MODELLING MAXIMUM RAINFALL AND WIND SPEED USING THE GENERALISED PARETO DISTRIBUTION: A CASE STUDY OF MARA, LIMPOPO PROVINCE OF SOUTH AFRICA

by

MATOME LESLEY SEBOLA
(201926712)

A RESEARCH PROJECT SUBMITTED FOR THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE

HONOURS

in

STATISTICS

in the

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH
SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES
FACULTY OF SCIENCE AND AGRICULTURE

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: MR SIMON MONYAI

NOVEMBER 2023

Declaration

I, **Matome Lesley Sebola**, hereby certify that I am the author of this research project and that I have not previously submitted it to the University of Limpopo or another educational institution in order to be considered for a degree. I affirm that the content, ideas, and analysis presented in this research are my own original work, except where duly acknowledged and referenced.

Signature: Sebola Date: 09/11/2023

Abstract

Extreme weather events are climate weather occurrences that are rare. To estimate the frequency of extreme rainfall and wind speed, the generalised Pareto distribution (GPD) is applied to rainfall data and wind speed data at Mara station for 32 years (1990-2023). The maximum likelihood estimation was used to estimate the parameters of the GPD. The Mann-Kendall test revealed an independence of rainfall data and a dependence on wind speed data. The study used the nested models of the GPD to find a model that is suitable for fitting the data. It was found through the likelihood ratio test that the Exponential distribution is more suitable for fitting rainfall and the Pareto Type II was more appropriate for fitting wind speed at Mara station. The return levels corresponding to the return periods of 10, 20, 40, 50, and 100 years were estimated and indicated a slight increase in rainfall and wind speed. It was estimated that the maximum rainfall of 157.6821 mm will be exceeded on average every 100 years, whereas the maximum wind speed is estimated to exceed 10.036 m/s on average once every 100 years. The study assists in the field of climatology and hydrology by providing insight into the behaviour of extreme rainfall and wind speed at Mara village.

Keywords: generalised Pareto distribution, maximum rainfall, maximum wind speed, return levels.

Dedication

First and foremost, I thank God for making this research project possible. I offer my heartfelt dedication to my mother, Miss Mohlatlego Monica Sebola for her astounding support and my grandmother, Mrs Mokgadi Ennie Sebola for her guidance. Your love and encouragement have been my guiding light.

Acknowledgments

I thank Mr Simon Monyai for his patience and remarkable supervision throughout this work. I sincerely thank South African Weather Service (SAWS) for supplying the data for this research project. I greatly appreciate Mr Hulisani Lutombo, Mr Kgothatso Makubyane, Miss Mapula Lekganyane, and Miss Maggie Kalembo's support, guidance and motivation. Your support is much appreciated.

My profound gratitude goes to the National Research Foundation (NRF) for their generous financial support, which made this research project possible.

Contents

List of Figures

List of Tables

Abbreviations

A-D	Anderson-Darling
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
BIC	Bayesian Information Criteria
BM	Block Maxima
CI	Confidence Interval
EVT	Extreme Value Theory
GEVD	Generalised Extreme Value Distribution
GPD	Generalised Pareto Distribution
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
J-B	Jarque-Bera
K-S	Kolmogorov-Smirnov
LLRB	Lower Limpopo River Basin
MCMC	Markov Chain Monte Carlo
M-K	Mann-Kendall
POT	Peak-Over Threshold
PP	Phillips Perron
SAWS	South African Weather Service

Chapter 1

Introduction and background

1.1 Introduction

Extreme weather events are climate weather occurrences that are rare. Gálfi et al. (2017) attests to the importance of statistical models in estimating the frequency of disastrous weather occurrences and mitigating their effects. The extremes of most climate events are particularly notable instances because they may play a substantial role in understanding the harm caused by extreme events (Diriba and Debusho, 2020). In recent years, a small village called Mara in Limpopo Province of South Africa, has experienced extreme weather events, with both heavy rain and strong winds causing serious harm to infrastructure and property. The maximum rainfall and wind speed that may occur in the area needs to be accurately modelled to manage the risks related to these weather events (Martins et al., 2020).

