainfall over the west of South Africa, particularly in the southern interior and decreases in rainfall in some places in the far north-east for the period 1921-2015.

MacKellar et al. (2014) found statistically significant decreases in rainfall over the central and north-eastern parts of the country in the autumn months and significant increases in the southern Drakensberg in spring and summer seasons. Easterling et al. (2000) showed a significant increase in heavy rainfall frequency over the threshold of 25.4 mm and 50.8 mm for the Western Cape and parts of KwaZulu-Natal provinces, respectively. Botai et al. (2018) investigated the spatial-temporary variability and trends of precipitation concentration across South Africa using TRMM 3B42 version 7 satellite precipitation data sets spanning 1998-2015. The results indicated that precipitation concentration across South Africa exhibits noticeable spatial-temporary variability. The authors concluded that findings from this study have important scientific and practical applications in hydrological hazard risk and soil erosion monitoring. Odiyo et al. (2015) investigated long-term changes and variability in daily rainfall and streamflow in the Luvuvhu River catchments in the Limpopo province, South Africa. The study applied linear regression method to compute trend in 5- and 10-year average rainfall. The authors further applied linear regression and M-K test statistic to detect trends for annual rainfall and streamflow data. The study showed a decreasing trend in 5- and 10-year mean rainfall. Results from linear regression and M-K test statistic are not different from those based on 5- and 10-year streamflow.

2.8 Extreme value theory and its applications

Extreme value theory (EVT) can be applied in fields where extreme events may occur, including climate change, insurance, finance and public health (Bensalah, 2000).

Sigauke and Bere (2017) carried out a study in South Africa based on the application of the GPD to the modelling of daily peak electricity demand. The POT approach with time varying covariates and threshold were used to model non-stationary time series. In their study, the GPD model showed a better fit to the data than the GEVD model (Sigauke and Bere, 2017). The scholars concluded that peak electricity demand is a major concern for utility companies. A study carried out by Diriba et al. (2015) modelled extreme daily temperature using GPD at Port Elizabeth, South Africa. The outcome of the study based on the return levels analysis showed that by the end of the 21st century, the extreme summer maximum temperature could be around 5°C more than the current one, while in winter the return level analysis indicated an increase of about 2°C.

Thomas et al. (2016) conducted a study on the application of EVT in public health. The study applied EVT to weekly rates of pneumonia and influenza deaths over 1979-2011 in Canada. Their findings revealed that an annual pneumonia and influenza death rate of 12 per 100,000 (the highest maximum observed) should be exceeded once over the next 30 years, and each year there should be a 3% risk that the pneumonia and influenza death rate will exceed this value. In their study, Jakata and Chikobvu (2019) modelled extreme risk of the South African financial index (J580) using GPD. The study showed that the upside risk of the financial index (J580) outweighs the downside risk.

2.9 Concluding remarks

This chapter has reviewed research undertaken by other researchers on rainfall, finance, public health, wind and temperature in various countries worldwide including South Africa. Models applied by different researchers have also been reviewed in this chapter.

Previous literature looked at modeliing climate data using stationary GEVD and stationary GPD. The present syudy will use similar methods applied by Gao et al. (2016) and Wi et al. (2016), to model monthly rainfall data for selected provinces of South Africa.

Chapter 3

Research methodology

3.1 Introduction

This chapter presents statistical and graphical tests used to analyse the monthly rainfall data for selected provinces of South Africa.

3.2 Data source and study area

Monthly rainfall data for Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo and Mpumalanga provinces were obtained from South Africa Weather Service (SAWS) and is a time series secondary data measured in millimeters (mm). The monthly rainfall data for the five selected provinces are summarised in Table 3.1.

Table 3.1: Summary of rainfall data for selected provinces of South Africa.

Provinces	Starting year	Ending year
Eastern Cape	1900	2017
Gauteng	1900	2017
KwaZulu-Natal	1900	2017
Limpopo	1904	2017
Mpumalanga	1904	2017

3.3 Parent distributions

Probability distributions are basic concepts in statistics (Amin et al., 2016). Monthly rainfall data for selected provinces of South Africa were assessed with five parent distributions in order to identify the appropriate distributions. The probability models explored include log-normal, Gumbel, Weibull, gamma and Pareto distributions.

3.3.1 Log-normal distribution

The log-normal distribution is a distribution of random variables with a normally distributed logarithm. The log-normal distribution model includes a random variable Y, and $\log$(Y) is normally distributed (Amin et al., 2016). The probability density function (PDF) and cumulative distribution are calculated using (3.1) and (3.2), respectively

$$f(x) = \frac{\exp\left[-\frac{1}{2}\,(\frac{\mathbf{ln}(x-\gamma)-\mu}{\sigma})^2\right]}{(x-\gamma)\sigma\sqrt{2\pi}}, \tag{3.1}$$

$$F(x) = \phi\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right) = \frac{1}{2}\left[\mathbf{erfc}\{-\frac{\ln(x-\gamma)-\mu}{\sigma\sqrt{2}}\}\right], \tag{3.2}$$

where μ is the shape parameter, σ is the scale parameter, γ is the location parameter and ϕ is the Laplace integral and erfc is the error function.

3.3.2 Gumbel distribution

According to Amin et al. (2016) the Gumbel distribution also referred to as the extreme value type I distribution has two forms, one is based on the smallest extreme (minimum case), and the other is based on the extreme (maximum case). In this study , the maximum case is employed and its PDF and CDF are given as:

$$f(x) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma} - \exp\left(-\frac{x-\mu}{\sigma}\right)\right), \tag{3.3}$$

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), \tag{3.4}$$

where σ and μ are the scale and location parameters, respectively.

3.3.3 Weibull distribution

The Weibull distribution is a two-parameter distribution with parameters α and β. Alam et al. (2018) mentioned Weibull as a commonly used frequency distribution in hydrology. The PDF and CDF for two-parameter Weibull distribution are given as:

$$f(x) = \frac{\alpha}{\beta}\left(\frac{\alpha}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \tag{3.5}$$

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\alpha}\right)\right], \tag{3.6}$$

where α is the shape parameter ($\alpha > 0$) and β is a scale parameter ($\beta > 0$).

3.3.4 Pareto distribution

The Pareto distribution is a continuous distribution with the following PDF and CDF, respectively

$$f(x;\alpha,\beta) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \tag{3.7}$$

$$F(x) = 1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha}, \tag{3.8}$$

where α is a shape parameter and β is scale parameter.

3.3.5 Gamma distribution

The gamma distribution function consists of three different types, 1-, 2-, 3-parameter gamma distribution (Aksoy, 2000). If the continuous random variable x fits to the PDF function of

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}; x \geqslant 0, \tag{3.9}$$

it is said that the variable x is 1-parameter gamma distributed, with the shape parameter α. In equation (3.9), $\Gamma(\alpha)$, the incomplete gamma function, is given by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx. \tag{3.10}$$

The distribution function has a form of the simple exponential distribution in the case of α = 1. If x in equation (3.9) is replaced by $\frac{x}{\beta}$, where β is the scale parameter, then the 2-parameter gamma distribution is obtained as

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}; x \geqslant 0, \tag{3.11}$$

which returns to the 1-parameter gamma distribution for β = 1. If x is replaced by $\frac{(x-\gamma)}{\beta}$, where γ is the location parameter, then the 3-parameter gamma distribution is obtained by

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} (x-\gamma)^{\alpha-1} e^{-\dfrac{x-\gamma}{\beta}}; x \geqslant \gamma. \tag{3.12}$$

3.4 Model selection

According to Acquah (2012), the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are widely used as criteria of model selection tool in many problems. AIC and BIC are employed in this study to find the best-fit family of probability distribution. The probability distribution with the lowest AIC and BIC will be considered as the best-fit probability distribution (Thiombiano et al., 2017; Mandal and Choudhury, 2015; Katz, 2013).

3.4.1 Akaike's information criterion (AIC)

AIC is one of the most commonly used information criteria. The idea of AIC is to select the model that minimises the negative likelihood penalised by the number of parameters as specified in the following equation:

$$\mathrm{AIC} = -2\log(L) + 2p, \tag{3.13}$$

where L refers to the likelihood under the fitted model and p is the number of parameters in the model (Acquah, 2012). Specifically, AIC is aimed at finding the best approximating model to the unknown true data generating process and its applications (Acquah, 2010)

3.4.2 Bayesian information criterion (BIC)

BIC is another widely used information criterion. BIC is usually explained in terms of the Bayesian theory, especially as an estimate of the Bayes factor for comparison of a model to the saturated model (Acquah, 2012). BIC is defined as:

$$\mathrm{BIC} = -2\log(L) + p\log(n), \tag{3.14}$$

where n is the sample size and L is the likelihood under the fitted model. BIC is designed to find the true model where the true model is assumed to be among the models being compared (Acquah, 2012).

3.5 Parameter estimation using maximum likelihood estimation method

The parameters of the probability distributions are estimated using the maximum likelihood estimator (MLE) method. Maximum likelihood turns out to be a widely applicable method that yields good estimates when sample sizes are large: maximum likelihood estimators are consistent, asymptotically normal, and asymptotically efficient (Pan and Fang, 2002). According to Jakata and Chikobvu (2019) and Chege et al. (2016), the joint density function of a sample size n that is independent and identically distributed (iid) is given as follows:

$$L(\theta|x_1, x_2, ..., x_n|\theta) = f(x_1|\theta)f(x_2|\theta)...f(x_n|\theta), \tag{3.15}$$

where θ are the parameters of the model and x_i are the observed variables. Thus the observed variables, x_i are known whereas the parameters given by θ are to be estimated. The estimated likelihood function is then given by

$$L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{3.16}$$

and the logarithmic likelihood function is given by the following:

$$\ln L(\theta|x_1, x_2, ..., x_n) = \sum_{i=1}^{n} f(x_i|\theta). \tag{3.17}$$

The estimated parameters are then given by the set which maximizes the likelihood function in (3.16) or (3.17).

3.6 Test for stationarity

Statistical theory offers a wide range of unit root test, with the most commonly used being augumented Dickey-Fuller (ADF) test, Phillips-Perron (PP) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Fedorová, 2016). However, in this study ADF, PP and KPSS are used to test whether the monthly rainfall data for selected provinces of South Africa are stationary.

3.6.1 Augmented Dickey-Fuller (ADF) test

The ADF test was employed in this study to check whether the monthly rainfall data for selected provinces of South Africa are stationary.

The ADF test is estimated under the following hypothesis:
H_0: there exists a unit root and the time series is non-stationary.
H_1: time series is stationary.

The ADF test consists of estimating the following regression model:

$$y_t = \beta + \beta_1 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \triangle Y_{t-1} + \epsilon_t, \tag{3.18}$$

where β is a constant, β_1 is the coefficient on time trend. The null hypothesis is $\delta = 1$, and the alternative hypothesis is $\delta \neq 1$, while ϵ_t is a pure white noise error term and the ADF follows an asymptotic distribution (Paparoditis and Politis, 2013).

3.6.2 Phillips-Perron (PP) Unit Root Tests

The Phillips-Perron (Liolios, 2015; Phillips and Perron, 1988) test is a more developed test, introduced in 1988 and it has the same null hypothesis with ADF tests and also uses the same critical values with it. The PP test makes a non-parametric correction to the t-statistic. The PP test involves the equation coming from Dickey-Fuller test:

$$\triangle Y_t = \mu + v + \lambda_t + \epsilon_t, \tag{3.19}$$

where ϵ_t is I(0) and it can be heteroscedastic. For this reason, the test estimates the equation:

$$y_t = y_{t-1} + v + \lambda_t + \epsilon_t. \tag{3.20}$$

The PP method estimates the non-augmented DF test equation and modifies the t-ratio of the coefficient, so that serial correlation does not affect the asymptotic distribution of the test statistic. The PP test is based on the statistic:

$$\bar{t}_\mu = t_\mu \left(\frac{\gamma_0}{f_0}\right)^{\frac{1}{2}} - \frac{T(f_0 - \gamma_0[se(\mu)]}{2f_0^{\frac{1}{2}} s}. \tag{3.21}$$

The PP test is estimated under the following hypothesis:
H_0: there is a unit root.
H_1: there is no unit root.

3.6.3 Kwiatkowski-Phillips-Schmidt-Shin (KPSS)

The present study employed KPSS unit root test model proposed by Poulos (2016) and Hobijn et al. (2004). This method assumes that a time series y_t can be decomposed as:

$$y_t = \zeta + r_t + \epsilon_t, \tag{3.22}$$

where ζ is a deterministic trend, r_t is a random walk and ϵ_t is a stationary error. The random walk component is expressed by the following equation:

$$r_t = r_{t-1} + u_t, \tag{3.23}$$

where

u_t is a random variable with mean 0 and variance σ_u^2.

If $\sigma_u^2 = 0$, the null hypothesis of stationary series is true.

If $\sigma_u^2 = 0$, and $\zeta = 0$, then the series is stationary about the value r_0.

If $\sigma_u^2 = 0$, and $\zeta \neq 0$, then the series is stationary about a trend.

The KPSS test is estimated under the following hypothesis:

H_0: Series does not have a unit root test or is stationary

H_1: Series has a unit root or is not stationary.

3.7 Trend test

This study used non-parametric Mann-Kendall (M-K) test statistic and time series plot to investigate the long-term trend of the monthly rainfall and their variability across the selected provinces.

3.7.1 Non-parametric Mann-Kendall (M-K) test statistic

Non-parametric M-K test statistic is frequently used to quantify the significance of monotonic trend in hydrometeorological time series (Wi et al., 2016; Da Silva et al., 2015). The Mann-Kendall test statistic is defined as

$$S = \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} \textbf{sgn}(e_i - e_j), \tag{3.24}$$

where n is the number of extreme values. If S is positive, then there is an increasing trend, but if S is negative, then there is a decreasing trend, and sgn$(e_i - e_j)$ is a sign function given by:

$$\text{sgn}(e_i - e_j) = \begin{cases} 1, & \text{if } e_i - e_j > 0, \\ 0, & \text{if } e_i - e_j = 0, \\ -1. & \text{if } e_i - e_j < 0. \end{cases} \tag{3.25}$$

Under the null hypothesis of no trend, the theoretical mean of S is 0 and its variance is given by

$$Var(S) = \left[n(n-1)(2n+5) - \sum_{p=1}^{g} t_p(t_p - 1)(2t_p + 5) \right] /18, \tag{3.26}$$

where g is the number of tied groups (a tied group is a set of sample data having the same value), and t_p is the number of data points in the p$^{\text{th}}$ tied group. If no tied group exist, this process can be ignored (Da Silva et al., 2015). In cases where the sample size $n > 30$, the normalised test statistic Z can be used to statistically quantify the significance of the trend. Z is calculated using the following equation:

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}, & \text{if } S > 0, \\ 0, & \text{if } S = 0, \\ \frac{S+1}{\sqrt{Var(S)}}, & \text{if } S < 0. \end{cases} \tag{3.27}$$

Positive values of Z indicate an increasing trend, while negative Z values show decreasing trends. In a one-tailed test at a significance level of α, the null hypothesis of no trend is rejected if $\mid Z \mid > z_\alpha$, where z is the standard normal variable. In this study, the significance level was set to be 5%.

3.7.2 Time series plots

A time series plot is simply a graph in which the data values are arranged sequentially in time. It is commonly used to give a pictorial view of the data series over time. time series plots and other plots such probability, quantile, return level, and density will be used as part of exploratory data analysis.

3.8 Test for normality

According to Adefisoye et al. (2016), there are several parametric and non-parametric methods of assessing whether data are normally distributed or not. They are split into two groups: graphical and statistical. The most frequently used techniques include: Quantile-Quantile (Q-Q) plots, cumulative, probability-probability (P-P) plots, Anderson–Darling test (AD), Shapiro–Wilk (SW) test, D'Agostino-Pearson K2 (DPK) test, chi-square test, Jarque-Bera (JB) test, kurtosis test, Shapiro-Francia (SF), skewness test, robust Jarque-Bera (RJB) test. In this study only JB, SW and chi-square methods are employed to check whether the monthly rainfall data are normally distributed or not. The SW test is one of the most popular test for normality assumption diagnostics which has good properties of power and it based on cerrelation withing given observations and associated normal scores (Das and Imon, 2016). Wuertz and Katzgraber (2005) and Adefisoye et al. (2016) stated that JB and chi-square test are likely most widely used procedure for testing normality.

3.8.1 Jarque-Bera (JB) test

The JB test statistic is expressed as:

$$JB = n\left(\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2-3)^2}{24}\right), \tag{3.28}$$

where $\sqrt{b_1}$ and b_2 are the skewness and kurtosis measures and are given by $\frac{m_3}{(m_2)^{3/2}}$ and $\frac{m_4}{(m_2)^3}$, respectively; and m_2, m_3 and m_4, are second, third and fourth central moments, respectively. The JB test statistic is chi-square distributed with two degrees of freedom.

The hypothesis test for the JB test procedure is
H_0: The monthly rainfall data is normally distributed, versus, H_1: The monthy rainfall data do not come from a normal distribution.

3.8.2 Shapiro–Wilk test (SW)

The SW test is of the form:

$$W = \frac{1}{D}\left[\sum_{i=1}^{m} a_i(x_{(n-i+1)} - x_{(i)}\right]^2, \tag{3.29}$$

where $m = \frac{n}{2}$ if n is even, while $m = \frac{(n-1)}{2}$ if n is odd. $D = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $x_{(i)}$ represent the i^{th} order statistic of a sample, the constants a_i are given by: $(a_1, a_2, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$ and m is given by $m = (m_1, m_2, ..., m_n)^T$ where $m_1, m_2, ..., m_n$ are the expected values of order statistics of independent and identically distributed (iid) random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics (Adefisoye et al., 2016).

The S-W test is estimated under the following hypothesis:
H_0: The monthly rainfall data is normally distributed
H_1: The monthly rainfall data does not come from a normal distribution.

3.8.3 Chi-square test

The chi-square goodness-of-fit is defined as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{3.30}$$

where (O_i) and (E_i) refer to the i^{th} observed and expected frequencies, respectively, and n is the number of groups. When the null hypothesis is true, the above test statistic follows a chi-square distribution with $k-1$ degrees of freedom (Adefisoye et al., 2016).

The chi-square test is estimated under the following hypothesis:
H_0: The monthly rainfall data are sampled from a normal distribution
H_1: The monthly rainfall data are not sampled from a normal distribution.

3.9 Extreme value theory techniques

In extreme value theory (EVT) two approaches exist: the block maxima (BM) and the peaks-over-threshold (POT) methods. According to Ferreira and De Haan (2015), the BM approach is an approach in EVT that consists of dividing the observation period into non-overlapping periods of equal sizes. Kajambeu (2016) and Coles et al. (2001) defined the POT as the method whereby the peak values from a continuous record for any period during which values exceed a certain threshold are extracted. In this study the BM in a changing climate and POT with time varying covariates and thresholds are utilised to model monthly rainfall of the five selected provinces of South Africa.

3.9.1 Stationary generalised extreme value distribution

Generalised extreme value distribution (GEVD) is the family of asymptotic distributions that describes the behaviour of extreme conditions. The GEVD consists of three extreme value distributions namely: Gumbel, Fréchet and Weibull families which are also referred to as type I, II and III extreme value distributions (Syafrina et al., 2019; Ngailo et al., 2016; Coles et al., 2001). The cumulative probability distribution for GEVD is of the form:

$$GEVD(x,\mu,\sigma,\xi) = \begin{cases} \exp - \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}; \xi \neq 0, \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right); \xi = 0, \end{cases} \tag{3.31}$$

where x are the extreme values from the blocks, μ, σ and ξ are the location, scale and shape parameters, respectively. For $\xi > 0$, we obtain the Fréchet distribution, for $\xi = 0$, we get the Gumbel distribution and for $\xi < 0$, we get the Weibull distribution.

3.9.2 Non-stationary generalised extreme value distribution

The non-stationary GEVD model is the fundamental modification of the stationary GEVD model (Syafrina et al., 2019). To account for non-stationary GEVD the location parameter μ and the scale parameter σ are assumed to vary with time t and possibly other covariates (Hundecha et al., 2008; Coles et al., 2001). The non-stationary GEVD is given by:

$$F(x;\mu(t),\sigma(t),\xi(t) = \exp - \left[1+\xi\frac{x-\mu(t)}{\sigma(t)}\right]^{-\frac{1}{\xi(t)}}. \tag{3.32}$$

In the simplest case, the following regression structures could be examined for the location and scale parameters:

$$\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2, \tag{3.33}$$

$$\sigma(t) = \exp(\sigma_0 + \sigma_1 t + \sigma_2 t^2, \xi(t) = \xi \tag{3.34}$$

allowing up to quadratic dependence on time t (Panagoulia et al., 2014).

3.9.3 Parameter estimation of non-stationary GEVD

Parameters of the non-stationary GEVD are estimated using the method of maximum likelihood (ML).

Maximum likelihood estimation method

For a sample of N observations, the ML of the time-dependent GEVD in (3.32) was determined by maximising the log-likelihood function, expressed with time-varying parameters:

$$\begin{aligned} l(\mu(t), \sigma(t), \xi) = -\sum_{t=1}^{N} \log \sigma(t) + \left(1 + \frac{1}{\xi}\right) \log \left[1 + \xi \left(\frac{x_i - \mu(t)}{\sigma(t)}\right)\right] \\ + \left[1 + \xi \left(\frac{x_i - \mu(t)}{\sigma(t)}\right)\right]^{-1/\xi}, \end{aligned} \tag{3.35}$$

where N is the number of years of observation. To obtain the GEVD parameter estimators that maximise equation (3.35) we used the interior algorithm based nonlinear optimisation as implemented in the MATLAB Optimisation Toolbox (Wi et al., 2016).

3.10 Goodness-of-fit

Goodness-of-fit test statistics are used for checking the validity of a specified or assumed probability distribution model. In this study, Kolmogorov-Smirnov (K-S) test, the Anderson-Darling (A-D) and graphical methods, were applied to identify the best model.

3.10.1 Kolmogorov-Smirnov (K-S) test

The K-S test, based on the empirical cumulative distribution function is used to decide if a sample comes from a hypothesised continuous distribution (Alam et al., 2018; Chikobvu and Chifurira, 2015; Sharma and Singh, 2010). The K-S statistic D is defined as the largest vertical difference between theoretical and the empirical cumulative distribution (ECDF) and is formulated as follows:

$$D_{max} = \max_{1 \leq i \leq n} \left(F(x_i) - \frac{i-1}{n}; \frac{i}{n} - F(x_i) \right), \tag{3.36}$$

where X_i are random samples, $i = 1, 2, ..., n$, and the CDF is

$$F_n(x) = \frac{1}{n} \left[Number\ of\ observations\ \ \leq x \right]. \tag{3.37}$$

The K-S test is estimated under the following hypothesis:
H_0: The monthly rainfall data follow a specified distribution
H_1: The monthly rainfall data do not follow the specified distribution.

3.10.2 Anderson-Darling (A-D)

The A-D test statistic (A^2) is defined as:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left[\ln F(X_i) + \ln(1 - F(X_{n-1+1})) \right]. \tag{3.38}$$

The A-D test is used to compare the fit of an observed CDF to an expected CDF. This test gives more weight to the tails of the distribution than the K-S test (Chikobvu and Chifurira, 2015; Sharma and Singh, 2010).

The A-D test is estimated under the following hypothesis:
H_0: The monthly rainfall data follow a specified distribution
H_1: The monthly rainfall data do not follow the specified distribution.

3.10.3 Graphical test

Alam et al. (2018) stated that graphical test is one of the most simple powerful techniques for selecting the best-fit model. To check if the time-dependent GEVD and GPD fit well to the monthly rainfall data, the following graphical tests were used.

Quantile-quantile (Q-Q) plots

Quantile-quantile (Q-Q) plot, is a comparison of an empirical form for estimating the exceedance and the inverse of fitted distribution function. Any departure from linearity indicates model failure in perfectly fitting the data (Iyamuremye et al., 2019; Alam et al., 2018).

Probability-probability (P-P) plots

Probability-probability (P-P) plot is a comparison of an empirical (usually percentage rank) and the fitted distribution function. In case of perfect fit, the data would line up on the diagonal of the probability plots (Iyamuremye et al., 2019; Alam et al., 2018).

Return level plots

In these plots the empirical estimates of the return level functions are added. If there is an agreement between the model-based curve and empirical estimates, then the model is suitable for the data (Iyamuremye et al., 2019; Alam et al., 2018).

3.10.4 Choice of preferred model

When time-dependent GEVD and GPD are considered with covariates, there are a number of possible models to select from (Osman et al., 2015). In order to select between model fits, a test of the likelihood ratio test also known as the deviance (D) statistic is used. For models $M_0 \subset M_i$, we define the D statistic as:

$$D = 2\{l_i(M_i) - l_0(M_0)\}, i = 1, 2, 3, ... \tag{3.39}$$

where $l_0(M_0)$ and $l_i(M_i)$ are the maximised log-likelihood under models M_0 and M_i respectively. The asymptotic distribution of D is given by χ_k^2 distribution with k degrees of freedom, where k is the difference in dimensionality of M_1 and M_0 . The calculated deviance statistic, D, is compared to critical values from χ_k^2 at α level of significance. Large values of D suggest that M_1 explains substantially more of the variation in the data than M_0 (Kajambeu, 2016; Osman et al., 2015; Coles et al., 2001).

3.11 Stationary generalised Pareto distribution (GPD)

The generalised Pareto distribution (GPD) is a peaks-over-threshold (POT) distribution which considers the maximum values exceeding a pre-determined threshold which is assumed to approximately follow a GPD (Maposa et al., 2014). Let X be a random variable. The CDF of the GPD (ξ, μ, σ) with shape parameter ξ, location parameter μ, and scale parameter σ is given by:

$$G_{\xi,\mu,\sigma}(x)) = \begin{cases} 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{\frac{-1}{\xi}}, \ \xi \neq 0 \\ 1 - \exp\left(-\frac{x-\mu}{\sigma}\right), \ \xi = 0. \end{cases} \tag{3.40}$$

For $\xi \geq 0$, the range is $\mu \leq x < \infty$, and for $\xi < 0$, $\mu < x < \mu - \sigma/\xi$. When $\xi = 0$, the GPD is the exponential distribution (Zhao et al., 2019; Pickands III, 1975).

3.11.1 Peaks-over-threshold with time-varying covariates

Let Y_t be a process with associated time-varying covariates X_t, for $t = 1, ..., n$, where n is the number of observations and let $\tau(t)$ be the time varying threshold. The extreme quantiles of Y_t are denoted by y_p and are conditional on the covariates X_t, and $P(Y_t > y_p | X_t = x_t)$ is the tail probability above the quantile y_p. On average, the high quantile y_p is exceeded approximately once every $\frac{1}{p}$ (Sigauke and Bere, 2017; Eastoe and Tawn, 2009). The observations y_t above $\tau(t)$ are assumed to follow a GPD, that is,

$$Y_t \sim \text{GPD}(\sigma(x_t), \xi(x_t)), \tag{3.41}$$

where $\sigma(x_t)$ is the scale parameter and $\xi(x_t)$ is the shape parameter depending on the time-varying covariates, X_t. The CDF is composed as follows:

$$G(y_t) = 1 - \left(1 + \frac{\xi(x_t)(y_t - \tau(t))}{\sigma(x_t)}\right)^{-\frac{1}{\xi(x_t)}}, \tag{3.42}$$

where $\xi(x_t) \neq 0$, y_t is the monthly rainfall data and the parameters are modelled as a function of the covariates.

3.11.2 Time-varying threshold

From the previous section we let $\tau(t)$ to be our time-varying threshold. Sigauke and Bere (2017) defined $\tau(t)$ as a penalised cubic smoothing spline with a positive shift factor and the function is given as:

$$\tau(t) = \sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int (f^{''}(t))^2 dt + u, \tag{3.43}$$

where y_i denotes our monthly rainfall data, λ is a smoothing parameter and $u \in \mathcal{R}$ is a shift factor which should be large enough to allow asymptotic condition to be satisfied when we fit the GPD. In this study, extremal mixture models were adopted to estimate the positive shift factor u.

3.11.3 Declustering

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

$$\eta_u = \frac{2\left[\sum_{n-1}^{N-1}(T_i = 1)\right]^2}{(N-1)\sum_{i=1}^{N-1}(T_i - 1)(T_i - 2)}, \tag{3.44}$$

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

3.11.4 Extremal mixture model

This section outlines the extremal mixture models in which the threshold is used as a parameter to be estimated together with the GPD parameters (Scarrott and MacDonald, 2012). Mixture models are divided into three parts, namely: parametric, semi-parametric and non-parametric. In this study, the non-parametric mixture models, were employed. According to Scarrott and MacDonald (2012) the major benift of the non-parametric approaches as compared to the parametric approaches is that the tail fit is robust to the bulk fit. In non-parametric mixture models the observations below the threshold are assumed to follow a non-parametric density $h(.|\lambda, \mathbf{X})$, which is dependent on parameter λ and the observation vector X. The excesses above the threshold are assumed to follow a GPD(σ_u, ξ).

Suppose the monthly rainfall data consist of a sequence of n iid observations

$X = x_i; i = 1, ..., n$, with distribution function F defined by

$$F(x|\lambda, u, \sigma_u, \xi, \mathbf{X}) = \begin{cases} (1-\phi_u)\frac{H(x|\lambda,\mathbf{X})}{H(u|\lambda,\mathbf{X})}, & \text{if } x \leq u, \\ (1-\phi_u) + \phi_u G(x|u, \sigma_u, \xi), & \text{if } x > u, \end{cases} \tag{3.45}$$

where $\phi_u G(x|u, \sigma_u, \xi)$ is the unconditional GPD function given by (3.41) (Scarrott and MacDonald, 2012).

Chapter 4

Data analysis and discussion

4.1 Introduction

Chapter 4 presents the analyses and interpretation of results using methods that were discussed in Chapter 3. This chapter is divided into various sections: descriptive statistics, stationarity tests, normality tests, parent distribution selection, trend analysis and model fitting.

4.2 Descriptive statistics

The descriptive statistics evaluated are the mean, standard deviation, median, kurtosis, skewness, minimum and the maximum monthly rainfall amount for each province. The summary of the descriptive statistics for each province is shown in Table 4.1.

From Table 4.1, the monthly rainfall data for each province has a mean value

$\bar{X} > Q_2$ (Median), indicating that the monthly rainfall data is positively skewed and this is confirmed by the positive values of skewness. Eastern Cape, Gauteng KwaZulu-Natal provinces have kurtosis greater than three which indicate heavy tails than a normal distribution, while Limpopo and Mpumalanga have kurtosis less than three which indicate lighter tails than a normal distribution.

The standard deviation for all the five provinces ranges from 31.28 to 57.23 mm per month. KwaZulu-Natal province has the highest standard deviation with the value of 57.23 mm per month which indicates a large variation in the monthly rainfall series, while Mpumalanga province has the lowest standard deviation of 31.28 mm per month which implies a small variation in the monthly rainfall series.

The minimum monthly rainfall ranges between 0.01 mm and 0.50 mm per month where Eastern Cape receives the highest minimum rainfall of 0.50 mm per month, while Gauteng and KwaZulu-Natal receive the lowest minimum rainfall of 0.01 mm per month.

The maximum monthly rainfall lie between 111.00 mm and 478.80 mm per month where KwaZulu-Natal receives the highest maximum monthly rainfall of 478.80 mm per month followed by Gauteng with the maximum rainfall of 438.10 mm per month. Mpumalanga receives the lowest maximum rainfall of 111.00 mm per month.

Table 4.1: The descriptive statistics of the monthly rainfall data.

Provinces	**Min**	**Max**	**Median**	**Mean**	**Std.dev**	**Kurt**	**Skew**
Eastern Cape	0.50	211.00	42.50	49.03	34.46	4.06	0.99
Gauteng	0.01	438.10	45.45	58.45	55.62	5.22	1.18
KwaZulu-Natal	0.01	478.80	67.75	73.92	57.23	5.68	1.10
Limpopo	1.00	112.00	45.00	46.74	31.40	1.93	0.26
Mpumalanga	1.00	111.00	47.00	48.19	31.28	1.85	0.16

Note: Min=Minimum, Max=Maximum, Std.dev=Standard deviation, Kurt= Kurtosis, Skew= Skewness.

4.3 Test for stationarity results

The augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests were used to check for stationarity of monthly rainfall data for selected provinces of South Africa. Table 4.2 shows the results of the ADF, PP and KPSS tests.

The ADF and PP tests were tested under the following hypotheses:
H_0: the series has a unit root.
H_1: the series is stationary.

The KPSS test was tested under the following hypothesis:
H_0: The series does not have a unit root test (or series is stationary).
H_1: The series has a unit root (or series is not stationary).

From Table 4.2 the p-values of the ADF test statistics for Eastern Cape, Limpopo and Mpumalanga are significant ($p < 0.05$), suggesting that the monthly rainfall data for these three provinces are stationary. The ADF p-values for Gauteng and KwaZulu-Natal are insignificant ($p > 0.05$), suggesting that the monthly rainfall data for these two provinces are not stationary at 5% level of significance.

Also, from Table 4.2 the p-values of the KPSS test for all five provinces are significant ($p < 0.05$), suggesting that the monthly rainfall data are not stationary. Furthermore, from table 4.2 the p-values of the PP test for all five provinces are significant ($p < 0.05$), implying that the monthly rainfall data are stationary.

Overall, based on all the stationarity test findings, we conclude that the monthly rainfall data are not stationary for the majority of the provinces.

Table 4.2: ADF, KPSS and PP stationarity test results of monthly rainfall data.

Provinces	**Test**	**Test statistic**	**p-value**
Eastern Cape	ADF	-3.7614	0.02092
	KPSS	3.7258	0.01
	PP	-1432	0.01
Gauteng	ADF	-2.6238	0.3143
	KPSS	4.205	0.01
	PP	-840.85	0.01
KwaZulu-Natal	ADF	-2.6452	0.3052
	KPSS	4.1714	0.01
	PP	-1003.5	0.01
Limpopo	ADF	-7.1461	0.01
	KPSS	1.8398	0.01
	PP	-1502.6	0.01
Mpumalanga	ADF	-8.1155	0.01
	KPSS	0.96204	0.03041
	PP	-1312.9	0.010

4.4 Test for normality results

In this study we formally tested for normality of the monthly rainfall data using the Jarque-Bera (JB), Shapiro-Wilk (SW) and chi-square tests. Table 4.3 presents the results of the normality tests.

The JB, SW and chi-square tests are evaluated under the following hypotheses: H_0: The monthly rainfall data are normally distributed, versus, H_1: The monthly rainfall data do not come from a normal distribution.

From Table 4.3, the results for all the three normality tests are significant ($p < 0.05$), which suggest that the monthly rainfall data do not come from a normal distribution.

Table 4.3: JB, SW and chi-square normality test test results of monthly rainfall data.

Provinces	**Test**	**Test statistic**	**p-value**
Eastern Cape	JB	298.15	<0.01
	SW	0.93113	<0.01
	Chi-square	34276	<0.01
Gauteng	JB	618.22	<0.01
	SW	0.88137	<0.01
	Chi-square	78541	<0.01
KwaZulu-Natal	JB	710.9	<0.01
	SW	0.92103	<0.01
	Chi-square	62693	<0.01
Limpopo	JB	83.244	<0.01
	SW	0.9511	<0.01
	Chi-square	29843	<0.01
Mpumalanga	JB	83.833	<0.01
	SW	0.95219	<0.01
	Chi-square	28739	<0.01

4.5 Results for parent distributions

Five candidate parent distributions, namely: Gumbel, Weibull, gamma, Pareto and log-normal distributions, were fitted to the monthly rainfall data. The parent distribution with the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) was considered as the best fitting parent distribution. From Table 4.4 the best fitting parent distribution for the following provinces: Eastern Cape, Kwazulu-Natal, Limpopo and Mpumalanga is the Weibull distribution since it has the lowest values of AIC and BIC, while for Gauteng the best fitting parent distribution was found to be gamma distribution.

Table 4.4: AIC and BIC results for selection of parent distributions.

Provinces	Probability distributions	AIC	BIC	loglikelihood
Eastern Cape	log-normal	13812.61	13823.13	-6904.31
	Weibull	**13609.13**	**13619.64**	-6802.56
	Gumbel	13759.98	13770.4	-6877.94
	gamma	13625.88	13636.4	-6810.94
	Pareto	13859.42	13869.93	-6927.71
Gauteng	log-normal	14672.78	14683.29	-7334.39
	Weibull	13995.70	14006.21	-6995.85
	Gumbel	15098.24	15108.75	-7547.12
	gamma	**13843.41**	**13853.92**	-6919.71
	Pareto	14356.93	14367.45	-7176.47
KwaZulu-Natal	log-normal	15333.89	15344.4	-7664.95
	Weibull	**14954.07**	**14964.58**	-7475.03
	Gumbel	15206.38	15216.89	-8187.64
	gamma	14979.70	14990.21	-7487.85
	Pareto	15022.14	15032.66	-7509.07
Limpopo	log-normal	14109.66	14120.17	-7052.83
	Weibull	**13583.36**	**13593.87**	-6789.68
	Gumbel	13750.84	13761.35	-6873.42
	gamma	13649.70	13660.21	-6822.85
	Pareto	13723.83	13734.34	-6859.91
Mpumalanga	log-normal	14201.52	14212.03	-7098.76
	Weibull	**13636.32**	**13646.83**	-6816.16
	Gumbel	13781.86	13792.37	-6888.93
	gamma	13717.04	13727.55	-6856.52
	Pareto	13810.14	13820.65	-6903.07

4.5.1 Diagnostic plots of the best fitting parent distributions in the five provinces.

Figures 4.1-4.5 illustrate the diagnostic plots of the parent distribution for each of the five provinces: Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo and Mpumalanga. The diagnostic goodness-of-fit plots presented for each are quantile-quantile (Q-Q), empirical and theoretical density, empirical and theoretical CDFs, and probability-probability (P-P) plots.

Figure 4.1 presents the Weibull distribution diagnostic plots for Eastern Cape. Both the Q-Q and P-P plots are reasonably linear with very few outliers suggesting a very good fit of the Weibull distribution to the monthly rainfall data. Also, the density and CDF plots show a very good fit of the Weibull distribution. Therefore, since all the diagnostic plots suggest a good fit, we conclude that the suitable parent distribution of the Eastern Cape monthly rainfall data belongs to the Weibull distribution.

Figure 4.2 presents the gamma distribution diagnostic plots for Gauteng. Both the Q-Q and P-P plots are reasonably linear with some outliers indicating a better fit of the gamma distribution to the monthly rainfall data. Also, the density and CDF plots show a very good fit of the gamma distribution. Therefore, since all the diagnostic plots suggest a good fit, we conclude that the suitable parent distribution of the Gauteng monthly rainfall data belongs to the gamma distribution.

Figure 4.3 presents the Weibull distribution diagnostic plots for KwaZulu-Natal. Both the Q-Q and P-P plots are reasonably linear with some outliers suggesting a very good fit of the Weibull distribution to the monthly rainfall data. Also, the density and CDF plots show a very good fit of the Weibull distribution. Therefore, since all the diagnostic plots suggest a good fit, we conclude that the suitable parent distribution of the KwaZulu-Natal monthly rainfall data belongs to the Weibull distribution.

Figure 4.4 presents the Weibull distribution diagnostic plots for Limpopo. Both the Q-Q and P-P plots are reasonably linear with some outliers indicating a better fit of the Weibull distribution to the monthly rainfall data. Also, the density and CDF plots show a very good fit of the Weibull distribution. Therefore, since all the diagnostic plots suggest a good fit, we conclude that the suitable parent distribution of the Limpopo monthly rainfall data belongs to the Weibull distribution.

Figure 4.5 presents the Weibull distribution diagnostic plots for Mpumalanga. Both the Q-Q and P-P plots are reasonably linear with some outliers suggesting a very good fit of the Weibull distribution to the monthly rainfall data. Also, the density and CDF plots show a very good fit of the Weibull distribution. Therefore, since all the diagnostic plots suggest a good fit, we conclude that the suitable parent distribution of the Mpumalanga monthly rainfall data belongs to the Weibull distribution.

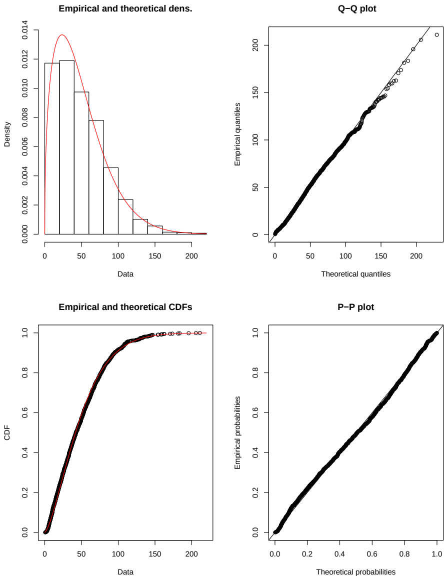

Figure 4.1: Weibull distribution diagnostic plots for Eastern Cape.

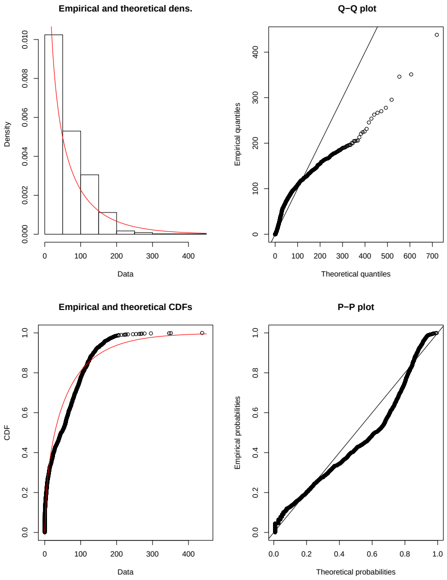

Figure 4.2: gamma distribution diagnostic plots for Gauteng.

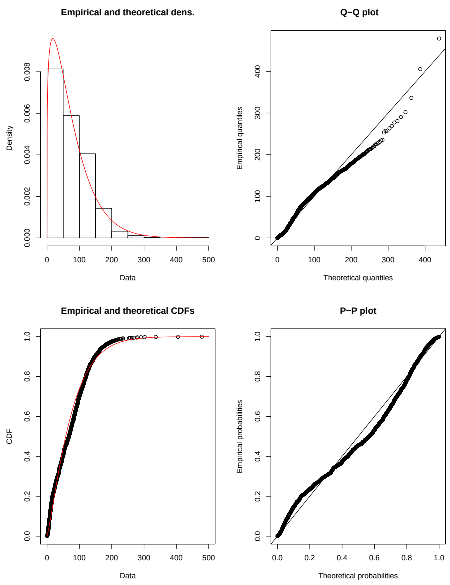

Figure 4.3: Weibull distribution diagnostic plots for KwaZulu-Natal.

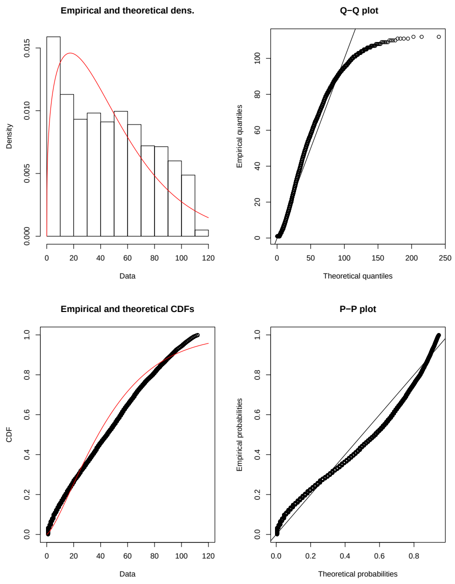

Figure 4.4: Weibull distribution diagnostic plots for Limpopo.

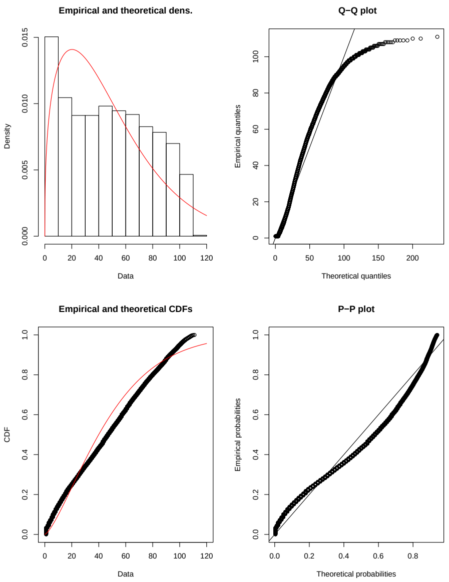

Figure 4.5: Weibull distribution diagnostic plots for Mpumalanga.

4.6 Trend analysis results

Mann-Kendall test statistic and time series plots were used to analyse the long-term trends of the monthly rainfall data for the five provinces. The Mann-Kendall test statistic results are presented in Table 4.5. The outcome of the Mann-Kendall test results revealed that in the Eastern Cape, Gauteng and KwaZulu-Natal provinces there were a significant monotonic decreasing long-term trends ($p < 0.05$ and τ was negative), while in Limpopo and Mpumalanga there were no significant monotonic decreasing long-term trends ($p > 0.05$ and τ was negative).