Extreme Value Theory (EVT) is a branch of statistics used to analyse extreme events (Coles, 2001; Acero et al., 2014; Ferreira and De Haan, 2015; Bhagwandin, 2017; Maposa et al., 2017; Nemukula et al., 2018). The two most common techniques for practical extreme value analysis are the block maxima technique and the peak-over-threshold (POT) technique (Chuangchid et al., 2013; Ferreira and De Haan, 2015). The first method mainly focuses on the distribution of block maxima, where data is blocked in equal blocks and high observations within each block are fitted the Generalised Extreme Value Distribution (GEVD) (Nemukula et al., 2018; Diriba and Debusho, 2020). The second method, which is the peak-over-threshold (POT), recognises high values over a high threshold and fits the generalised Pareto distribution (GPD) to them (Pickands III, 1975; Gilli and Këllezi, 2006). This study presents the POT technique to model extreme rain and strong wind.

The POT approach determines the value beyond which the data can be classified as extreme (Coles, 2001; Castillo and Hadi, 1997). The threshold choice plots and the mean excess plot are two techniques that can be used to calculate this threshold (Keef et al., 2013; Nemukula et al., 2018; Maposa et al., 2021). Those chosen observations over a threshold have a probability distribution that is roughly a GPD (Pickands III, 1975; Scarrott and MacDonald, 2012). Once the threshold value is identified, the GPD can be fitted to the excesses above the threshold (McNeil and Saladin, 1997; Harris, 2005; Scarrott and MacDonald, 2012; Wu, 2014). This involves estimating the two parameters of the GPD: the scale parameter, which determines the scale of the distribution and the shape parameter, which determines the tail behaviour of the distribution, (de Zea Bermudez and Kotz, 2010; Wu, 2014). Modelling maximum rainfall and wind speed using the GPD can help better understand the nature of extreme rare weather events and generate more precise forecasts about the possibility of such events recurring in the future (Westra et al., 2014).

1.2 Problem statement

Extreme weather occurrences pose risks to the community and have a significant effect on people's quality of life (Jentsch and Beierkuhnlein, 2008; Hasan et al., 2012; Khavarian-Garmsir et al., 2019). A study by Dyson and Van Heerden (2001) determined the effects of the February 2000 rain and strong winds on the northern region of South Africa, leaving people homeless and some dead. Flooding as a result of heavy rains puts human lives in danger, damages homes and infrastructure, and destroys crops and livestock (Stephenson et al., 2008; Khan, 2011; IPCC, 2018; Martins et al., 2020). Strong winds may result in downed trees and power lines, transportation delays, car damage, and possibly injuries or death (Bourque et al., 2006; Stephenson et al., 2008).

At least seven people were killed by wildfire in Knysna, Western Cape of South Africa, in 2017, and hundreds of residences were burned (eNCA, 2019). Furthermore, the wind caused the fire to expand into a region that had not yet burnt, and it nearly destroyed the entire town. Around 70 people were killed by flooding and mudslides in KwaZulu-Natal, South Africa in 2019 because of heavy rainfall (TimesLive, 2019). Extreme events in South Africa demonstrate the vulnerability of society to extreme weather and climate change (Diriba and Debusho, 2020).

Stephenson et al. (2008) attests that a useful statistical technique that is frequently used to model extreme phenomena like floods, hurricanes, and windstorms is the GPD. Using the GPD, a statistical model is created to characterise and forecast extreme wind and rainfall events at Mara village. The objective is to investigate how extreme weather events behave and shed light on the underlying causes of these phenomena (Stephenson et al., 2008; Wu et al., 2017). The likelihood of future extreme wind and rainfall events is estimated using the model (Wu et al., 2017; Outten and Sobolowski, 2021). Planning for floods, landslides, and other disasters is a benefit of being able to predict extremes of wind and rainfall.