Figures 4.6-4.10 illustrate the monthly rainfall data time series plots for Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo and Mpumalanga provinces. The time series plots in Figures 4.6-4.10 do not exhibit any significant discrible long-term trends for all the provinces. This justifies the use of Mann-Kendall test to help uncover the hidden long-term trends in the monthly rainfall series in Table 4.5.

Table 4.5: Results for Mann-Kendall test statistic.

Provinces	**M-K test statistic**	**Kendall's tau**	**p-value**
Eastern Cape	-4.130	-0.073	0.01
Gauteng	-3.057	-0.054	0.002
KwaZulu-Natal	-2.399	-0.043	0.016
Limpopo	-0.832	-0.015	0.405
Mpumalanga	-0.487	-0.009	0.626

4.6.1 Time series plot for the five selected provinces of South Africa.

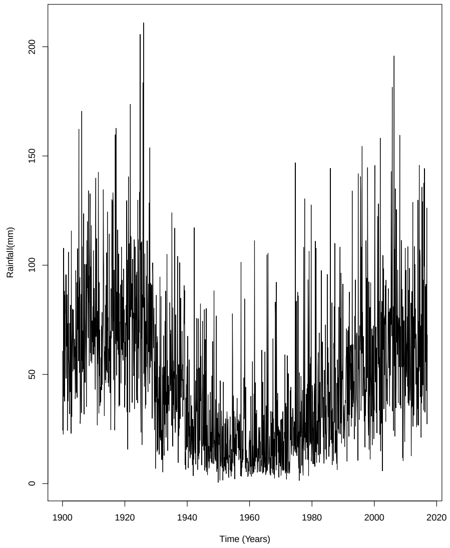

Figure 4.6: Time series plot for Eastern Cape monthly rainfall, 1900-2017.

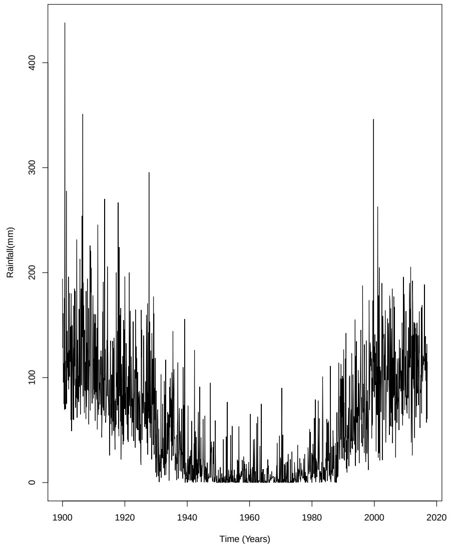

Figure 4.7: Time series plot for Gauteng, monthly rainfall, 1900-2017.

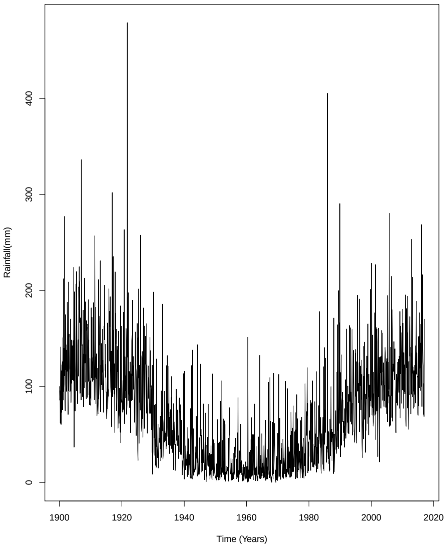

Figure 4.8: Time series plot for KwaZulu-Natal, monthly rainfall, 1900-2017.

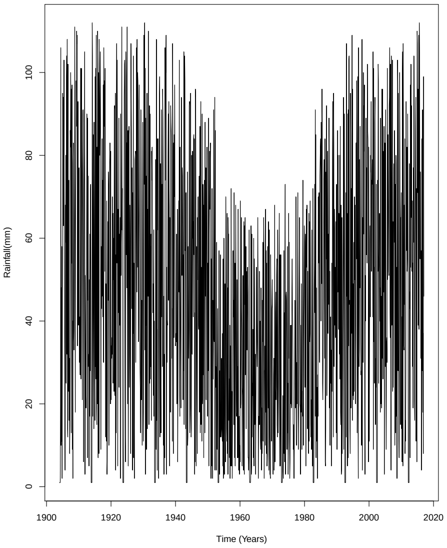

Figure 4.9: Time series plot for Limpopo, monthly rainfall, 1904-2017.

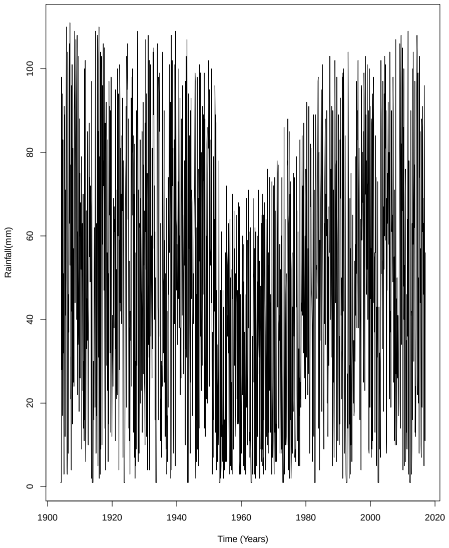

Figure 4.10: Time series plot for Mpumalanga, monthly rainfall, 1904-2017.

4.7 Non-stationary GEVD modelling of annual block maxima rainfall data

The time series plots of the annual block maxima rainfall series are shown in Figures 4.11-4.15. There seems to be rather strong evidence for a positive long-term trend over the years, for all the provinces. A substantial part of the variability in the data can probably be explained by a systematic variation in rainfall over the years. One way of capturing this trend is by allowing the GEVD location and scale parameters to vary with time. From Figures 4.11-4.15, a simple linear trend in time seems plausible for our annual maximum rainfall X_t, and we can use the model

$$X_t \sim GEV(\mu(t), \sigma(t), \xi), \tag{4.1}$$

where $\mu(t)$ and $\sigma(t)$ are the time-dependent location and scale parameters, respectively.

In the present study eight models are proposed for the non-stationary GEVD: $M_1, M_2, M_3, M_4, M_5, M_6, M_7$ and M_8. The reference model is denoted by M_0 and is the stationary GEVD. Model M_1 has a linear trend in the location parameter such that $\mu(t) = \mu_0 + \mu_1 t$, $\sigma(t) = \sigma$ and $\xi(t) = \xi$; Model M_2 has a linear trend in the scale parameter such that $\mu(t) = \mu$, $\log \sigma(t) = \exp(\sigma_0 + \sigma_1 t)$ and $\xi(t) = \xi$; Model M_3 has a linear trend in both location and scale parameters such that $\mu(t) = \mu_0 + \mu_1 t$, $\log \sigma(t) = \exp(\sigma_0 + \sigma_1 t)$ and $\xi(t) = \xi$; Model M_4 has a nonlinear quadratic trend in the location parameter and a linear trend in scale parameter such that $\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2$, $\log \sigma(t) = \exp(\sigma + \sigma_1 t)$ and $\xi(t) = \xi$; Model M_5 has a linear trend in the location parameter and a nonlinear quadratic trend in the scale parameter such that $\mu(t) = \mu_0 + \mu_1 t$, $\log \sigma(t) = \exp(\sigma_0 + \sigma_1 t + \sigma_2 t^2)$

and $\xi(t) = \xi$; Model M_6 has a nonlinear quadratic trend in both location and scale parameters such that $\mu(t) = \mu_0+\mu_1 t+\mu_2 t^2$, $\log\sigma(t) = \exp(\sigma_0+\sigma_1 t+\sigma_2 t^2)$ and $\xi(t) = \xi$; Model M_7 has a nonlinear quadratic trend in the location parameter only with no variation in scale such that $\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2$, $\sigma(t) = \sigma$ and $\xi(t) = \xi$; Model M_8 has a nonlinear quadratic trend in the scale parameter with no variation in the location parameter such that $\mu(t) = \mu$, $\log\sigma(t) = \exp(\sigma_0 + \sigma_1 t + \sigma_2 t^2)$ and $\xi(t) = \xi$.

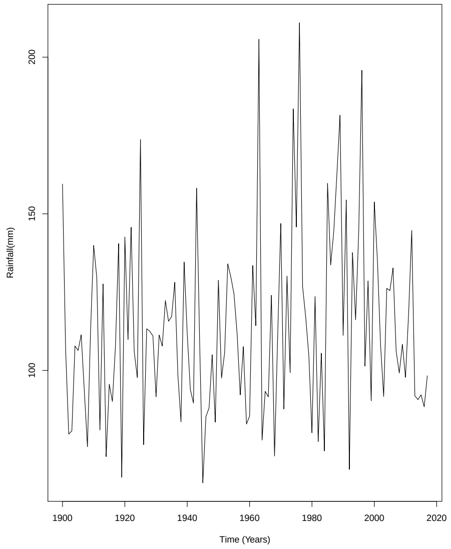

Figure 4.11: Time series plot showing the annual block maximum rainfall in mm observed in Eastern Cape, 1900–2017.

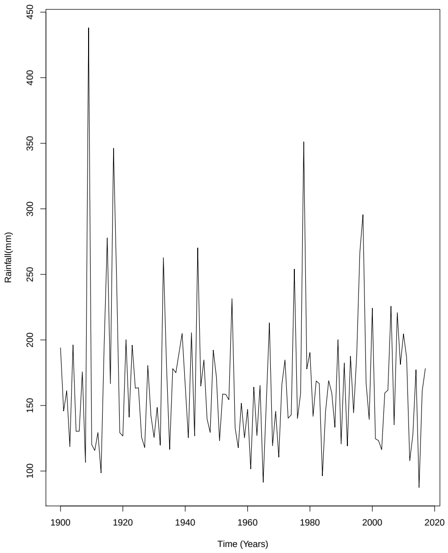

Figure 4.12: Time series plot showing the annual block maximum rainfall in mm observed in Gauteng, 1900–2017.

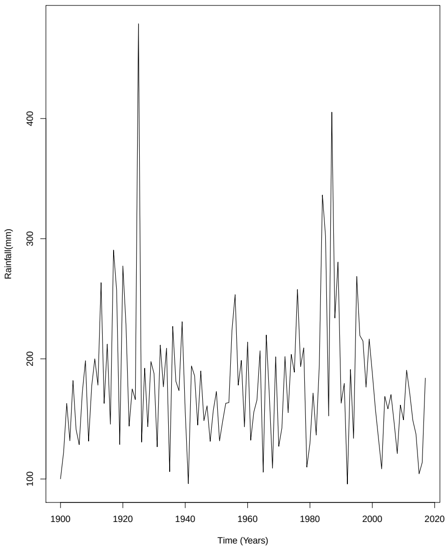

Figure 4.13: Time series plot showing the annual block maximum rainfall in mm observed in KwaZulu-Natal, 1900–2017.

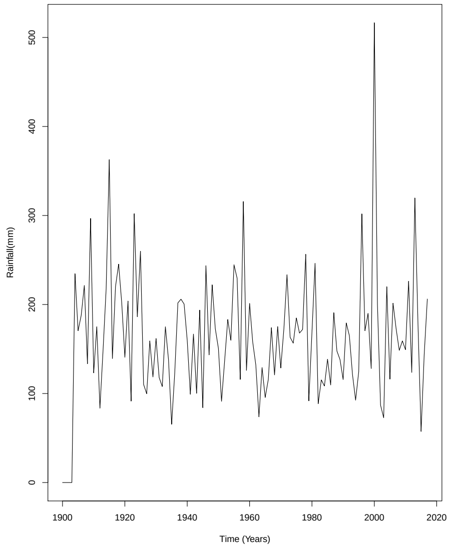

Figure 4.14: Time series plot showing the annual block maximum rainfall in mm observed in Limpopo, 1904–2017.

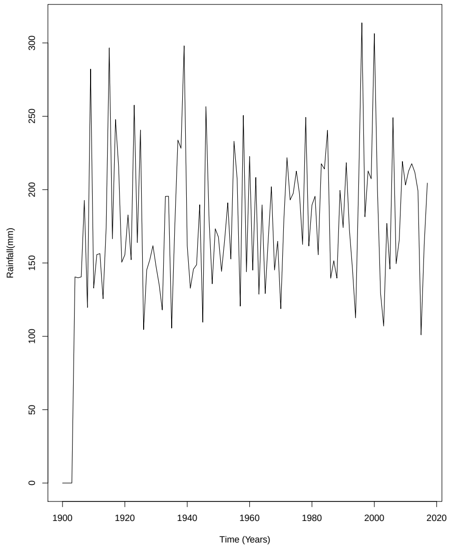

Figure 4.15: Time series plot showing the annual block maximum rainfall in mm observed in Mpumalanga, 1904–2017.

4.7.1 Eastern Cape

The stationary GEVD model for Eastern Cape data (i.e., model M_0) has a maximum negative log-likelihood (NLLH) of 556.765 (see Table 4.6). A GEVD model with linear trend in the location parameter (i.e., M_1) has a maximum NLLH of 555.820. The deviance statistic for comparing these two models is therefore, D=2(556.769-555.820)=1.898, which is small compared to $\chi_1^2(0.05) = 3.841$. Thus, allowing for a linear trend in the location parameter does not improve on our stationary GEVD model, M_0. Therefore, M_1 is not a worth model to consider.

Consider the pair of models (M_0, M_2) from Table 4.6. The deviance statistic is 2(556.769-555.724)=2.090, which is small compared to $\chi_1^2(0.05) = 3.841$. Thus, allowing for a linear trend in the scale parameter does not improve on our stationary GEVD model, therefore, we reject model M_2 and conclude that is not worthwhile to allow for a linear trend in the scale parameter.

From Table 4.6, the deviance statistics of model pairs (M_0, M_3) and (M_0, M_7) are 2.478 and 1.442, respectively. Since both values of the deviance statistics are smaller than $\chi_2^2(0.05) = 5.991$, it implies that both models do not provide any improvement in fit over the stationary GEVD model. The other model pairs from Table 4.6 (M_0, M_4) and (M_0, M_5), have deviance statistics of 1.864 and 0.452, respectively. These results revealed that model M_4, which allows for nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter, does not provide an improvement in fit over the stationary GEVD model since the value of the deviance statistic (1.864) is small as compared to the value of $\chi_3^2(0.05) = 7.815$. On the other hand, model M_5, which allows for linear trend in location parameter and a nonlinear quadratic trend in the scale parameter, does not provide an improvement in fit over the stationary GEVD model since the value of the deviance statistic is smaller than the value of $\chi^2_3(0.05) = 7.815$.

The nonlinear quadratic model pair (M_0, M_6), which allows for nonlinear quadratic trend in both location and scale parameters, does not improve our stationary GEVD model since the deviance statistic , D=1.37, is small compared to $\chi^2_4(0.05) = 9.488$. Again in Table 4.6, the model pair (M_0, M_8), which allows for nonlinear quadratic trend in scale parameter with no variation in location parameter, has a deviance statistic of 0.354, which is too small compared to the critical value of 5.991 with 2 degrees of freedom. Thus, allowing for a quadratic trend in the scale parameter with no variation in the location parameter does not improve on our stationary GEVD.

Table 4.6: Non-stationary GEVD models for Eastern Cape for the period 1900-2017.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**
M_0	100.782	0	0	23.244	0	0	-0.012	556.769
M_1	95.768	0.086	0	23.057	0	0	-0.013	555.820
M_2	100.7005	0	0	22.328	0.017	0	-0.019	556.724
M_3	94.955	0.102	0	20.715	0.041	0	-0.022	555.530
M_4	98.043	-0.039	0.001	20.928	0.036	0	-0.010	555.837
M_5	99.323	0.002	0	18.126	0.310	-0.003	0.092	556.543
M_6	96.672	-0.003	0.001	17.7333	0.246	-0.002	0.095	556.084
M_7	98.475	-0.033	0.001	22.982	0	0	-0.002	556.048
M_8	99.499	0	0	18.263	0.305	-0.003	0.091	556.592

Key: NLLH = negative log-likelihood.

Overall, the final model for Eastern Cape is the stationary GEVD model, M_0. The general model for Eastern Cape is given by

$$GEV(x, \mu, \sigma, \xi) = \exp - \left[1 - 0.012\left(\frac{x - 100.782}{23.244}\right)\right]^{\frac{1}{0.012}}. \tag{4.2}$$

The shape parameter (-0.012) for the model, M_0, indicates that the rainfall data for Eastern Cape can be modelled by the Weibull class of distribution since the shape parameter $\xi < 0$. The diagnostic plots for the stationary GEVD model in (4.2) are presented in Figure 4.16. The diagnostic plot results in Figure 4.16 show that the stationary GEVD model, M_0, is the best fit for the Eastern Cape monthly rainfall data.

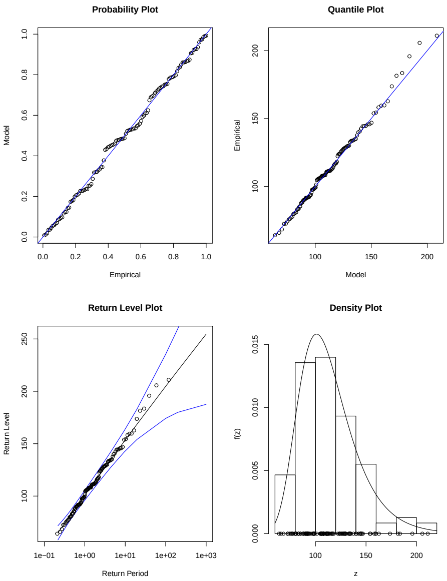

Figure 4.16: Diagnostic plots for the stationary GEVD best fitting model for Eastern Cape province.

Goodness-of-fit test for Eastern Cape GEVD model

The goodness-of-fit test based on Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests were performed in order to check if the maximum monthly rainfall data for Eastern Cape follow a stationary GEVD model. Table 4.7 presents the results of the K-S and A-D goodness-of-fit tests for the selected stationary GEVD model for the Eastern Cape.

The hypotheses are formulated as follows H_0: The monthly rainfall data follow a specified distribution, and H_1: The monthly rainfall data do not follow the specified distribution.

Since the p-values for both the K-S and A-D tests are greater than the 5% level of significance, α = 0.05, we conclude that the maximum monthly rainfall for Eastern Cape follow the specified stationary GEVD.

Table 4.7: Goodness-of-fit for Eastern Cape (1900-2017).

Test	Test statistic	p-value
K-S	0.056844	0.8403205
A-D	0.1935343	0.8918115

4.7.2 Gauteng

The model pairs (M_0, M_1) and (M_0, M_2) from Table 4.8 have the same critical value of $\chi^2_1(0.05) = 3.841$ with the deviance statistic values of 0.022 and 0.250 for M_1 and M_2, respectively. Since the value of the deviance statistics for M_1 (0.022) and M_2 (0.250) are smaller than the critical value of 3.841, we conclude that both models do not provide any improvement in fit over the stationary GEVD model.

From Table 4.8, the deviance statistics of model pairs (M_0, M_3) and (M_0, M_7) are 0.272 and 0.130, respectively. Since the values of the deviance statistics for both model pairs are smaller than $\chi^2_2(0.05) = 5.991$, it implies that both models do not provide any improvement in fit over the stationary GEVD model. The model pair (M_0, M_6) from Table 4.8 has $\chi^2_4(0.05) = 9.488$ and a deviance statistic value of 1.706. Since the deviance statistic value (1.706) is smaller than the critical value of 9.488, we conclude that model M_6 does not provide any improvement in fit over the stationary GEVD model.

The other pairs from Table 4.8, i.e. (M_0, M_4) and (M_0, M_5), have deviance statistics of 0.254 and 1.704, respectively. These results revealed that model M_4, which allows for nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter, does not improve on the stationary GEVD model since the value of the deviance statistic (1.864) is small as compared to the value of $\chi_3^2(0.05) = 7.815$. On the other hand, model M_5, which allows for linear trend in the location parameter and a nonlinear quadratic trend in the scale parameter, does not provide any improvement on the stationary GEVD model because the value of the deviance statistic is smaller than the critical value of $\chi_3^2(0.05) = 7.815$. The model pair (M_0, M_8), which allows for nonlinear quadratic trend in scale parameter with no variation in location parameter, has a deviance statistic of 1.710, which is small compared to the critical value of 5.991 with 2 degrees of freedom. Thus, allowing for a quadratic trend in the scale parameter with no variation in the location parameter does not improve on the stationary GEVD model, therefore, model M_8 is not worthwhile.

Table 4.8: Non-stationary GEVD models for Gauteng for the period 1900-2017.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**
M_0	141.929	0	0	34.705	0	0	0.117	612.516
M_1	142.629	-0.012	0	34.669	0	0	0.118	612.505
M_2	141.690	0	0	32.345	0.032	0	0.128	612.391
M_3	140.811	0.015	0	32.138	0.039	0	0.129	612.380
M_4	141.474	0.016	0.000	32.514	0.0319	0	0.133	612.389
M_5	142.238	-0.002	0	39.864	-0.303	-0.003	0.108	611.664
M_6	141.590	0.007	0.000	39.640	-0.300	0.003	0.105	611.663
M_7	142.800	0.007	0.000	34.573	0	0	0.125	612.451
M_8	142.117	0	0	39.789	-0.302	0.003	0.108	611.661

Key: NLLH = negative log-likelihood.

The best fit model for Gauteng is the stationary GEVD model, M_0, and is given by

$$GEV(x, \mu, \sigma, \xi) = \exp - \left[1 + 0.117\left(\frac{x - 141.292}{34.705}\right)\right]^{\frac{-1}{0.117}} . \qquad (4.3)$$

The shape parameter (0.117) for the stationary GEVD model, M_0, indicates that the rainfall data for Gauteng can be modelled using Fréchet distribution class of distribution since the shape parameter $\xi > 0$. The diagnostic plots for the stationary GEVD model in (4.3) are presented in Figure 4.17. The diagnostic plot results in Figure 4.17 reveal that the stationary GEVD model, M_0, in the Fréchet distribution of attraction is the best fit for Gauteng monthly rainfall data. These findings are consistent with parent distribution selection findings which revealed that Gauteng monthly rainfall data cannot be best modelled by a distribution in the Weibull domain of attraction.

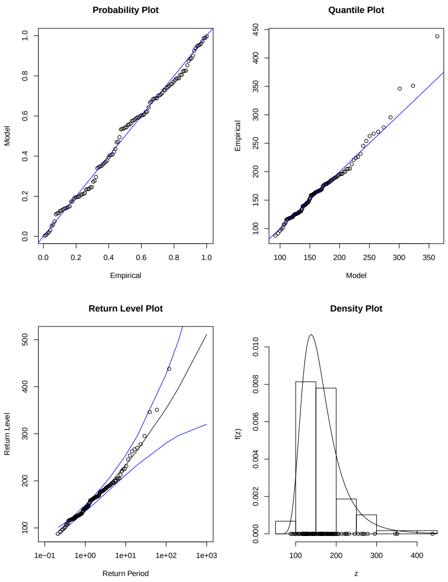

Figure 4.17: Diagnostic plots for the stationary GEVD best fitting model for Gauteng province.

Goodness-of-fit test for Gauteng GEVD model

Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests were used to determine whether maximum monthly rainfall data for Gauteng follow a stationary GEVD. Table 4.9 presents the results of the K-S and A-D goodness-of-fit tests for Gauteng stationary GEVD model.

The results from Table 4.9 show that the p-values for both the K-S and A-D tests are not significant (p $>$ 0.05). Therefore, we conclude that the maximum monthly rainfall for Gauteng province follow the specified stationary GEVD.

Table 4.9: Goodness-of-fit for Gauteng (1900-2017).

Test	Test Statistic	p-value
K-S	0.0673058	0.6590246
A-D	0.3562733	0.4519496

4.7.3 KwaZulu-Natal

Consider the model pairs (M_0, M_1) and (M_0, M_2) from Table 4.10. The critical value for both pairs is $\chi_1^2(0.05) = 3.841$ with respective deviance statistic values of (0.210 and 0.026) for the two model pairs. The pairs (M_0, M_1) and (M_0, M_2) do not provide any improvement in fit over the stationary GEVD model since the deviance statistic values (0.210) and (0.026) are smaller than the critical value of 3.841 with 1 degree of freedom.

Consider the model pair (M_0, M_3) from Table 4.10 with $\chi_2^2(0.05) = 5.991$ and deviance statistic of 0.224 which is too small compared to the critical value of 5.991 with 2 degrees of freedom. Thus, allowing for linear trend in the location and scale parameter is not worthwhile over the stationary GEVD model. The other pairs from Table 4.10 (M_0, M_4) and (M_0, M_5) have deviance statistics of 0.624 and 0.226, respectively. These results revealed that model M_4, which allows for nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter, is not worthwhile over the stationary GEVD model since the value of the deviance statistic (0.624) is small as compared to the value of $\chi_3^2(0.05) = 7.815$. On the other hand, model M_5, which allows for linear trend in the location parameter and a nonlinear quadratic trend in the scale parameter, does not provide any improvement in fit over the stationary

GEVD model because the value of the deviance statistic (0.226) is smaller than the value of $\chi_3^2(0.05) = 7.815$.

The model pairs (M_0, M_7) and (M_0, M_8) in Table 4.10 share a critical value of $\chi_2^2(0.05) = 5.991$ with deviance statistic values of 2.248 and -0.176 for M_7 and M_8, respectively. Since the values of the deviance statistics are smaller than the critical value of 5.991 with 2 degrees of freedom, it implies that both models do not provide any improvement in fit over the stationary GEVD model.

The model pair (M_0, M_6), which allows for nonlinear quadratic trend in both the location and scale parameters in Table 4.10, has a deviance statistic of 0.598 which is too small compared to the critical value of 9.488 with 4 degrees of freedom. Thus, allowing for a quadratic trend in both the location and scale parameters does not improve on our stationary GEVD model.

Table 4.10: Non-stationary GEVD models for KwaZulu-Natal for the period 1900-2017.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**
M_0	153.756	0	0	39.560	0	0	0.070	624.418
M_1	156.383	-0.044	0	39.518	0	0	0.070	624.313
M_2	153.791	0	0	38.808	0.012	0	0.071	624.405
M_3	156.817	-0.051	0	40.195	-0.011	0	0.070	624.306
M_4	158.193	-0.002	-0.0007	41.398	-0.003	0	0.009	624.106
M_5	157.029	-0.006	0	40.021	0.002	-0.0001	0.007	624.305
M_6	157.126	-0.005	-0.0008	39.892	0.036	-0.0006	0.098	624.119
M_7	146.685	0.464	-0.004	39.308	0	0	0.066	623.294
M_8	153.260	0	0	38.741	0.005	0.0000	0.011	624.506

Key: NLLH = negative log-likelihood.

Overall, the final best model for KwaZulu-Natal is the stationary GEVD model,

M_0. The general model for KwaZulu-Natal is given by

$$GEV(x,\mu,\sigma,\xi) = \left\{ \exp - \left[1 + 0.070\left(\tfrac{x-153.756}{39.560}\right)\right]^{\frac{-1}{0.070}} \right. . \qquad (4.4)$$

The shape parameter (0.070) for the model M_0, indicates that the rainfall data for KwaZulu-Natal can be modelled using Fréchet class of distributions since the shape parameter $\xi > 0$. The diagnostic plots for the stationary GEVD model in (4.4) are presented in Figure 4.18. The results in Figure 4.18 show that the stationary GEVD model, M_0, is the best fit for KwaZulu-Natal maximum monthly rainfall data since all the four diagnostic plots suggest a reasonable good fit for the stationary GEVD model.

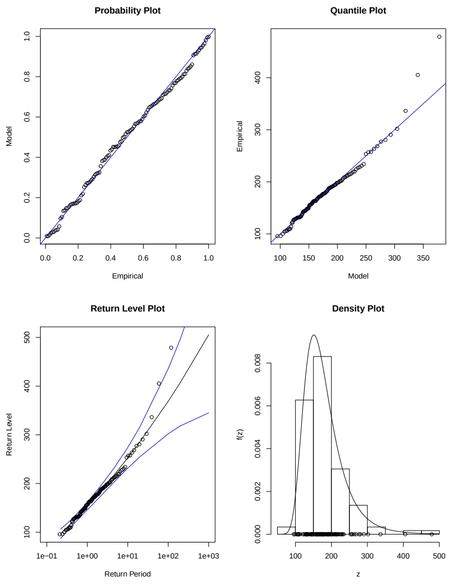

Figure 4.18: Diagnostic plots for the stationary GEVD best fitting model for KwaZulu-Natal province.

Goodness-of-fit test for KwaZulu-Natal GEVD model

Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests were used to determine whether maximum monthly rainfall data for KwaZulu-Natal follow a stationary GEVD model. Table 4.11 presents the K-S and A-D goodness-of-fit tests results for KwaZulu-Natal GEVD.

From Table 4.11 the p-values for both K-S and A-D tests are insignificant ($p > 0.05$) at 5% level of significance. Thus, we conclude that the maximum monthly rainfall for KwaZulu-Natal follow the specified stationary GEVD model.

Table 4.11: Goodness-of-fit for KwaZulu-Natal (1900-2017).

Test	Test Statistic	p-value
K-S	0.04470146	0.9724252
A-D	0.3284819	0.5135279

4.7.4 Limpopo

The stationary GEVD model for Limpopo data (i.e., model M_0) has a maximum NLLH of 669.707. A GEVD model with linear trend in the location parameter (i.e., M_1) has a maximum NLLH of 666.705 (see Table 4.12). The deviance statistic for comparing these two models is therefore D=2(669.707-666.705)=6.004, which is greater than the critical value of 3.841 with 1 degree of freedom and therefore model M_1 provides an improvement in fit over the stationary GEVD model. The likelihood ratio test for μ_1 = 0 has p-value= 0.005, which is significant at 5% level of significance ($p < 0.05$). This clearly shows that the non-stationary GEVD model is worthwhile and does provide an improvement in fit over the stationary GEVD model.

Consider the pair of models (M_0, M_2) from Table 4.12. The deviance statistic is 2(669.707-665.327)=8.760, which is large compared to $\chi_1^2(0.05) = 3.841$. Thus, allowing for a linear trend in the scale parameter does improve on our stationary GEVD model. The likelihood ratio test for σ_1 = 0 has p-value of 0.001, implying that the linear trend in the scale parameter is significant at 5% level of significance ($p < 0.05$), which indicates that model M_2 is important and does provide an improvement in fit over the stationary GEVD model.

From Table 4.12, the pair of models (M_0, M_3), has the deviance statistic of 11.014, which is greater than the critical value of 5.991 with two degrees of freedom, implying that model M_3 provides an improvement in fit over the stationary GEVD model. The likelihood ratio test for μ_1 = 0 has p-value = 0.067, which indicates that the likelihood ratio test is not significant at 5% level of significance (p $>$ 0.05), while the likelihood ratio test for σ_1 = 0 has p-value = 0.013, which indicates that the likelihood ratio test is significant at 5% level of significance (p $<$ 0.05).

The other pairs from Table 4.12, (M_0, M_4) and (M_0, M_5), have deviance statistic values of 19.040 and 7.900, respectively. These results revealed that model M_4, which allows for nonlinear quadratic trend in the location parameter and linear trend in the scale parameter, provides an improvement in fit over the stationary GEVD model since the value of the deviance statistic (19.040) is larger as compared to the value of $\chi^2_3(0.05) = 7.815$. The likelihood ratio test for μ_1 = 0 has p-value= 0.001, for μ_2 = 0 it has p-value of 0.002, and for σ_1 = 0 it has p-value= 0.034, which are all significant at 5% level of significance (p $<$ 0.05). On the other hand, model M_5 which allows for linear trend in the location parameter and a nonlinear quadratic trend in the scale parameter, provides an improvement in the stationary GEVD model since the value of the deviance statistic is greater than the value of $\chi^2_3(0.05) = 7.815$. The likelihood ratio test for μ_1 = 0 has p-value= 0.236, which is not significant at 5% level of significance (p $>$ 0.05), while the likelihood ratio test for σ_1 = 0, and σ_2 = 0, all have p-values $<$ 0.001, which are both significant at 5% level of significance (p $<$ 0.05).

The model pair (M_0, M_6), which allows for nonlinear quadratic trend in both the location and scale parameters in Table 4.12, has a deviance statistic of

9.046 which is small compared to the critical value of 9.488 with 4 degrees of freedom. Thus, allowing for a quadratic trend in both the location and scale parameters is not worthwhile in fit over the stationary GEVD model M_0. The likelihood ratio test for μ_1 = 0 has p-value = 0.145, and for μ_2 = 0 has p-value = 0.185, which is insignficant at 5% level of significance (p $>$ 0.05), while the likelihood ratio test for σ_1 = 0, and σ_2 = 0, all have p-values $<$ 0.001, which are both significant at 5% level of significance (p $<$ 0.05).

Consider the model pair (M_0, M_7) in Table 4.12 with deviance statistic of 15.820, which is greater than the critical value of $\chi_2^2(0.05) = 5.991$, indicating that the non-stationary GEVD model provides an improvement in fit over the stationary GEVD model. The likelihood ratio tests for μ_1 = 0, and μ_2 = 0 have p-values $<$ 0.001, which indicates that the likelihood ratio tests are significant at 5% level of significance (p $<$ 0.05) for the quadratic trend in the location parameter with no variation in the scale parameter. This implies that the non-stationary GEVD model is worthwhile and does give an improvement in fit over the stationary GEVD model.

Consider the model pair (M_0, M_8) from Table 4.12 with $\chi_2^2(0.05) = 5.991$ and deviance statistic of 9.338. The likelihood ratio tests for σ_1= 0 and σ_2 = 0 have p-values $<$0.001. These results show that the nonlinear quadratic trend in scale parameter with no variation in the location parameter is significant at 5% level of significance (p $<$ 0.05). The deviance statistic (9.338) is greater than the critical value of 5.991, which implies that the non-stationary GEVD model, M_8, is important and does provide an improvement in fit over the stationary GEVD model.

Table 4.12: Non-stationary GEVD models for Limpopo for the period 1904-2017.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**
M_0	132.224	0	0	65.611	0	0	-0.097	669.707
M_1	105.813	0.423	0	61.752	0	0	-0.067	666.705
M_2	133.060	0	0	78.463	-0.289	0	-0.030	665.327
M_3	115.204	0.258	0	73.135	-0.218	0	-0.036	664.200
M_4	74.261	2.073	-0.015	65.293	-0.178	0	0.040	660.187
M_5	122.793	0.132	0	105.672	-1.988	0.014	0.105	655.757
M_6	107.754	0.732	-0.005	99.611	-1.756	0.013	0.094	655.184
M_7	62.447	2.432	-0.017	54.826	0	0	0.047	661.797
M_8	133.880	0	0	107.223	-2.009	0.015	0.008	665.038

Key: NLLH = negative log-likelihood.

Overall, Limpopo has five competing non-stationary GEVD models: M_1, M_2, M_4, M_7 and M_8, for which only two models were considered based on their deviance statistic values as main and alternative best models. The best non-stationary GEVD model is M_4, which has a nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter, and is given by

$$GEV(x, \mu, \sigma, \xi) = \left\{ \exp - \left[1 + 0.040 \left(\tfrac{x-74.261}{65.293}\right)\right]^{\frac{-1}{0.040}} \right. . \tag{4.5}$$

The alternative non-stationary GEVD model is M_7, which has a nonlinear quadratic trend in the location parameter and no variation in the scale parameter, and is given by:

$$GEV(x, \mu, \sigma, \xi) = \left\{ \exp - \left[1 + 0.047 \left(\tfrac{x-62.447}{54.826}\right)\right]^{\frac{-1}{0.047}} \right. . \tag{4.6}$$

The shape parameters in (4.5) and (4.6), that is, 0.040 and 0.047 for the models M_4 and M_7, respectively, are poisitive, which indicates that the rainfall data for Limpopo can be modelled using Fréchet distribution class since the shape parameter $\xi > 0$. The diagnostic plots for the stationary GEVD model in (4.5) are presented in Figure 4.19. The results in Figure 4.19 show that model M_4 is the best fit for Limpopo maximum monthly rainfall data since the two diagnostic plots indicate a reasonable good fit for the non-stationary GEVD model with a nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter.

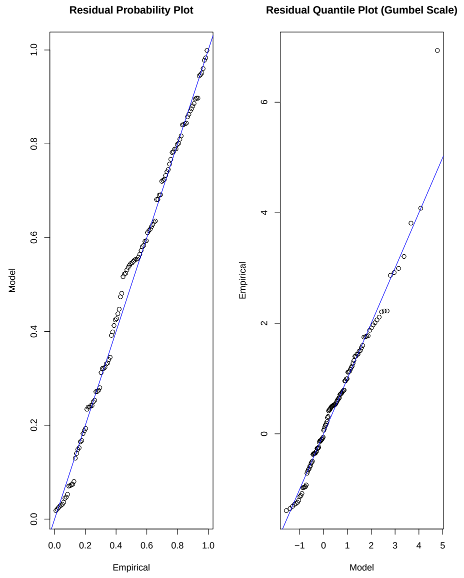

Figure 4.19: Diagnostic plots for the non-stationary GEVD best fitting model for Limpopo province.

Goodness-of-fit test for Limpopo non-stationary GEVD model

Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests were used to determine whether maximum monthly rainfall data for Limpopo follows the non-stationary GEVD model, M_4. Table 4.13 presents the K-S and A-D goodness-of-fit tests.

From Table 4.13, the p-value for the K-S test is insignificant (p $>$ 0.05), implying that the maximum monthly rainfall for Limpopo follows the non-sationary GEVD model, while the results from the A-D test suggest that the maximum monthly rainfall for Limpopo do not follow the specified non-stationary GEVD model (p $<$ 0.05).

Table 4.13: Goodness-of-fit for Limpopo (1904-2017).

Test	Test Statistic	p-value
K-S	0.07362455	0.5445211
A-D	1.133259	0.005549523

4.7.5 Mpumalanga

The model pairs (M_0, M_1) and (M_0, M_2) in Table 4.14 share the critical value of $\chi^2_1(0.05) = 3.841$ with repective deviance statistic values of 10.008 and 7.236 for the two pairs. The two pairs have p-values of 0.001 and 0.003 for μ_1 = 0 and σ_1 = 0, respectively for model M_1 and M_2. These results revealed that the model pairs (M_0, M_1) and (M_0, M_2) are significant at 5% level of significance (p $<$ 0.05). The deviance statistic values for the two models are large in comparison to $\chi^2_1(0.05) = 3.841$. Thus, we conclude that models M_1 and M_2 provide a significant improvement over the stationary GEVD model, M_0.

From Table 4.14, the pair of models (M_0, M_3) has a deviance statistic of 19.530, which is greater than the critical value of 5.991 with two degrees of freedom, implying that model M_3 provides an improvement in fit over the stationary GEVD model. The likelihood ratio tests for μ_1 = 0 and σ_1 = 0 have p-values $<$ 0.001, which indicates that the likelihood ratio tests are significant at 5% level of significance (p $<$ 0.05) for both the location and scale parameters, implying that the non-stationary GEVD model is important and does provide an improvement in fit over the stationary GEVD model.

The other model pairs from Table 4.14, (M_0, M_4) and (M_0, M_5), have deviance statistic values of 23.330 and 23.898, respectively. These results revealed that model M_4, which allows for nonlinear quadratic trend in the location parameter and linear trend in the scale parameter, is worthwhile over the stationary GEVD model since the value of the deviance statistic (23.330) is greater than the value of $\chi^2_3(0.05) = 7.815$. The likelihood ratio test for μ_1 = 0 has p-value= 0.392, and for μ_2 = 0 it has p-value of 0.096, which are both not significant at 5% level of significance (p $>$ 0.05), but the likelihood ratio test for σ_1 = 0 has p-value $<$ 0.001, which is significant at 5% level of significance (p $<$ 0.05). On the other hand, model M_5 which allows for linear trend in the location parameter and a nonlinear quadratic trend in the scale parameter, provides an improvement in fit over the stationary GEVD model since the value of the deviance statistic is greater than the value of $\chi^2_3(0.05) = 7.815$. The likelihood ratio test for μ_1 = 0, σ_1 = 0 and σ_2 = 0, all have p-values $<$ 0.001, which are significant at 5% level of significance (p $<$ 0.05).

The model pair (M_0, M_6) in Table 4.14, which allows for nonlinear quadratic trend in both the location and scale parameters, has a deviance statistic of 24.512 which is greater than the critical value of 9.488 with 4 degrees of freedom. Thus, allowing for a quadratic trend in both location and the scale parameters is worthwhile in fit over the stationary GEVD model, M_0. The likelihood ratio test for μ_1= 0 has p-value = 0.499, and μ_2 = 0 has p-value = 0.303, which is insignficant at 5% level of significance (p $>$ 0.05), while the likelihood ratio tests for σ_1 = 0 and σ_2 = 0 all have p-values $<$ 0.001, which are significant at 5% level of significance (p $<$ 0.05).

Consider the model pair (M_0, M_7) in Table 4.14, with a deviance statistic of 6.394 which is greater than the critical value of $\chi^2_2(0.05) = 5.991$. These results show that the non-stationary GEVD model provides an improvement in fit over the stationary GEVD model. The likelihood ratio test for μ_1 = 0 it has p-value = 0.369 and for μ_2 = 0 it has p-value = 0.449, which are both not significant at 5%

level of significance (p $>$ 0.05). This implies that model M_7, with a quadratic trend in the scale parameter and no variation in the location parameter is not worthwhile over the stationary GEVD model.

Consider the model pair (M_0, M_8) from Table 4.14 with $\chi^2_2(0.05) = 5.991$ and deviance statistic value of 29.150. The likelihood ratio tests for σ_1= 0 and σ_2 = 0 have p-values <0.001. These results show that the nonlinear quadratic trend in scale parameter with no variation in the location parameter is significant at 5% level of significance (p $<$ 0.05). The deviance statistic (29.150) is greater than the critical value of 5.991, which implies that the non-stationary GEVD model, M_8, is important and does provides an improvement in fit over the stationary GEVD model.

Table 4.14: Non-stationary GEVD models for Mpumalanga for the period 1904-2017.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**
M_0	155.612	0	0	59.246	0	0	-0.325	643.234
M_1	124.429	0.512	0	54.437	0	0	-0.261	638.230
M_2	159.885	0	0	71.217	-0.274	0	-0.228	639.616
M_3	131.503	0.428	0	69.312	-0.268	0	-0.240	633.469
M_4	114.907	1.296	-0.007	71.758	-0.310	0	-0.207	631.569
M_5	161.943	-0.031	0	113.977	-2.386	0.017	-0.006	631.285
M_6	145.643	-0.001	0.002	108.921	-2.381	0.017	0.076	630.978
M_7	145.591	0.324	-0.0009	58.266	0	0	-0.328	640.037
M_8	161.734	0	0	100.364	-1.981	0.014	-0.161	628.659

Key: NLLH = negative log-likelihood.

In general, in Mpumalanga there were five competing non-stationary GEVD models: M_1, M_2, M_3, M_5 and M_8, for which only two models were considered based on their deviance statistic values as main and alternative best models. The best non-stationary GEVD model is M_8, which has a nonlinear quadratic trend in the scale parameter and no variation in the location parameter, and is given by

$$GEV(x, \mu, \sigma, \xi) = \left\{ \exp - \left[1 - 0.014 \left(\tfrac{x-161.734}{100.364}\right)\right]^{\frac{1}{0.161}} \right. . \tag{4.7}$$

The alternative non-stationary GEVD model, is M_5, which has a linear trend in location parameter and nonlinear quadratic trend in scale paramater and is given by:

$$GEV(x, \mu, \sigma, \xi) = \left\{ \exp - \left[1 - 0.006 \left(\tfrac{x-161.943}{113.977}\right)\right]^{\frac{1}{0.006}} \right. . \tag{4.8}$$

The shape parameter in (4.7) and (4.8), that is, -0.161 and -0.006 for the respective models M_8 and M_5 are negative, which indicates that the rainfall data for Mpumalanga can be modelled using Weibull distribution class since the shape parameter $\xi < 0$. The diagnostic plots for the non-stationary GEVD model in (4.7) are presented in Figure 4.20. The results in Figure 4.20 show that the non-stationary GEVD model, M_8, is the best fit for Mpumalanga maximum monthly rainfall data because the two diagnostic plots suggest a reasonable good fit for the non-stationary GEVD model with a quadratic trend in the scale parameters and no variation in other parameters.