1.3 Motivation of the study

The choice of using the GPD in this study is motivated by its suitability for modeling extreme values. The GPD is commonly used in hydrology, meteorology, and climatology to analyse rare events that exceed a certain threshold. There is lack of literature applying GPD to extreme rainfall and wind speed at Mara station. Furthermore, Mara village residents rely on agriculture for their livelihood. By farming, they maintain food security (Khumbane, 2004). With the means to support their families, these people run small businesses that are solely dependent on farming. Crops and livestock are negatively impacted when weather events like heavy rain and strong winds occur (Vining, 1990; Roncoli et al., 2002). Their businesses are impacted, which promotes poverty in the community. Understanding and predicting extreme weather events are crucial for disaster preparedness, risk assessment, and climate change adaptation strategies. The study is motivated by the need to understand and minimise damages that can be caused by heavy rains and strong winds.

1.4 Purpose of the study

The purpose of this study is to analyse extreme rainfall and strong winds in Mara village, Limpopo province in South Africa from January 1990 to March 2023, and find a suitable model to estimate these extreme events.

1.4.1 Aim

The study aims to model maximum monthly rainfall and wind speed using the generalised Pareto distribution.

1.4.2 Objectives

The objectives of the study are to:

1. Use the peak over threshold method to fit the GPD to Mara data.
2. Model the tail behaviour of extreme wind speed and rainfall that exceeds a predetermined high threshold for Mara station in Limpopo province.
3. Determine the best model out of the nested models of the GPD for both rainfall and wind speed data.
4. Estimate return levels corresponding to return periods.

1.4.3 Research questions

This project will present the following questions:

1. Is the maximum rainfall and wind data stationary and independent?
2. Can an extreme value model, such as the generalised Pareto distribution, be used to model extreme maximum wind and rainfall?
3. Which nested model best fits the rainfall and wind datasets?
4. What effects does the anticipated return level have on rainfall and wind planning?

1.5 Scientific contribution

This study sheds light on how severe events behave. Decision makers in the field of climatology and hydrology will gain knowledge about high rain and strong winds and use the results to prepare for the worst. Engineers can also build safe structures that can resist the highest anticipated loads by using the study's findings. Furthermore, knowledge about rainfall and wind patterns can help farmers prepare for planting, irrigation, crop damage or loss.

1.6 Structure of the the study

The study is organised into five chapters. The introduction, problem statement, study motivation, aim and objectives, and scientific contribution are all presented in Chapter 1. A collection of relevant literature on wind and rainfall is included in Chapter 2. The study's methodology is provided in Chapter 3. The results of analysing the data are discussed in Chapter 4 while Chapter 5 contains the study's conclusion and recommendations.

Chapter 2

Literature review

2.1 Introduction

In this chapter, the history behind wind speed and rainfall patterns in some parts of the world, South Africa, and Limpopo Province, is reviewed. A review of what other researchers have done is discussed in this chapter. Many researchers applied EVT to predict extreme events like temperature, rainfall, air pollution and many others (Dosio, 2017; Abdullah et al., 2019). Many nations worldwide have been impacted by the change in climate (Masereka et al., 2018; Nemukula et al., 2018). This change has impacted how the weather behaves, including the wind, rain, and temperature. To stop the loss of lives and needless spending, more study on climate change needs to be done.

2.2 Rainfall in the world

For farmers as well as all other living beings in the world, the rainy season is thrilling. Farmers may rest easy knowing that their crops will get enough water to produce high-quality products. But in some places or countries, there is anomalous rainfall, which causes floods instead of the usual rainfall (Goudenhoofdt et al., 2017). In February 2000, approximately 800 people and 20,000 cattle died in Mozambique due to floods, with people left homeless and their land and crops destroyed (Williams et al., 2008; Christie and Hanlon, 2001).

Oliver and Mung'atu (2018) conducted research utilising the GEVD to estimate extreme maximum rainfall in Kigali City. The extreme maximum rainfall data for Kigali city were analysed, and it was best fitted by the Gumbel distribution. In their analysis, the greatest return rainfall amount was quite high, which is extremely uncommon to occur. It was anticipated that each return period's data value would occasionally exceed the maximum value possible. For the same station's maximum monthly rainfall data, the comparable return period was approximately 11 years.