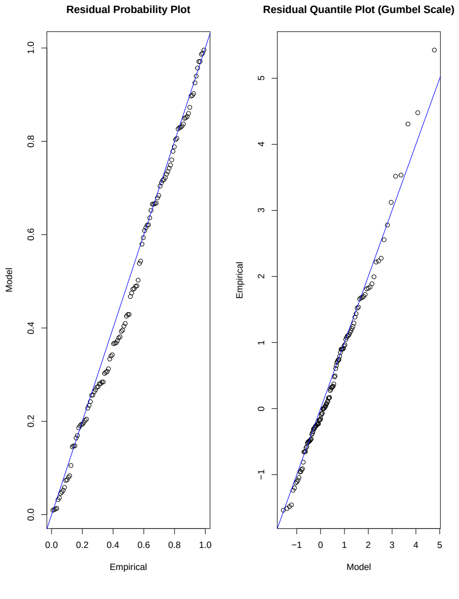

Figure 4.20: Diagnostic plots for the non-stationary GEVD best fitting model for Mpumalanga province.

Goodness-of-fit test for Mpumalanga non-stationary GEVD model

Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests were used to determine whether maximum monthly rainfall data for Mpumalanga follow the non-stationary GEVD model, M_8. Table 4.15 presents the K-S and A-D goodness-of-fit test results for Mpumalanga non-stationary GEVD model, M_8.

From Table 4.15, the p-value for the K-S test is insignificant (p $>$ 0.05), implying that the maximum monthly rainfall for Mpumalanga follows the specified non-stationary GEVD model. On the other hand, the results from the A-D test contradict the results from the K-S test.

Table 4.15: Goodness-of-fit for Mpumalanga (1904-2017).

Test	Test Statistic	p-value
K-S	0.08991587	0.2957988
A-D	1.791518	0.0001310069

4.8 Non-stationary GPD modelling of monthly rainfall peaks over a fixed threshold.

This section presents the results of the non-stationary GPD models for monthly rainfall peaks over a fixed threshold. Results for the stationary GPD model, M_0, and non-stationary GPD models M_1 and M_2 are presented. The appropriate threshold is found based on the mean residual life plots and the parameter stability plots. Since the exceedances above the threshold could not be assumed to be independent from each other, declustering of the cluster maxima was performed. The MLE method was used to estimate the parameters of all GPD models.

4.8.1 Eastern Cape

The mean residual life plot is used to help with the identification of a threshold u (Attalides, 2015; Davison and Smith, 1990). The mean residual life plot in Figure 4.21 is not easy to interpret for threshold selection, hence we use the parameter stability plots in Figure 4.22. After examining Figure 4.22, the threshold u = 55 mm was chosen because it is where the parameters appear to stabilise. Therefore, of the 533 threshold exceedances for our Eastern Cape monthly rainfall data we have 129 clusters as illustrated in Figure 4.23 and the extreme observations above the threshold are indicated by the red dots.

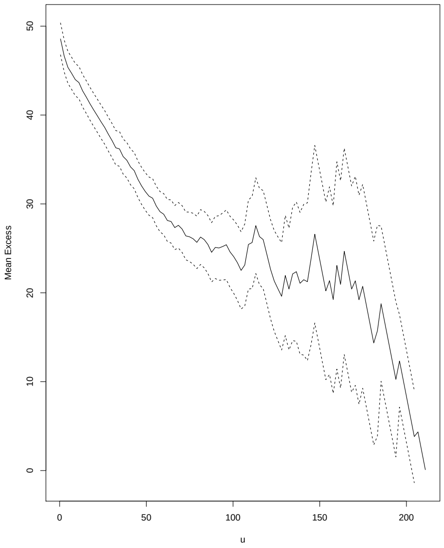

Figure 4.21: Mean residual life plot for the monthly rainfall data for Eastern Cape.

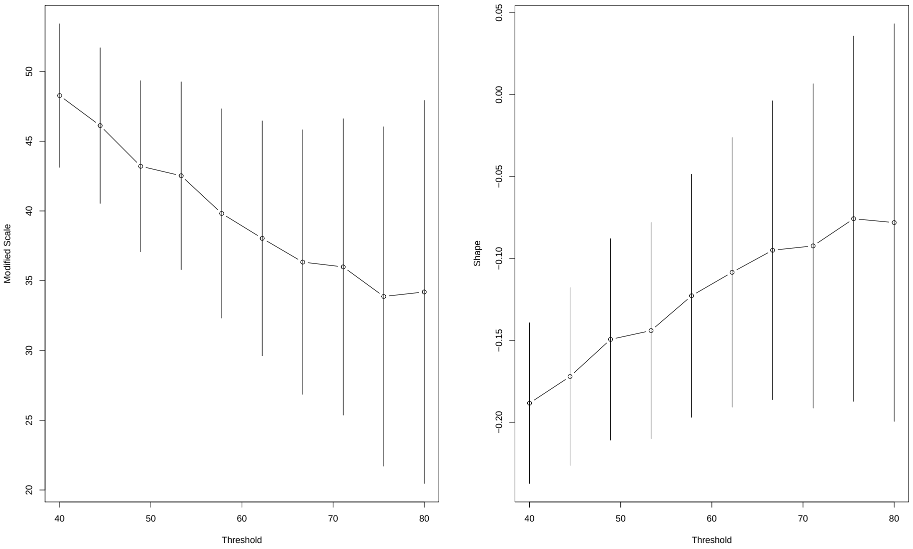

Figure 4.22: Threshold choice or parameter stability plots for Eastern Cape.

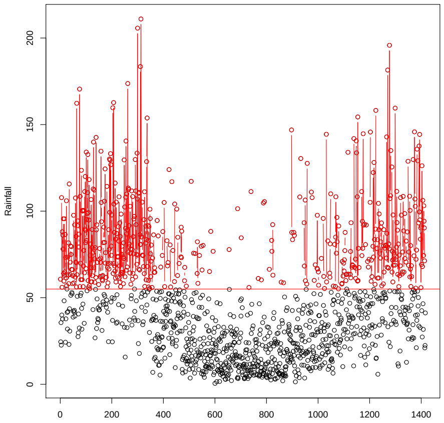

Figure 4.23: The Eastern Cape GPD fitted to cluster maxima (excesses) of the maximum monthly rainfall. The extreme observations above the threshold are indicated by the red dots.

The GPD and non-stationary GPD were applied to monthly rainfall exceedences over a threshold of u=55 mm, and the estimates of the scale and shape parameters of the GPD models are presented in Table 4.16.

The deviance statistic value for the model pair (M_0, M_1) in Table 4.16 is 0.428, which is small relative to $\chi_1^2(0.05) = 3.841$. Thus, there is no significant evidence of a linear trend in the scale parameter of the GPD model.

The nonlinear quadratic model pair (M_0, M_2) from Table 4.16 has a deviance statistic value of 0.520 which is too small compared to the critical value of 5.911 with 2 degrees of freedom. Thus, the non-stationary model, M_2, does not provide an improvement over the stationary GPD model, M_0.

Table 4.16: Parameter estimates and negative log-likelihood of the GPD models for Eastern Cape (1900-2017).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
M_0	34.480	0	0	-0.142	2344.112	(-0.2106,-0.073)
M_1	33.373	0.001	0	-0.141	2343.898	(-0.208,-0.074)
M_2	33.586	0.0002	0.000	-0.140	2343.852	(-0.207,-0.073)

Key: NLLH = negative log-likelihood.

Overall, the best model for Eastern Cape is the stationary GPD model, M_0. The stationary GPD model is given by

$$G(y) = 1 - \left[1 - 0.142\left(\frac{y-55}{32.309}\right)\right]^{\frac{1}{0.142}}. \tag{4.9}$$

The shape parameter (-0.142) is significantly different from zero (p $<$ 0.001) for model M_0, implying that the distribution of exceedances over the 55 mm threshold for Eastern Cape is short-tailed negative Weibull and does not come from exponential distribution family and confidence interval (CI) is signifacantly different from zero. Figure 4.24 shows the diagnostic plots for the stationary GPD model in (4.9). The results of the four diagnostic plots all suggest that the stationary GPD model is a good fit for the Eastern Cape monthly rainfall peaks-over-threshold (POT) data.

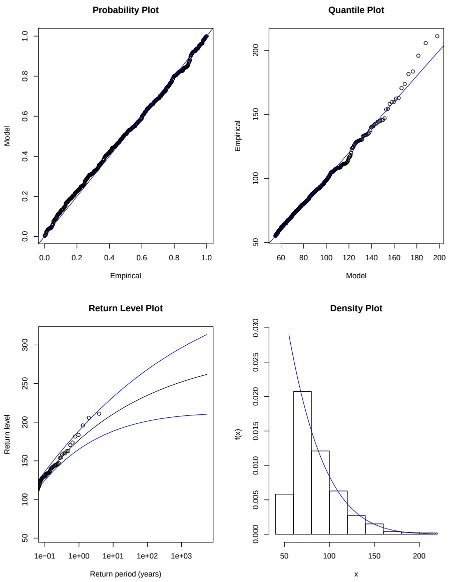

Figure 4.24: Diagnostic plots for the stationary GEV at Eastern Cape.

4.8.2 Gauteng

The mean residual life plot is used to aid the identification of a threshold u (Attalides, 2015; Davison and Smith, 1990). The mean residual life plot in Figure 4.25 is not easy to interpret for threshold selection, hence the parameter stability plots were used. After examining Figure 4.26, the threshold u = 72 mm was chosen because it is where the parameters appear to stabilise. Therefore, of the 521 threshold exceedances for our Gauteng monthly rainfall data, we have 30 clusters as illustrated in Figure 4.27 and the extreme observations above the threshold are indicated by the red dots.

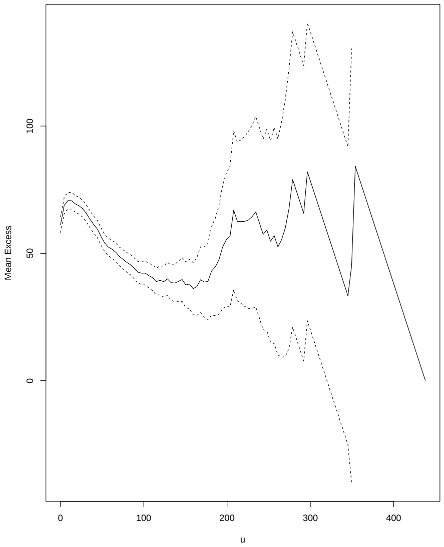

Figure 4.25: Mean residual life plot for the monthly rainfall data for Gauteng.

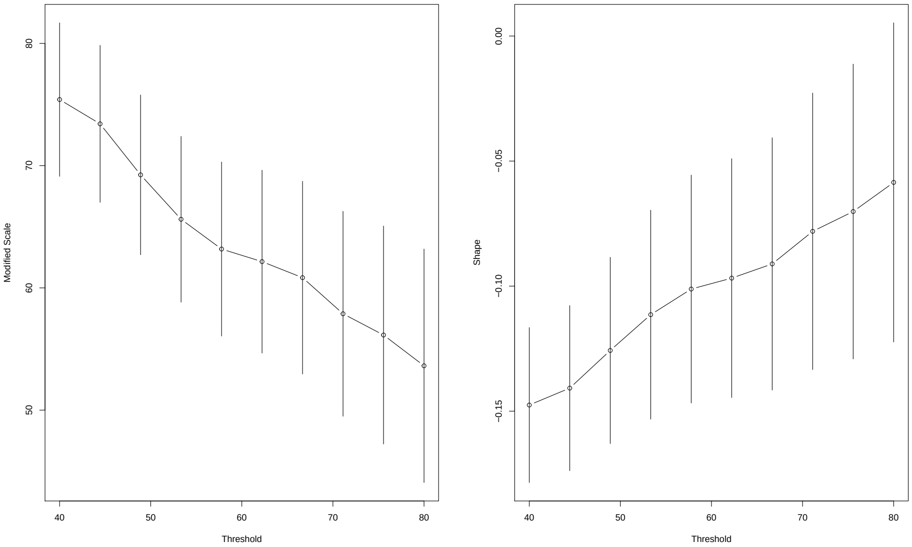

Figure 4.26: Threshold choice or parameter stability plots for Gauteng.

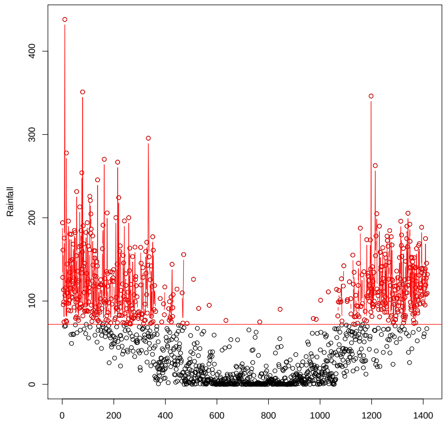

Figure 4.27: The Gauteng GPD fitted to cluster maxima (excesses) of the maximum monthly rainfall. The extreme observations above the threshold are indicated by the red dots.

The GPD and non-stationary GPD were applied to monthly rainfall exceedences over a threshold of u=72 mm, and the estimates of the scale and shape parameters of the GPD models are presented in Table 4.17.

The model pair (M_0, M_1) from Table 4.17 has a deviance statistic of 2.484 which is small in comparison to a critical value of 3.841 with 1 degree of freedom. Thus, there is no significant evidence of a linear trend in the scale parameter of the GPD model.

From Table 4.17, the model pair (M_0, M_2) has a critical value of $\chi_2^2(0.05) = 5.991$, with a deviance statistic value of 22.246. These results show that the nonlinear quadratic trend in the scale parameter is worthwhile over the non-stationary GPD model. The likelihood ratio test for σ_1= 0 it has p-value $<$0.001, and for σ_2= 0 it has p-value $<$ 0.001, which is significant at 5% level of significance (p $<$ 0.05).

Table 4.17: Parameter estimates and negative log-likelihood of the GPD models for Gauteng (1900-2017).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
M_0	52.006	0	0	-0.076	2539.720	(-0.133,-0.019)
M_1	56.688	-0.006	0	-0.092	2538.478	(-0.149,-0.035)
M_2	73.331	-0.124	0.000	-0.097	2528.597	(-0.151,-0.040)

Key: NLLH = negative loglikehood.

The proposed model for Gauteng based on the results is the non-stationary GPD model, M_2, with a nonlinear quadratic trend in the scale parameter. The non-stationary GPD model for Gauteng is given in (4.10)

$$G\left(\sigma(t), \xi; y_t, t\right) \quad = \quad 1 - \left(1 + \frac{-0.097 y_t}{\exp\left(73.331 - 0.124t + 0.000t^2\right)}\right)^{\frac{1}{0.097}}. \quad \textbf{(4.10)}$$

The shape parameter (-0.097) is significantly different from zero (p $<$ 0.001)

for model M_2, implying that the distribution of exceedances over the 72 mm threshold for Gauteng was short-tailed negative Weibull and does not come from exponential distribution family and confidence interval (CI) is signifacantly different from zero. Figure 4.28 shows the diagnostic plots for the non-stationary GPD model in (4.10) and the results suggest a reasonably good fit of the non-stationary GPD with a quadratic trend in the scale parameter for the Gauteng monthly rainfall POT data.

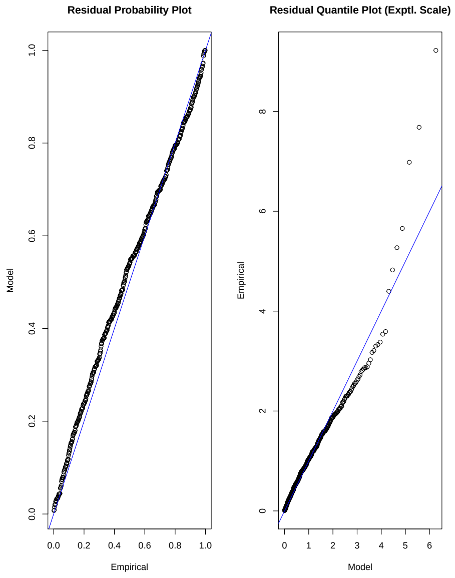

Figure 4.28: Diagnostic plots for the non-stationary GPD model (with a nonlinear quadratic trend in the scale parameter) for Gauteng.

4.8.3 KwaZulu-Natal

The mean residual plot (Figure 4.29) and parameter stability plots in Figures 4.30 were used to come up with a reasonably high threshold of 67 mm for KwaZulu-Natal province which was selected in such a way that it is high enough for the asymptotic theorem to be considered accurate and low enough to have adequate data to estimate the GPD parameters.

After examining Figure 4.30, the threshold u = 67 mm was chosen because it is where the parameters appear to stabilise. Therefore, of the 713 threshold exceedances for our KwaZulu-Natal monthly rainfall data, we have 84 clusters as illustrated in Figure 4.31 and the extreme observations above the threshold are indicated by the red dots.

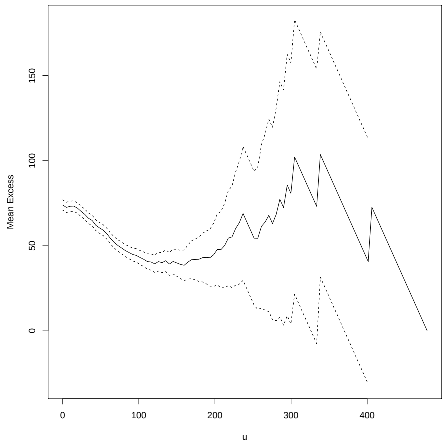

Figure 4.29: Mean residual life plot for the monthly rainfall data for KwaZulu-Natal.

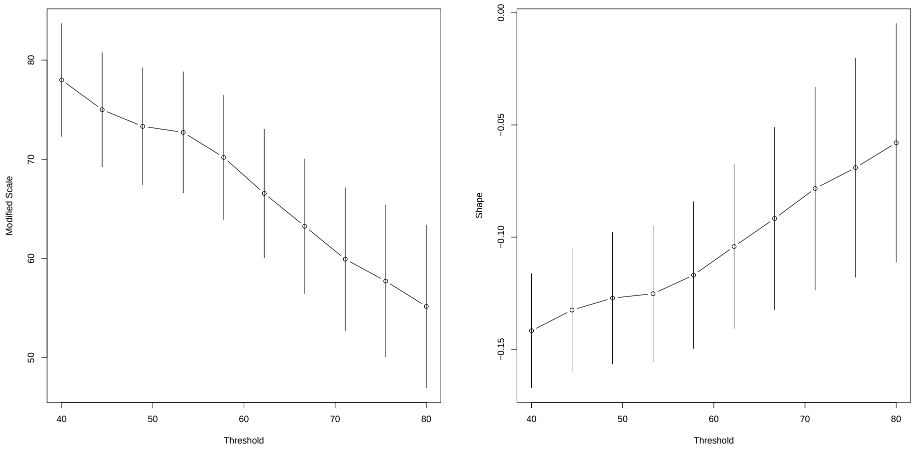

Figure 4.30: Threshold choice or parameter stability plots for KwaZulu-Natal.

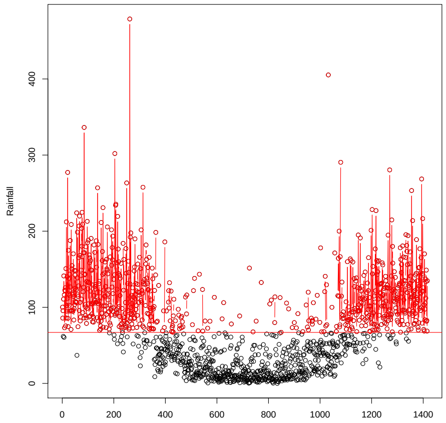

Figure 4.31: The KwaZulu-Natal GPD fitted to cluster maxima (excesses) of the maximum monthly rainfall. The extreme observations above the threshold are indicated by the red dots.

The GPD and non-stationary GPD were applied to monthly rainfall exceedences over a threshold of u=67 mm, and the estimates of the scale and shape parameters of the GPD models are presented in Table 4.18.

Consider the model pair (M_0, M_1) in Table 4.18 with a deviance statistic of 4.496, which is large in comparison to a critical value of 3.841 with 1 degree of freedom. These results reveal overwhelming evidence that the non-stationary GPD model provides an improvement in fit over the stationary GPD model. The likelihood ratio test for σ_1= 0 has p-value= 0.018, which is significant at 5% level of significance ($p < 0.05$) for the linear trend in the log-scale parameter.

The other model pair ((M_0, M_2) from Table 4.18 has a deviance statistic value of 2.806 which is which is small relative to $\chi_2^2(0.05) = 5.991$. Thus, there is no evidence of a nonlinear quadratic trend in the scale parameter of the GPD model.

Table 4.18: Parameter estimates and negative log-likelihood of the GPD models for KwaZulu-Natal (1900-2017).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
M_0	56.807	0	0	-0.090	3528.859	(-0.131,-0.049)
M_1	62.119	-0.007	0	-0.096	3526.611	(-0.137,-0.055)
M_2	60.613	-0.002	0.000	-0.096	3527.456	(-137,-0.055)

Key: NLLH = negative log-likelihood.

The proposed best model for KwaZulu-Natal based on the findings of this study is the non-stationary GPD model, M_1, with a linear trend in the scale parameter. The non-stationary GPD model for KwaZulu-Natal is given by

$$G\left(\sigma(t), \xi; y_t, t\right) = 1 - \left(1 + \frac{-0.096 y_t}{\exp\left(62.119 - 0.007t\right)}\right)^{\frac{1}{0.096}}. \quad \textbf{(4.11)}$$

The shape parameter (-0.096) is significantly different from zero ($p < 0.001$) for model M_1, implying that the distribution of exceedances over the 67 mm threshold at KwaZulu-Natal is short-tailed negative Weibull and does not come from exponential distribution family and confidence interval (CI) is signifacantly different from zero. Figure 4.32 shows the diagnostic plots for the non-stationary GPD model (with a linear trend in the scale parameter) in (4.14). The results of the diagnostic plots for the non-stationary GPD with a linear trend in the scale parameter suggest that the selected non-stationary GPD model is a reasonably good fit for the KwaZulu-Natal POT data.

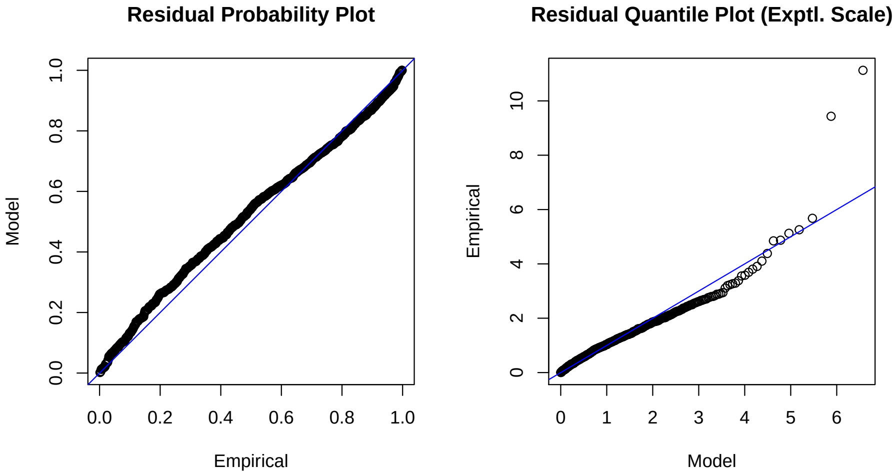

Figure 4.32: Diagnostic plots for the non-stationary GPD model (with the non-linear quadratic trend in the scale parameter) at KwaZulu-Natal.

4.8.4 Limpopo

The mean residual plot (Figure 4.33) and parameter stability plots in Figures 4.34 were used to come up with a reasonably high threshold of 53 mm for Limpopo province which was chosen in such a way that it is high enough for the asymptotic theorem to be considered accurate and low enough to have adequate data to estimate the GPD parameters.

After examining Figure 4.34, the threshold u = 53 mm was chosen because it is where the parameters appear to stabilise. Therefore, of the 631 threshold exceedances for our Limpopo monthly rainfall data, we have 328 clusters as illustrated in Figure 4.35 and the extreme observations above the threshold are indicated by the red dots.

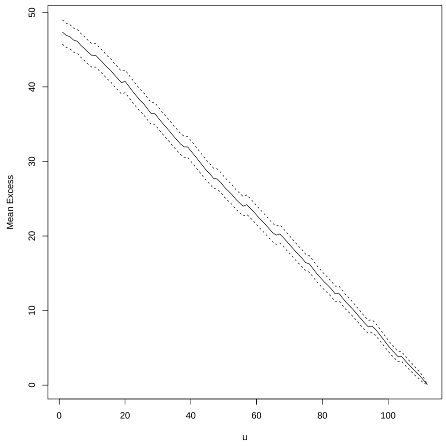

Figure 4.33: Mean residual life plot for the monthly rainfall data for Limpopo.

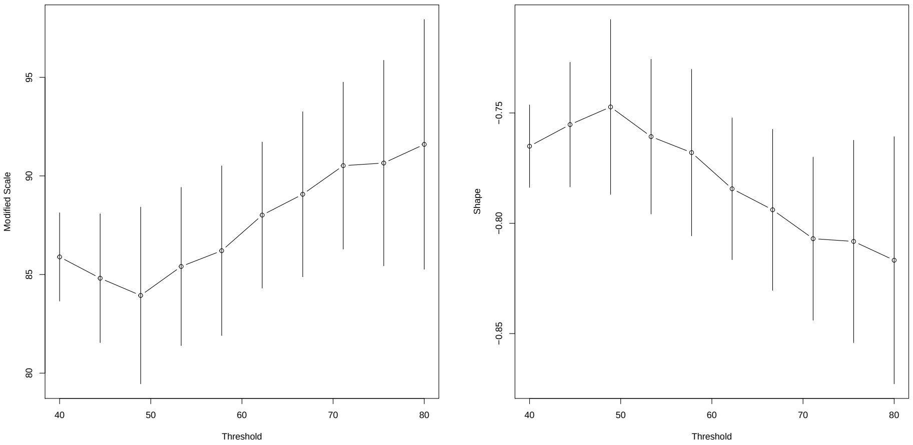

Figure 4.34: Threshold choice or parameter stability plots for Limpopo.

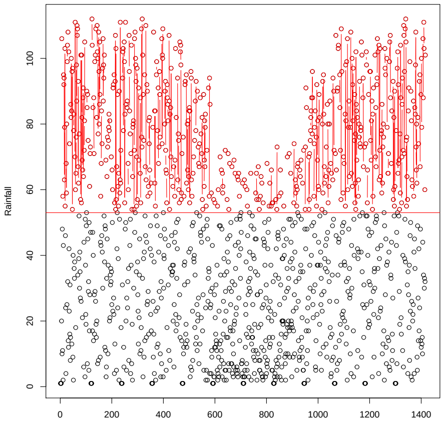

Figure 4.35: The Limpopo GPD fitted to cluster maxima (excesses) of the maximum monthly rainfall. The extreme observations above the threshold are indicated by the red dots.

The GPD and non-stationary GPD were applied to monthly rainfall exceedences over a threshold of u=53 m, and the estimates of the scale and shape parameters of the GPD models are presented in Table 4.19.

From Table 4.19, the deviance statistic of model pair (M_0, M_1) is 0.138 which is too small compared to the critical value of 3.841 with 1 degree of freedom. Thus, there is no evidence of a linear trend in the scale parameter of the GPD model.

Consider the model pair (M_0, M_2) in Table 4.19 which has a critical value of $\chi^2_2(0.05) = 5.991$, with the deviance statistic value of 17.980. These results show that the nonlinear quadratic trend in the scale parameter is worthwhile over the stationary GPD model. The likelihood ratio test for σ_1= 0 has p-value <0.001, and for σ_2= 0 has p-value $<$ 0.001, which is significant at 5% level of significance (p $<$ 0.05).

Table 4.19: Parameter estimates and negative log-likelihood of the GPD models for Limpopo (1904-2017).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
M_0	45.886	0	0	-0.775	2386.330	(-0.801,-0.750)
M_1	45.891	-0.0001	0	-0.773	2386.261	(-0.773,-0.772)
M_2	51.378	-0.003	0.000	-0.749	2377.340	(-0.759,-0.739)

Key: NLLH = negative log-likelihood.

In general, the best model for Limpopo based on the findings of this study, is the non-stationary GPD model, M_2. The non-stationary GPD model for Limpopo is given by

$$G\left(\sigma(t), \xi; y_t, t\right) = 1 - \left(1 + \frac{-0.749 y_t}{\exp\left(51.378 - 0.003t + 0.000t^2\right)}\right)^{\frac{1}{0.749}}. \quad \textbf{(4.12)}$$

The shape parameter (-0.749) is significantly different from zero (p $<$ 0.001)

for model M_2, indicating that the distribution of exceedances over the 53 mm threshold for Limpopo is short-tailed negative Weibull and does not come from exponential distribution family and confidence interval (CI) is signifacantly different from zero. Figure 4.36 shows the diagnostic plots for the non-stationary GPD model (with a nonlinear quadratic trend in the scale parameter) in (4.12). The results of the diagnostic plots for the non-stationary GPD with a linear trend in the scale parameter suggest that the selected non-stationary GPD model is a reasonably good fit for the limpopo POT data.

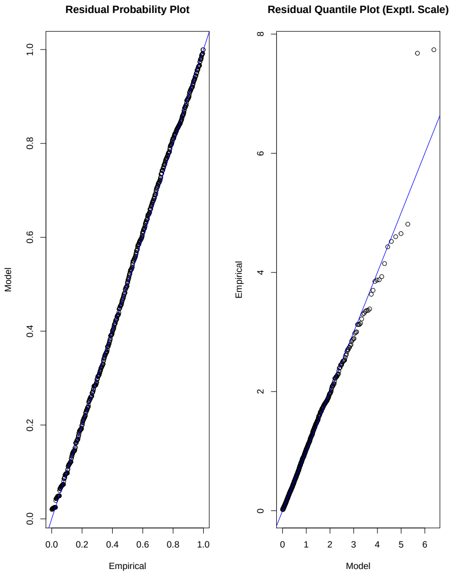

Figure 4.36: Diagnostic plots for the non-stationary GPD model (with the non-linear quadratic trend in the scale parameter) for Limpopo.

4.8.5 Mpumalanga

The mean residual plot in Figure 4.37 and parameter stability plots in Figures 4.38, were used to come up with a reasonably high threshold of 50 mm for Mpumalanga province which was selected in such a way that it is high enough for the asymptotic theorem to be considered accurate and low enough to have adequate data to estimate the GPD parameters.

After examining Figure 4.38, the threshold u = 50 mm was chosen because it is where the parameters appear to stabilise. Therefore, of the 658 threshold exceedances for our Mpumalanga monthly rainfall data, we have 313 clusters as illustrated in Figure 4.39 and the extreme observations above the threshold are indicated by the red dots.

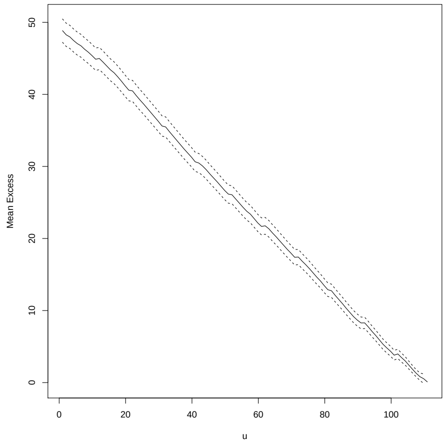

Figure 4.37: Mean residual life plot for the monthly rainfall data for Mpumalanga.

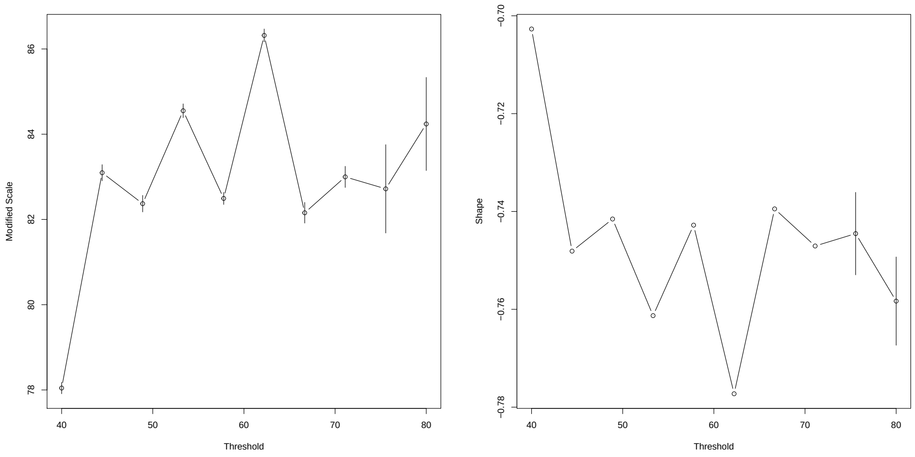

Figure 4.38: Threshold choice or parameter stability plots for Mpumalanga.

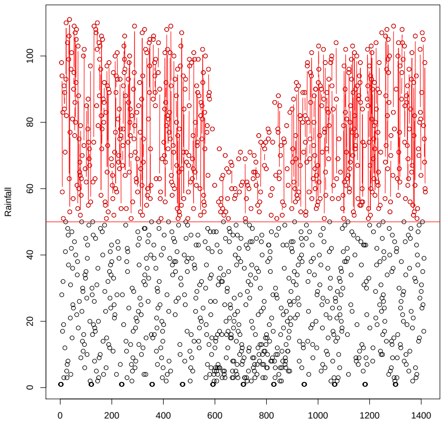

Figure 4.39: The Mpumalanga GPD fitted to cluster maxima (excesses) of the maximum monthly rainfall. The extreme observations above the threshold are indicated by the red dots.

The GPD and non-stationary GPD were applied to monthly rainfall exceedences over a threshold of u=50 mm, and the estimates of the scale and shape parameters of the GPD models are presented in Table 4.20.

Consider the model pair (M_0, M_1) in Table 4.20 with a deviance statistic of 10.276, which is large in comparison to a critical value of 3.841 with 1 degree of freedom. These results reveal overwhelming evidence that the non-stationary GPD model provides an improvement in fit over the stationary GPD model. The likelihood ratio test for σ_1= 0 has p-value$<$ 0.001, which is significant at 5% level of significance (p $<$ 0.05) for the linear trend in the scale parameter.

The nonlinear quadratic model pair (M_0, M_2) from Table 4.20 has a deviance statistic value of -21.968 which is too small compared to the critical value of 5.911 with 2 degrees of freedom. Thus, the non-stationary model M_2 does not provide an improvement over the stationary GPD model, M_0.

Table 4.20: Parameter estimates and negative log-likelihood of the GPD models for Mpumalanga (1904-2017).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
M_0	47.591	0	0	-0.779	2686.836	(-0.779,-0.778)
M_1	49.961	-0.001	0	-0.817	2681.698	(-0.816,-0.815)
M_2	45.584	0.014	0.000	-0.751	2697.820	(-0.751,-0.750)

Key: NLLH = negative log-likelihood.

Overall, the best model for Mpumalanga based on the results of this study is the non-stationary GPD model, M_1, with a linear trend in the scale parameter. The non-stationary GPD model for Mpumalanga is given in (4.13)

$$G\left(\sigma(t), \xi; y_t, t\right) \quad = \quad 1 - \left(1 + \frac{-0.817 y_t}{\exp\left(49.961 - 0.001t\right)}\right)^{\frac{1}{0.817}}. \qquad \textbf{(4.13)}$$

The shape parameter (-0.817) is significantly different from zero (p $<$ 0.001)

for model M_1, indicating that the distribution of exceedances over the 50 mm threshold for Mpumalanga is short-tailed negative Weibull and does not come from exponential distribution family and confidence interval (CI) is signifacantly different from zero. Figure 4.40 shows the diagnostic plots for the non-stationary GPD model (with a linear trend in the scale parameter) in (4.13). The results from Figure 4.40 reveal a reasonably good fit of the non-stationary GPD model to the Mpumalanga POT data.

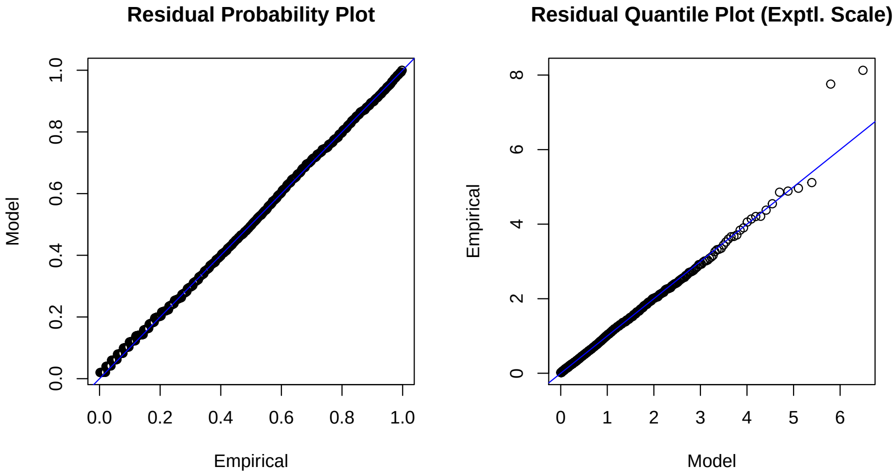

Figure 4.40: Diagnostic plots for the non-stationary GPD model (with a linear trend in the scale parameter) at Mpumalanga.

4.9 Modelling monthly rainfall data using a GPD with time-varying threshold

4.9.1 Eastern Cape

Figure 4.41 is a time series plot of the monthly rainfall data for Eastern Cape with a time-varying threshold, which is a penalised cubic smoothing spline. The smoothing parameter lamda (λ) is selected based on the generalised cross validation (GCV) criterion. The estimated value for λ is $\hat{\lambda}$ = 0.00003596222. We then determine a sufficiently high threshold by fitting a non-parametric extremal mixture model and exceedances are declustered using Ferro and Segers (2003) intervals estimator method. Figure 4.42 shows threshold estimation using a non-parametric extremal mixture model where a kernel density is fitted to the bulk model and a GPD fitted to the upper end-point of the model, with the vertical line indicating the estimated threshold. The estimated threshold is u=127.600 for the Eastern Cape GPD.

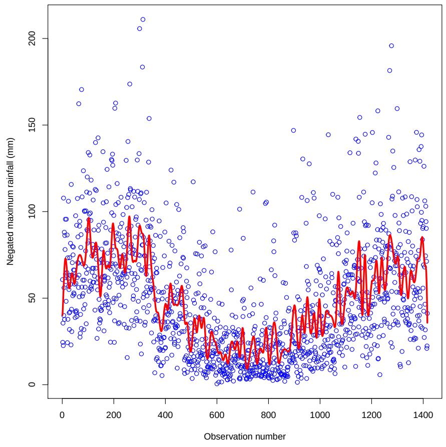

Figure 4.41: A time series plot of the Eastern Cape monthly rainfall data with a time-varying threshold, which is a penalised cubic smoothing spline. Blue dots show the negated observations and the red line shows the smoothing parameter.

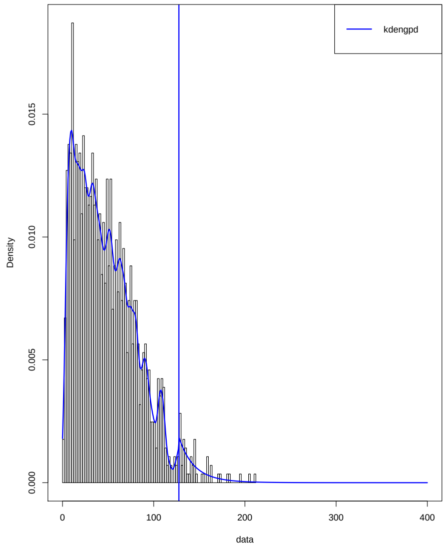

Figure 4.42: Threshold estimation for Eastern Cape using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the tail of the distribution (u=127.600).

The GPD model was fitted to cluster maxima using the threshold estimated by the non-parametric extremal mixture model. Table 4.21 presents the maximum likelihood estimates of GPD parameters and the estimated threshold for Eastern Cape.

Table 4.21: Parameter estimates of Eastern Cape GPD fitted to cluster maxima of the monthly rainfall.

Threshold	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
127.600	18.983 (4.641)	0.052 (0.195)	183.810	(-0.330,0.434)

Key: NLLH = negative log-likelihood.

From Table 4.21, scale and shape parameters are found to be 18.983 and 0.052, respectively, with standard errors in parentheses. To ensure that the scale parameter is positive (σ >0), we use the transformation, $\sigma = e^{18.983}$. The positive value of the shape parameter ($\hat{\xi}$=0.052) indicates that the monthly rainfall data follows a Pareto distribution. The 95% confidence interval (CI) is not signifacantly different from zero, indicating that the monthly rainfall data for Eastern Cape can be modelled by the exponential family of distribution. The diagnostic plots in Figure 4.43 show an appropriate fit of the GPD with time-varying threshold for the Eastern Cape province.

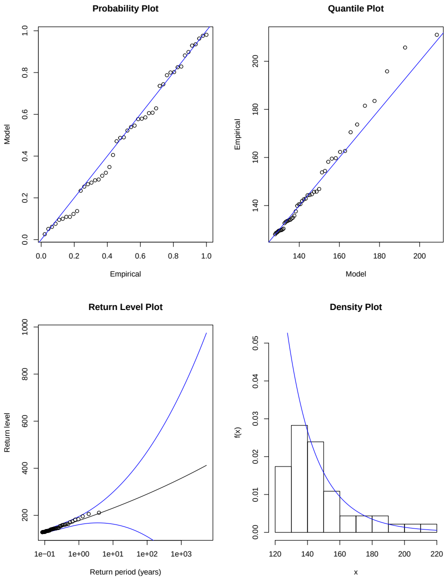

Figure 4.43: Diagnostic plots for the Eastern Cape GPD fitted to cluster maxima.

4.9.2 Gauteng

Figure 4.44 is a time series plot of the monthly rainfall data for Gauteng with a time-varying threshold, which is a penalised cubic smoothing spline. The smoothing parameter lamda (λ) is selected based on the GCV criterion. The estimated value for λ is $\hat{\lambda}$ = 0.00001716462. An initial threshold is set at zero after fitting time-varying threshold and only positive observations (excesses) above zero are considered. We then determine a sufficiently high threshold by fitting a non-parametric extremal mixture model and exceedances are declustered using Ferro and Segers (2003) intervals estimator method.

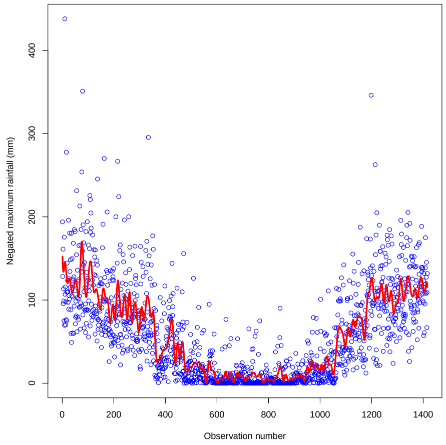

Figure 4.44: A time series plot of the Gauteng monthly rainfall data with a time-varying threshold, which is a penalised cubic smoothing spline. Blue dots show the negated observations and the red line shows the smoothing parameter.

The monthly rainfall for Gauteng is initial detrended using a penalised cubic smoothing spline in (3.44). The kernel density technique was applied to estimate the value of the threshold, u =122.892. Figure 4.45 shows threshold estimation using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the upper end-point of the model. The estimated threshold is u=122.892 for the Gauteng GPD

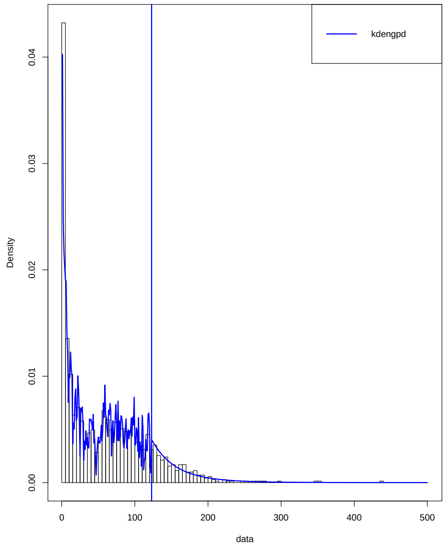

Figure 4.45: Threshold estimation using a non-parametric extremal mixture model where a kernel density is fitted to the bulk model and a GPD fitted to the tail of the distribution (u=122.892).

The GPD model was fitted to cluster maxima using the threshold estimated by the non-parametric extremal mixture model. Table 4.22 presents maximum likelihood estimates of GPD parameters and the estimated threshold for Gauteng.

Table 4.22: Parameter estimates of Gauteng GPD fitted to cluster maxima of the monthly rainfall.

Threshold	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
122.892	34.888 (3.502)	0.084 (0.071)	927.166	(-0.055,0.223)

Key: NLLH = negative log-likelihood.

From Table 4.22, scale and shape parameters are found to be 34.888 and 0.084, respectively, with standard errors in parenthesis. To ensure that the scale parameter is positive (σ >0), we use the transformation, $\sigma = e^{34.888}$. The positive value of the shape parameter ($\hat{\xi}$=0.084) indicates that the monthly rainfall data follows a Pareto distribution.The 95% confidence interval (CI) is not signifacantly different from zero, indicating that the monthly rainfall data for Gauteng can be modelled by the exponential family of distribution. The diagnostic plots in Figure 4.46 show an appropriate fit of the GPD with time-varying threshold for the Gauteng province.

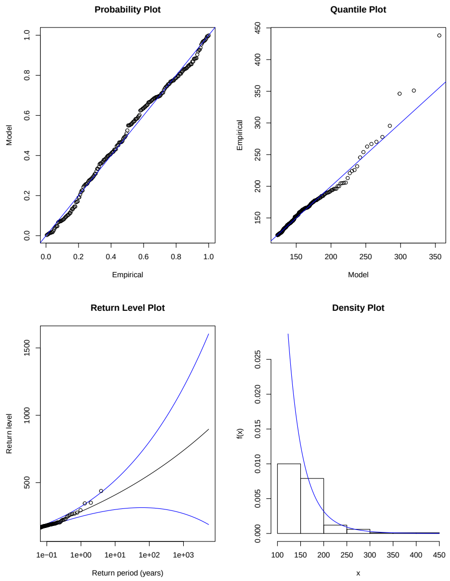

Figure 4.46: Diagnostic plots for the Gauteng GPD fitted to cluster maxima.

4.9.3 KwaZulu-Natal

Figure 4.47 is a time series plot of the monthly rainfall data for KwaZulu-Natal with a time-varying threshold, which is a penalised cubic smoothing spline. The smoothing parameter lamda (λ) is selected based on the GCV criterion. The estimated value for λ is $\hat{\lambda}$ = 0.002645513. The first threshold is set at zero after fitting time-varying threshold and only positive observations (excesses) above zero are considered. We then determine a sufficiently high threshold by fitting a non-parametric extremal mixture model and exceedances are declustered using Ferro and Segers (2003) intervals estimator method. Figure 4.48 shows threshold estimation using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the upper end-point of the model. The estimated threshold is u=148.499 for KwaZulu-Natal GPD.

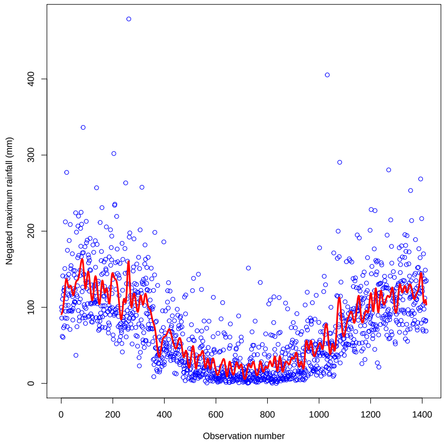

Figure 4.47: A time series plot of the KwaZulu-Natal monthly rainfall data with a time-varying threshold, which is a penalised cubic smoothing spline. Blue dots show the negated observations and the red line shows the smoothing parameter.

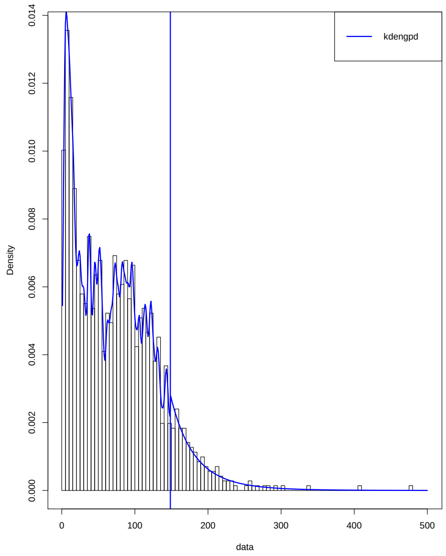

Figure 4.48: Threshold estimation for KwaZulu-Natal using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the tail of the distribution (u=148.499).

The GPD model was fitted to cluster maxima using the threshold estimated by the non-parametric extremal mixture model. Table 4.23 shows maximum likelihood estimates of GPD parameters and the estimated threshold for KwaZulu-Natal.

Table 4.23: Parameter estimates of KwaZulu-Natal GPD fitted to cluster maxima of the monthly rainfall.

Threshold	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
148.499	34.929 (4.196)	0.116 (0.086)	663.056	(-0.053,0.285)

Key: NLLH = negative log-likelihood.

From Table 4.23, scale and shape parameters are found to be 34.929 and 0.116, respectively, with standard errors in parentheses. To ensure that the scale parameter is positive (σ >0), we use the transformation, $\sigma = e^{34.929}$. The positive value of the shape parameter ($\hat{\xi}$=0.116) indicates that the monthly rainfall data follows a Pareto distribution. The 95% confidence interval (CI) is not signifacantly different from zero, indicating that the monthly rainfall data for KwaZulu-Natal can be modelled by the exponential family of distribution. The diagnostic plots in Figure 4.49 show an appropriate fit of the GPD with a time-varying threshold for the KwaZulu-Natal province.

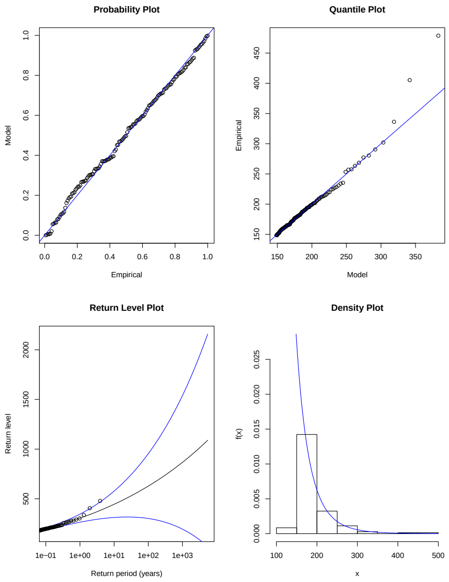

Figure 4.49: Diagnostic plot for the KwaZulu-Natal GPD fitted to cluster maxima.

4.9.4 Limpopo

Figure 4.50 is a time series plot of the monthly rainfall data for Limpopo with a time-varying threshold, which is a penalised cubic smoothing spline. The smoothing parameter lamda (λ) is selected based on the GCV criterion. The estimated value for λ is $\hat{\lambda}$ = 0.009853259. The first threshold is set at zero after fitting time-varying threshold and only positive observations (excesses) above zero are considered. We then determine a sufficiently high threshold by fitting a non-parametric extremal mixture model and exceedances are declustered using Ferro and Segers (2003) intervals estimator method. Figure 4.51 shows threshold estimation using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the upper end-point of the model. The estimated threshold is u=93 for Limpopo GPD.

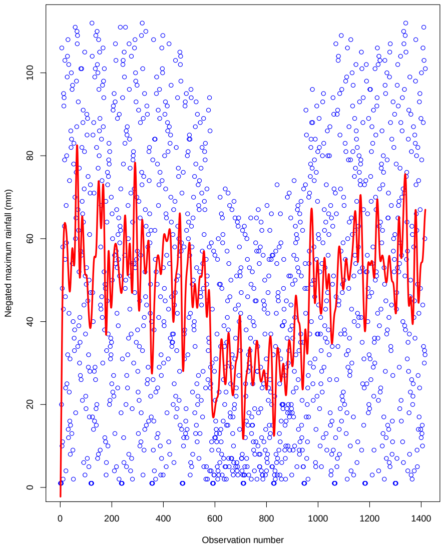

Figure 4.50: A time series plot of the Limpopo monthly rainfall data with a time-varying threshold, which is a penalised cubic smoothing spline. Blue dots show the negated observations and the red line shows the smoothing parameter.

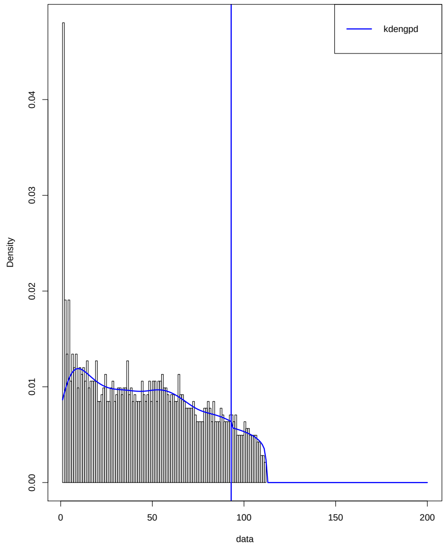

Figure 4.51: Threshold estimation for Limpopo using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the tail of the distribution (u=93).

The GPD model was fitted to cluster maxima using the threshold estimated by the non-parametric extremal mixture model. Table 4.24 presents the maximum likelihood estimates of GPD parameters and the estimated threshold for Limpopo.

Table 4.24: Parameter estimates of GPD fitted to cluster maxima of the monthly rainfall.

Threshold	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**	95% CI for ξ
93	16.577 (0.858)	-0.868 (0.044)	390.978	(-0.954,-0.782)

Key: NLLH = negative log-likelihood.

From Table 4.24, scale and shape parameters are found to be 34.929 and 0.116, respectively, with standard errors in parentheses. To ensure that the scale parameter is positive (σ >0), we use the transformation, $\sigma = e^{16.577}$. A negative value of the shape parameter reveals evidence that the monthly rainfall data for Limpopo belongs to Weibull family and confidence interval (CI) is signifacantly different from zero. The diagnostic plots in Figure 4.52 show an appropriate fit of the GPD with a time-varying threshold for the Limpopo province.

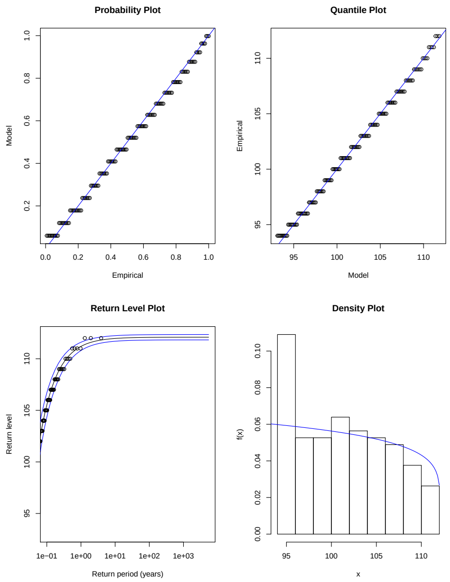

Figure 4.52: Diagnostic plots for the Limpopo GPD fitted to cluster maxima.

4.9.5 Mpumalanga

Figure 4.53 is a time series plot of the monthly rainfall data for Mpumalanga with a time-varying threshold, which is a penalised cubic smoothing spline. The smoothing parameter lamda (λ) is selected based on the GCV criterion. The estimated value for λ is $\hat{\lambda}$ = 0.00001018565. An initial threshold is set at zero after fitting time-varying threshold and only positive observations (excesses) above zero are considered. We then determine a sufficiently high threshold by fitting a non-parametric extremal mixture model and exceedances are declustered using Ferro and Segers (2003) intervals estimator method.

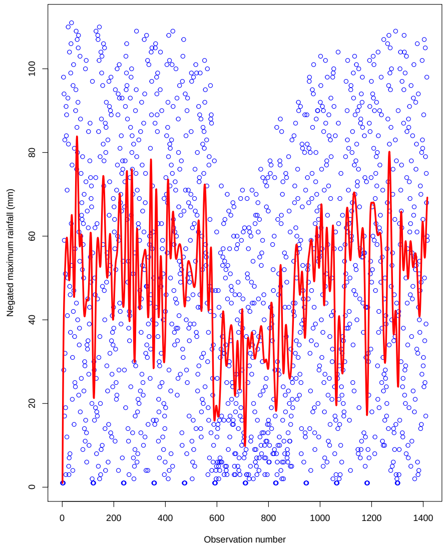

Figure 4.53: A time series plot of the Mpumalanga monthly rainfall data with a time-varying threshold, which is a penalised cubic smoothing spline. Blue dots show the negated observations and the red line shows the smoothing parameter.

The monthly rainfall for Mpumalanga is initial detrended using a penalised cubic smoothing spline in (3.44). The kernel density technique was applied to estimate the value of the threshold, u =92.996. Figure 4.54 shows threshold estimation using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the upper end-point of the model. The estimated threshold is u=92.996 for Mpumalanga GPD.

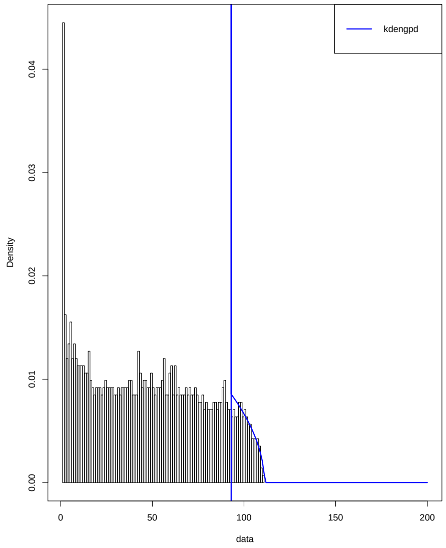

Figure 4.54: Threshold estimation for Mpumalanga using a non-parametric extremal mixture model, where a kernel density is fitted to the bulk model and a GPD fitted to the tail of the distribution (u=92.996).

The GPD model is fitted to cluster maxima using the threshold estimated by the non-parametric extremal mixture model. Table 4.25 shows maximum likelihood estimates of GPD parameters and the estimated threshold for the Mpumalanga province.

Table 4.25: Parameter estimates of Mpumalanga GPD fitted to cluster maxima of the monthly rainfall.

Threshold	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**	95 % CI for ξ
92.996	12.240 (1.034)	-0.673 (0.060)	410.595	(-0.791,-0.555)

Key: NLLH = negative log-likelihood.

From Table 4.25, scale and shape parameters are found to be 12.240 and -0.673, respectively, with standard errors in parentheses. To ensure that the scale parameter is positive (σ >0), we use the transformation, $\sigma = e^{12.240}$. A negative value of the shape parameter reveals evidence that the monthly rainfall data for Mpumalanga belongs to Weibull family of distribution and confidence intervals (CI) is significantly different from zero. The diagnostic plots in Figure 4.55 show an appropriate fit of the GPD with a time-varying threshold for the Mpumalanga province.

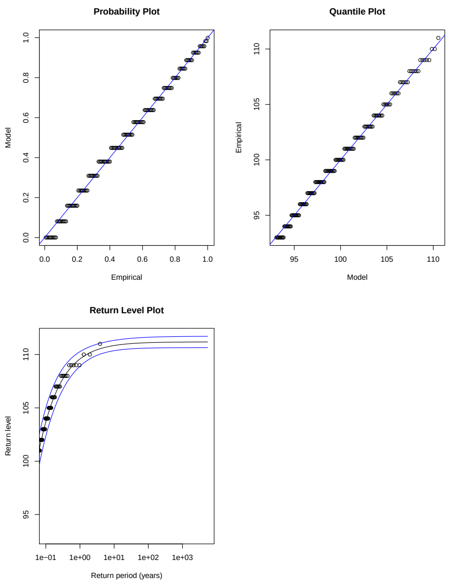

Figure 4.55: Diagnostic plots for the Mpumalanga GPD fitted to cluster maxima.

Chapter 5

Conclusion

5.1 Introduction

Over the past few decades, floods have been the most common and serious disasters in most countries worldwide, including South Africa. These disasters are mainly caused by the occurrence of extreme maximum rainfall. This dissertation can contribute towards understanding the environment and climate change in general. The aim of this study was to analyse the monthly rainfall data obtained from the South African Weather Service (SAWS) using various statistical techniques.

This dissertation sets out to model monthly rainfall data in selected provinces of South Africa using extreme value distributions. The five provinces considered in this study are: Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo and Mpumalanga. This chapter summarises the major findings of the study based on the data analysed. The chapter also gives recommendations for future work in the field of extreme value theory.

5.2 Dissertation summary

This study investigated five candidate parent distributions: gamma, Gumbel, log-normal, Pareto and Weibull, to establish the best-fit probability distribution for each of the five provinces. Augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) statistical tests were used to test for stationarity. Non-parametric Mann-Kendall (M-K) test and time series plots were used to investigate the long-term trends of the monthly rainfall and their variability across the selected provinces. The study also employed Jarque-Bera (JB), Shapiro–Wilk (SW) and chi-square tests methods to check whether the monthly rainfall data were normally distributed.

This research used non-stationary generalised extreme value distribution (GEVD) and non-stationary generalised Pareto distribution (GPD) with both fixed and time-varying thresholds in modelling extreme maximum monthly rainfall data for the five provinces. The deviance statistic and likelihood ratio test were used to select the best-fit model among non-stationary GEVD and non-stationary GPD families, while the maximum likelihood estimation method was used to obtain the estimates of the parameters. Model adequacy was checked using Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests.

5.3 Conclusion

This section summarises the concluding remarks of the analyses that were performed in Chapter 4. Firstly, the Weibull distribution provided the best-fit probability parent distribution for Eastern Cape, KwaZulu-Natal, Limpopo and Mpumalanga provinces based on the value of the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), while the best-fit probability parent distribution for Gauteng province was found to be the gamma distribution.

The p-values of the ADF test statistics for Eastern Cape, Limpopo and Mpumalanga are significant ($p < 0.05$), suggesting that the monthly rainfall data for these three provinces are stationary. The ADF p-values for Gauteng and KwaZulu-Natal are insignificant ($p > 0.05$), suggesting that the monthly rainfall data for these two provinces are not stationary at 5% level of significance, while the p-values of the KPSS test for all five provinces are significant ($p < 0.05$), suggesting that the monthly rainfall data are not stationary. Furthermore, the p-values of the PP test for all five provinces are significant ($p < 0.05$), implying that the monthly rainfall data are stationary. Findings from JB, SW and chi-square normality tests revealed that the monthly rainfall data do not come from a normal distribution. The findings of the Mann-Kendall trend test suggested that in Eastern Cape, Gauteng and Kwazulu-Natal provinces there was a significant monotonic decreasing trend, while in Limpopo and Mpumalanga provinces there was an insignificant monotonic decreasing trend.

The study presented an application of non-stationary GEVD in modelling maximum monthly rainfall data. The stationary GEVD was found as the best distribution model for Eastern Cape, Gauteng and KwaZulu-Natal provinces. Furthermore, model fitting supported non-stationary GEVD models for maximum monthly rainfall with nonlinear quadratic trend in the location parameter and a linear trend in the scale parameter for Limpopo, while in Mpumalanga the non-stationary GEVD model, which has a nonlinear quadratic trend in the scale parameter and no variation in the location parameter fitted well to the monthly rainfall data.

The negative values of the shape parameters for Eastern Cape and Mpumalanga, indicate that the data follow the Weibull distribution class, while the positive values of the shape parameters for Gauteng, KwaZulu-Natal and Limpopo, suggest that the data follow the Fréchet distribution class.

The study also presented an application of non-stationary GPD with a fixed threshold in modelling monthly rainfall excesses data. The stationary GPD provided the best-fit model for Eastern Cape, while the non-stationary GPD model with a linear trend in the scale parameter was found as the best distribution for KwaZulu-Natal and Mpumalanga provinces and the non-stationary GPD model with a nonlinear quadratic trend in the scale parameter was found as the best distribution for Gauteng and Limpopo. The shape parameters of the stationary GPD and non-stationary GPD were all negative, suggesting that the distribution of exceedances above the predetermined thresholds in the selected provinces is short-tailed negative Weibull distribution family.

The study further investigated a GPD with time-varying thresholds in modelling monthly rainfall excesses data. The data were detrended using cubic regression smoothing spline and the GPD was fitted with the threshold that was estimated using the non-parametric extremal mixture models. Findings indicate that the monthly rainfall data for Eastern Cape, Gauteng and KwaZulu-Natal comes from the exponential distribution family. On the other hand, for Limpopo and Mpumalanga, evidence suggested that the monthly rainfall data belong to the Weibull family.

5.4 Limitations of the dissertation

The monthly rainfall data were obtained from the South African Weather Service (SAWS) for the period 1900-2017. However, monthly rainfall data for Limpopo and Mpumalanga were recorded for the period 1904 up to 2017. This research focused only on five selected provinces and this was also considered as the study limits because it is important to conduct a study with all the provinces of South Africa. Another limitation of the study concerns scarce literature on modelling rainfall data using extreme value theory techniques with time-varying threshold in South Africa.

5.5 Recommendations for future work

Future research work could consider using the Bayesian estimation method to obtain estimates of the GEVD and GPD parameters. Future studies could also consider including covariates such as Southern Oscillation Index (SOI) to model monthly rainfall, as well as extending the study to cover all the nine provinces of South Africa. Future research could also explore the use of multivariate extreme value theory (MEVT) including spatial extremes in analysing rainfall data.

References

ACERO, F. J., GALLEGO, M. C., AND GARCÍA, J. A. (2012). Multi-day rainfall trends over the Iberian Peninsula. *Theoretical and Applied Climatology*, **108** (3-4), 411–423.

ACERO, F. J., GARCÍA, J. A., AND GALLEGO, M. C. (2011). Peaks-over-threshold study of trends in extreme rainfall over the Iberian Peninsula. *Journal of Climate*, **24** (4), 1089–1105.

ACQUAH, H. D.-G. (2010). Comparison of Akaike information criterion AIC and Bayesian information criterion BIC in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, **2** (1), 001–006.

ACQUAH, H. D.-G. (2012). A bootstrap approach to evaluating the performance of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Agricultural Sciences, Belgrade*, **57** (2), 99–110.

ADEFISOYE, J., GOLAM KIBRIA, B., AND GEORGE, F. (2016). Performances of several univariate tests of normality: An empirical study. *Journal of Biometrics and Biostatistics*, **7**.

AKSOY, H. (2000). Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, **24** (6), 419–428.

ALAM, M., EMURA, K., FARNHAM, C., AND YUAN, J. (2018). Best-fit probability distributions and return periods for maximum monthly rainfall in Bangladesh. *Climate*, **6** (1), 1–16.

ALAM, M., TORIMAN, M., SIWAR, C., AND TALIB, B. (2011). Rainfall variation and changing pattern of agricultural cycle. *American Journal of Environmental Sciences*, **7** (1), 82–89.

ALEXANDER, M. (2018). South Africa's weather and climate. Last accessed: 03.03.2020.
URL: *https://southafrica-info.com/land/south-africa-weather-climate/*

AMIN, M., RIZWAN, M., AND ALAZBA, A. (2016). A best-fit probability distribution for the estimation of rainfall in northern regions of Pakistan. *Open Life Sciences*, **11** (1), 432–440.

ATTALIDES, N. (2015). *Threshold-based extreme value modelling*. Ph.D. thesis, University College London.

BENSALAH, Y. (2000). *Steps in applying extreme value theory to finance: a review*. Bank of Canada.

BHARTI, V. (2015). *Investigation of extreme rainfall events over the northwest Himalaya Region using satellite data*. Ph.D. thesis, University of Twente.

BOTAI, C. M., BOTAI, J. O., AND ADEOLA, A. M. (2018). Spatial distribution of temporal precipitation contrasts in South Africa. *South African Journal of Science*, **114** (7-8), 70–78.

BOUDRISSA, N., CHERAITIA, H., AND HALIMI, L. (2017). Modelling maximum daily yearly rainfall in northern Algeria using generalized extreme value distributions from 1936 to 2009. *Meteorological Applications*, **24** (1), 114–119.

CHEGE, C. K., MUNGAT'U, J. K., AND NGESA, O. (2016). Estimating the extreme financial risk of the Kenyan Shilling versus US Dollar exchange rates. *Science Journal of Applied Mathematics and Statistics*, **4** (6), 249–255.

CHIKOBVU, D. AND CHIFURIRA, R. (2015). Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, **111** (9-10), 1–8.

CHIKODZI, D., MURWENDO, T., AND SIMBA, F. M. (2013). Climate change and variability in southeast Zimbabwe: Scenarios and societal opportunities. *American Journal of Climate Change*, **2** (3), 36–49.

CHU, L., MCALEER, M., AND CHANG, C.-H. (2013). *Statistical modelling of extreme rainfall in Taiwan*. Technical report. Atlantis Press.

CHU, P.-S., ZHAO, X., RUAN, Y., AND GRUBBS, M. (2009). Extreme rainfall events in the Hawaiian islands. *Journal of Applied Meteorology and Climatology*, **48** (3), 502–516.

COLES, S., BAWA, J., TRENNER, L., AND DORAZIO, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer. London.

CONNAUGHTON, C., HERMAN, J., JOHANSEN, A., KAWABATA, E., KERR, R., PEGG, M., REIZENSTEIN, J., SAKRAJDA, P., TAWN, N., AND WHINCOP, L. (2017). *African Drought Risk Pay-Out Benchmarking*. ESGI30, Univesity of Warwick.

DA SILVA, R. M., SANTOS, C. A., MOREIRA, M., CORTE-REAL, J., SILVA, V. C., AND MEDEIROS, I. C. (2015). Rainfall and river flow trends using Mann-Kendall and Sen's slope estimator statistical tests in the Cobres River basin. *Natural Hazards*, **77** (2), 1205–1221.

DAS, K. R. AND IMON, A. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, **5** (1), 5–12.

DAVISON, A. C. AND SMITH, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, **52** (3), 393–425.

DE WAAL, J. H. (2012). *Extreme rainfall distributions: Analysing change in the Western Cape*. Ph.D. thesis, Stellenbosch University.

DE WAAL, J. H., CHAPMAN, A., AND KEMP, J. (2017). Extreme 1-day rainfall distributions: Analysing change in the Western Cape. *South African Journal of Science*, **113** (7-8), 1–8.

DIRIBA, T. A., DEBUSHO, L. K., AND BOTAI, J. (2015). *Modeling extreme daily temperature using generalized Pareto distribution at Port Elizabeth, South Africa*. In: Annual Proceedings of the South African Statistical Association Conference, volume 2015. South African Statistical Association (SASA), pp. 41–48.

DU PLESSIS, J. AND SCHLOMS, B. (2017). An investigation into the evidence of seasonal rainfall pattern shifts in the Western Cape, South Africa. *Journal of the South African Institution of Civil Engineering*, **59** (4), 47–55.

DYSON, L. L. (2009). Heavy daily-rainfall characteristics over the Gauteng province. *Water SA*, **35** (5).

EASTERLING, D. R., EVANS, J., GROISMAN, P. Y., KARL, T. R., KUNKEL, K. E., AND AMBENJE, P. (2000). Observed variability and trends in extreme climate events: A brief review. *Bulletin of the American Meteorological Society*, **81** (3), 417–426.

EASTOE, E. F. AND TAWN, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25–45.

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** (1), 23–36.

FEDOROVÁ, D. (2016). Selection of unit root test on the basis of length of the time series and value of AR (1) parameter. *Statistika*, **96** (3), 3.

FERREIRA, A. AND DE HAAN, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, **43** (1), 276–298.

FERRO, C. A. AND SEGERS, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65** (2), 545–556.

FLOODLIST (2018). Climate change impacts fragile river ecosystems. Last accessed: 03.03.2020.
URL: *https://www.smcyinternationalfamily.org/climate-change-impacts-fragile-river-ecosystems/*

GAO, M., MO, D., AND WU, X. (2016). Nonstationary modeling of extreme precipitation in China. *Atmospheric Research*, **182**, 1–9.

GUILBERT, J. (2016). *The Impacts Of Climate Change On Precipitation And Hydrology In The Northeastern United States*. Ph.D. thesis, University of Vermont.

GYAMFI, C., NDAMBUKI, J., AND SALIM, R. (2016). A historical analysis of rainfall trend in the Olifants basin in South Africa. *Earth Sciences Research Journal*, **5**, 129–142.

HANUM, H., WIGENA, A. H., DJURAIDAH, A., AND MANGKU, I. W. (2015). Modeling extreme rainfall with gamma-Pareto distribution. *Applied Mathematical Sciences*, **9** (121), 6029–6039.

HANUM, H., WIGENA, A. H., DJURAIDAH, A., AND MANGKU, I. W. (2017). The application of modeling gamma-Pareto distributed data using GLM gamma in estimation of monthly rainfall with TRMM data. *Sriwijaya Journal of Environment*, **2** (2), 40–45.

HOBIJN, B., FRANSES, P. H., AND OOMS, M. (2004). Generalizations of the kpss-test for stationarity. *Wiley Online Library*, **58** (4), 483–502.

HOLLOWAY, A. J., FORTUNE, G., CHASI, V., BECKMAN, T., PHAROAH, R., POOLMAN, E., PUNT, C., ZWEIG, P., ET AL. (2010). *RADAR Western Cape 2010: Risk and Development Annual Review*. Technical report, Disaster Mitigation for Sustainable Livelihoods Programme (DiMP).

HUNDECHA, Y., ST-HILAIRE, A., OUARDA, T., EL ADLOUNI, S., AND GACHON, P. (2008). A nonstationary extreme value analysis for the assessment of changes in extreme annual wind speed over the Gulf of St. Lawrence, Canada. *Journal of Applied Meteorology and Climatology*, **47** (11), 2745–2759.

HUSAK, G. J., MICHAELSEN, J., AND FUNK, C. (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, **27** (7), 935–944.

IYAMUREMYE, E., MUNG'ATU, J., MWITA, P., ET AL. (2019). Extreme value modelling of rainfall using poisson-generalized Pareto distribution: A case study Tanzania. *International Journal of Statistical Distribution and Applications*, **5** (3), 67–75.

JAKATA, O. AND CHIKOBVU, D. (2019). Modelling extreme risk of the South African financial index (j580) using the generalised Pareto distribution. *Journal of Economic and Financial Sciences*, **12** (1), 1–7.

KAJAMBEU, R. (2016). *Modelling flood heights of the Limpopo River at Beit-*

bridge Border Post using extreme value distributions, MSc dissertation, University of Venda.

KATZ, R. W. (2013). *Statistical methods for nonstationary extremes*. In: Extremes in a Changing Climate. Springer, pp. 15–37.

KRUGER, A. (2006). Observed trends in daily precipitation indices in South Africa: 1910–2004. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **26** (15), 2275–2285.

KRUGER, A. C. AND NXUMALO, M. (2017). Historical rainfall trends in South Africa: 1921–2015. *Water SA*, **43** (2), 285–297.

LIOLIOS, E. (2015). Google trends as a predictive tool for the sales of the apple.

MACKELLAR, N., NEW, M., AND JACK, C. (2014). Observed and modelled trends in rainfall and temperature for South Africa: 1960-2010. *South African Journal of Science*, **110** (7-8), 1–13.

MANDAL, S. AND CHOUDHURY, B. (2015). Estimation and prediction of maximum daily rainfall at Sagar Island using best fit probability models. *Theoretical and Applied Climatology*, **121** (1-2), 87–97.

MANHIQUE, A., REASON, C., SILINTO, B., ZUCULA, J., RAIVA, I., CONGOLO, F., AND MAVUME, A. (2015). Extreme rainfall and floods in Southern Africa in January 2013 and associated circulation patterns. *Natural Hazards*, **77** (2), 679–691.

MAPOSA, D. (2019). *Fitting a generalised extreme value distribution to four candidate annual maximum flood heights time series models in the lower Limpopo River basin of Mozambique. In: Recent Advances in Flood Risk Management*. IntechOpen, London.

MAPOSA, D., COCHRAN, J., LESAOANA, M., AND SIGAUKE, C. (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach. *Natural Hazards and Earth System Sciences Discussions*, **2** (8), 5401–5425.

MAPOSA, D., COCHRAN, J. J., AND LESAOANA, M. (2016). Modelling nonstationary annual maximum flood heights in the lower Limpopo River basin of Mozambique. *Jàmbá: Journal of Disaster Risk Studies*, **8** (1).

MASEREKA, E. M., OCHIENG, G. M., AND SNYMAN, J. (2018). Statistical analysis of annual maximum daily rainfall for Nelspruit and its environs. *Jàmbá: Journal of Disaster Risk Studies*, **10** (1), 1–10.

MAZVIMAVI, D. (2008). Investigating possible changes of extreme annual rainfall in Zimbabwe. *Hydrology & Earth System Sciences Discussions*, **5** (4).

MAZVIMAVI, D. (2010). Investigating changes over time of annual rainfall in Zimbabwe. *Hydrology and Earth System Sciences*, **14** (12), 2671–2679.

MÉLICE, J.-L. AND REASON, C. J. (2007). Return period of extreme rainfall at George, South Africa. *South African Journal of Science*, **103** (11-12), 499–501.

MOSASE, E. AND AHIABLAME, L. (2018). Rainfall and temperature in the Limpopo River basin, Southern Africa: Means, variations, and trends from 1979 to 2013. *Water SA*, **10** (4), 364.

MUCHURU, S., LANDMAN, W. A., DEWITT, D., AND LÖTTER, D. (2014). Seasonal rainfall predictability over the Lake Kariba catchment area. *Water SA*, **40** (3), 461–470.

MZEZEWA, J., MISI, T., AND VAN RENSBURG, L. (2010). Characterisation of rainfall at a semi-arid ecotope in the Limpopo province (South Africa) and its implications for sustainable crop production. *Water SA*, **36** (1), 20–26.

NAMITHA, M. AND RAVIKUMAR, V. (2018). Analysis of extreme rainfall events and calculation of return levels using generalised extreme value distribution. *International Journal of Pure and Applied Bioscience*, **6** (6), 1309–1316.

NAMITHA, M. AND VINOTHKUMAR, V. (2019). Derivation of the intensity-duration-frequency curve for annual maxima rainfall using generalised extreme value distribution. *International Journal of Current Microbiology and Applied Sciences*, **8** (1), 2626–2632.

NASH, D. J., PRIBYL, K., KLEIN, J., NEUKOM, R., ENDFIELD, G. H., ADAMSON, G. C., AND KNIVETON, D. R. (2016). Seasonal rainfall variability in southeast Africa during the nineteenth century reconstructed from documentary sources. *Climatic change*, **134** (4), 605–619.

NEL, W. (2009). Rainfall trends in the KwaZulu-Natal Drakensberg region of South Africa during the twentieth century. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **29** (11), 1634–1641.

NGAILO, J., REUDER, J., RUTALEBWA, E., NYIMVUA, S., AND MESQUITA, D. (2016). Modelling of extreme maximum rainfall using extreme value theory for Tanzania. *International Journal of Scientific and Innovative Mathematical Research*, **4** (3), 34–45.

ODIYO, J. O., MAKUNGO, R., AND NKUNA, T. R. (2015). Long-term changes and variability in rainfall and streamflow in Luvuvhu River Catchment, South Africa. *South African Journal of Science*, **111** (7-8), 1–9.

ODUNIYI, O. S. (2013). *Climate change awareness: A case study of small scale maize farmers in Mpumalanga province, South Africa*. Ph.D. thesis, University of South Africa.

OLOFINTOYE, O., SULE, B., AND SALAMI, A. (2009). Best-fit probability distribution model for peak daily rainfall of selected cities in Nigeria. *New York Science Journal*, **2** (3), 1–12.

OSMAN, Y. Z., FEALY, R., AND SWEENEY, J. (2015). Modelling extreme temperatures in ireland under global warming using a hybrid peak-over-threshold and a generalised pareto distribution approach. *International Journal of Global Warming*, **7** (1), 21–47.

PAN, J.-X. AND FANG, K.-T. (2002). Maximum likelihood estimation. *In Growth curve models and statistical diagnostics*. Springer, pp. 77–158.

PANAGOULIA, D., ECONOMOU, P., AND CARONI, C. (2014). Stationary and nonstationary generalized extreme value modelling of extreme precipitation over a mountainous area under climate change. *Environmetrics*, **25**, 29–43.

PAPARODITIS, E. AND POLITIS, D. N. (2013). The asymptotic size and power of the augmented Dickey-Fuller test for a unit root.

PHILLIPS, P. C. AND PERRON, P. (1988). Testing for a unit root in time series regression. *Biometrika*, **75** (2), 335–346.

PICKANDS III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, **3** (1), 119–131.

PINDURA, T. H. (2016). *An assessment of water security and hydrology resources in the face of climate variability: The case study of Nkonkobe Local Municipality, Eastern Cape, South Africa*. Ph.D. thesis, University of Fort Hare.

POULOS, M. (2016). Determining the stationarity distance via a reversible stochastic process. *PloS one*, **11** (10).

PYLE, D. M. AND JACOBS, T. L. (2016). The Port Alfred floods of 17–23 October 2012: A case of disaster (mis) management? *Jàmbá: Journal of Disaster Risk Studies*, **8** (1).

Rapolaki, R. S., Blamey, R. C., Hermes, J. C., and Reason, C. J. (2019). A classification of synoptic weather patterns linked to extreme rainfall over the Limpopo River basin in Southern Africa. *Climate Dynamics*, 1–15.

Rapolaki, R. S. and Reason, C. J. (2018). Tropical storm Chedza and associated floods over south-eastern Africa. *Natural Hazards*, **93** (1), 189–217.

Roghani, R., Soltani, S., and Bashari, H. (2016). Influence of southern oscillation on autumn rainfall in Iran (1951–2011). *Theoretical and Applied Climatology*, **124** (1-2), 411–423.

Sauka, S. (2016). *Flood risk assessment of the Crocodile River, Mpumalanga*. Ph.D. thesis, University of the Witwatersrand.

SAWS) (2019). Seasonal climate watch. Last accessed: 03.03.2020.
URL: *http://www.weathersa.co.za/images/data/longrange/gfcsa/scw.pdf*

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, **10** (1), 33–60.

Sharma, M. A. and Singh, J. B. (2010). Use of probability distribution in rainfall analysis. *New York Science Journal*, **3** (9), 40–49.

Shi, Q., Li, B., and Alexiadis, S. (2012). Testing the real interest parity hypothesis in six developed countries. *International Research Journal of Finance and Economics*, **86**, 168–180.

Sigauke, C. and Bere, A. (2017). Modelling non-stationary time series using a peaks over threshold distribution with time varying covariates and threshold: An application to peak electricity demand. *Energy*, **119**, 152–166.

Singo, L., Kundu, P., Mathivha, F., and Odiyo, J. (2016). Evaluation of flood risks using flood frequency models: A case study of Luvuvhu River catchment in Limpopo province, South Africa. *WIT Transactions on The Built Environment*, **165**, 215–226.

SINGO, L., KUNDU, P., ODIYO, J., MATHIVHA, F., AND NKUNA, T. (2012). *Flood frequency analysis of annual maximum stream flows for Luvuvhu River Catchment, Limpopo Province, South Africa*. University of Venda.

SLABBERT AND SLATTER (2019). Eskom blames cyclone Idai for SA's power outages. Last accessed: 03.03.2020.
URL: *https://city-press.news24.com/News/eskom-blames-cyclone-idai-for-sas-power-outages-20190317*

SMITH, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 367–377.

SYAFRINA, A., NORZAIDA, A., AND AIN, J. J. (2019). Stationary and nonstationary generalized extreme value models for monthly maximum rainfall in Sabah. *In :Journal of Physics: Conference Series*, volume **1366**. IOP Publishing, p. 012106.

SYAFRINA, A., ZALINA, M., AND JUNENG, L. (2015). Historical trend of hourly extreme rainfall in Peninsular Malaysia. *Theoretical and Applied Climatology*, **120** (1-2), 259–285.

THIOMBIANO, A. N., EL ADLOUNI, S., ST-HILAIRE, A., OUARDA, T. B., AND EL-JABI, N. (2017). Nonstationary frequency analysis of extreme daily precipitation amounts in Southeastern Canada using a peaks-over-threshold approach. *Theoretical and Applied Climatology*, **129** (1-2), 413–426.

THOMAS, D. S., TWYMAN, C., OSBAHR, H., AND HEWITSON, B. (2011). Adaptation to climate change and variability: Farmer responses to intra-seasonal precipitation trends in South Africa. *In: African Climate and Climate Change*. Springer, pp. 155–178.

THOMAS, M., LEMAITRE, M., WILSON, M. L., VIBOUD, C., YORDANOV, Y., WACKERNAGEL, H., AND CARRAT, F. (2016). Applications of extreme value theory in public health. *PloS One*, **11** (7).

UNOCHA (2019). Mozambique and Zimbabwe – Tropical Cyclone Idai causes death and destruction. Last accessed: 03.03.2020.
URL: *http://www.icfm.world/News/News/174/mbique-and-Zimbabwe-—Tropical-Cyclone-Idai-Causes-Death-and-Destruction*

WI, S., VALDÉS, J. B., STEINSCHNEIDER, S., AND KIM, T.-W. (2016). Nonstationary frequency analysis of extreme precipitation in South Korea using peaks-over-threshold and annual maxima. *Stochastic Environmental Research and Risk Assessment*, **30** (2), 583–606.

WUERTZ, D. AND KATZGRABER, H. G. (2005). Precise finite-sample quantiles of the jarque-bera adjusted lagrange multiplier test. *arXiv preprint math/0509423*.

ZENGENI, R., KAKEMBO, V., AND NKONGOLO, N. (2016). Historical rainfall variability in selected rainfall stations in Eastern Cape, South Africa. *South African Geographical Journal*, **98** (1), 118–137.

ZHAO, X., ZHANG, Z., CHENG, W., AND ZHANG, P. (2019). A new parameter estimator for the generalized Pareto distribution under the peaks-over-threshold framework. *Mathematics*, **7** (5), 406.

ZIN, W. Z. W., JEMAIN, A. A., AND IBRAHIM, K. (2009). The best fitting distribution of annual maximum rainfall in Peninsular Malaysia based on methods of L-moment and LQ-moment. *Theoretical and Applied Climatology*, **96** (3-4), 337–344.

Appendices

SOME SELECTED R CODES

R code for fitting non-stationary GEVD using ismev package

```
install.package(“ismev”)
library(ismev)
attach(EC.MAX.VALUESMax)
head(EC.MAX.VALUESMax)
```

Now to fit the GEV to allow for a linear trend in μ location only, we type:

```
tail(EC.MAX.VALUESMax)ti=matrix(ncol=1,nrow=118)
ti[,1]=seq(1,118,1)
ti=gev.fit(EC.MAX.VALUESMax,ydat = ti,mul = 1, sigl=NULL)
gev.diag(ti)
```

We can also create a quadratic linear trend in location only model, and linear in scale ti2=matrix(ncol=2,nrow=118)

```
ti2[,1]=seq(1,118,1)
ti2[,2]=(ti2[,1])**2
ti=gev.fit(EC.MAX.VALUESMax,ydat = ti2,mul=c(1,2),sigl = 1)
gev.diag(ti)
```

R code for fitting non-stationary GPD

Mean residual life plot and Threshold choice plots

mean residual life plot:

```
mrp=mrl.plot(data)
tcplot(data, u.range = c(40, 80), nt=10 )
```

decluster the sequence by using the automatic declustering method

```
ei <- extremalIndex(data, threshold = 55)
ei
dc <- declust(ei)
par(mfrow=c(1,1))
plot(dc,col=“blue”, xlab=“Time”, ylab=“Rainfall”)
dc
dc <- declust(data, threshold = 55)
freq=gpd.fit(data,threshold=55)
gpd.diag(freq)
```

Calculating t-ratios and p-values

```
tb1=abs((-0.1424708)/0.03459958); tb1
pt(tb1,118,lower.tail=FALSE)
ti=matrix(ncol=1,nrow=1416)
ti[,1]=seq(1,1416,1)
ti=gpd.fit(data,threshold=55,ydat=ti,sigl=1)
gpd.diag(ti)
```

We can also create a quadratic linear trend in location only model, and linear in scale

```
ti2=matrix(ncol=2,nrow=1416)
ti2[,1]=seq(1,1416,1)
ti2[,2]=(ti2[,1])**2
ti=gpd.fit(data,threshold=55,ydat = ti2,sigl = c(1,2))
gpd.diag(ti)
```

R code for fitting cubic regression smoothing splines using ismev

```
attach(data)
head(data)
tail(data)
win.graph()
fitting cubic regression smoothing splines
library(ismev)
fitting cubic regression smoothing splines
plot(data,xlab=“Observation number”, ylab=“Negated maximum rainfall (mm)”,
col=“blue”)
lines(smooth.spline(time(data),data, spar=0.1),col=“red”,lwd=3)
plot(data, type=“p”, ylab=“Negated maximum rainfall (mm)”, col=“blue”,xlab=
“Observation number”)
lines(smooth.spline(time(data), data, spar=0.59369),col=“red”, lwd=3)
smooth.spline(time(data), data) GCV
r2=residuals((smooth.spline(time(data), data, spar=0.59369)))
plot(r2,col=“blue”,ylab=“Residuals observations”, xlab=“Observation number”)
r2pos ≤ r2[r2 > 0]
plot(r2pos, ylab=“Residuals above time-varying threshold (positive residuals)”,
col=“blue”, xlab=“Observation number”)
tail(r2pos)
```

R code for fitting non-parametric extremal mixture model using evmix package

Nonparametric extreme value mixture models

Example fit kernel density

```
attach(data)
install.package(“evmix”)
library(evmix)
win.graph()
fit = fkdengpd(data, phiu = FALSE, std.err = FALSE)
hist(data,breaks=100, freq = FALSE, main=””,xlim = c(0,400))
dataa = seq(0,400, 1)
lines(dataa, dkdengpd(dataa, data, fit$lambda, fit$u, fit$sigmau, fit$xi, fit$phiu),
col=“blue”, lwd =2)
abline(v = fit$u, col=“blue”, lwd = 2)
legend(“topright”, “kdengpd”, col = “blue”,lty = 1, lwd = 2)
box()
fit
```