Hossain et al. (2021) applied GEVD to predict data on high rainfall and its return period in Tasmania, Australia. The sensitivity of several parameter estimation strategies, which are frequently applied in the GEVD application, was examined in their study. For four distinct timeframes, four alternative approaches were used to estimate the GEVD parameters. In the majority of the Tasmanian rainfall stations examined, the Fréchet distribution was appropriate. Martins et al. (2020) studied how maximum rainfall events in Uruguaina were analysed using the GPD. To give the highest possible rainfall levels, the GPD was effectively fitted in all months. There was no clear upward trend or temporal correlation in the maximum monthly rainfall. For the return periods of 2, 5, 10, 30, 50, and 100 years, rainfall estimates from January to December were computed. With rainfall exceeding 170 mm every 100 years, April was the month with the highest estimate. When it rained roughly 90 mm every 100 years in July, the return level was at its lowest.

A time-heterogeneous generalised Pareto distribution was fitted by (Maposa et al., 2017) to the flood heights in Mozambique's lower Limpopo River basin (LLRB). The data from the basin were used in their study for the first time to apply models of extreme value theory related to climate change. For parameter estimation of the non-stationary GPD, the maximum likelihood approach was applied. The non-stationary GPD model having a linear trend in the scale parameter was used to generate the study's findings, which showed a very significant influence of climate change on the basin. At all three sites, the time-heterogeneous GPD methods outperformed the time-homogeneous GPD methods, indicating that a non-stationary GPD models are valuable and offer a better fit than the time-homogeneous GPD models.

Nkrumah et al. (2017) conducted a study in Ghana on the analysis of extreme rainfall and temperature values. Analysis was done on daily temperature and rainfall data from January 1960 to December 2012. The southern (Greater Accra), central (Ashanti Region), and northern (Northern Region) regions of the country were the focus of their study. As opposed to Ashanti and the Greater Accra Area, which both reported temperatures of 38.9°C, the Northern Region had a higher temperature of 42.8°C. The Weibull and Fréchet families of distributions were found to be suitable for modelling the extreme occurrences of temperature and rainfall in Ghana. Every five years, Accra, Ashanti, and the Northern Areas are expected to experience high temperatures of 34.7°C, 34.66°C, and 39.6°C respectively. Accra's probabilities of reaching temperatures of 39°C and 22°C were calculated to be 0.001 and 0.014, respectively.

Boudrissa et al. (2017) modelled the annual maximum daily precipitation in northern Algeria using the generalized extreme value DEVD to understand rainfall behavior and establish forecasting models for climate risk prevention. The research found that the Gumbel distribution is suitable for stations like Algiers and Miliana, while the Fréchet distribution fits better for the Oran station. Parameters were estimated using the maximum likelihood method, and return levels were calculated for various return periods. The findings suggested that for Algiers and Miliana, around 100 years must elapse to observe rainfall levels of 181.9 mm and 173 mm per day respectively, while for Oran, it's about 109.54 mm.

2.3 Rainfall in South Africa

South Africa experiences annual rainfall on average of about 450mm, which is significantly less than the global average of 860mm, and beside for south coast and southwestern, where rain falls year-round during the winter season, most of the rainfall occurs during the summer seasons (Alabi et al., 2019; Muller, 2022). In South Africa over the past ten years, the prevalence of extreme disasters such as catastrophic floods have damaged property and infrastructure (Diriba and Debusho, 2020). Nemukula et al. (2018) examined the application of r-largest order statistics to the modelling of South Africa's daily average temperature. Their research's findings supported the negative Weibull distribution as a best fit for the data and indicated that the average daily temperature will rise in the years to come.

Masingi and Maposa (2021) investigated long-term monthly rainfall variability in selected provinces of South Africa, aiming to understand trends, stationarity, and extreme value distributions in the context of climate change. The analysis revealed non-normality and non-stationarity in the rainfall distributions across the provinces. The paper employed the non-stationary generalized extreme value distribution (GEVD) to model monthly rainfall extremes, finding that the stationary GEVD is suitable for Eastern Cape, Gauteng, and KwaZulu-Natal, while non-stationary models with trends in location and scale parameters fit Limpopo and Mpumalanga data. Additionally, the distribution classes differed across provinces, with negative shape parameters indicating Weibull distribution for Eastern Cape and Mpumalanga and positive shape parameters suggesting Fréchet distribution for Gauteng, KwaZulu-Natal, and Limpopo. Model diagnostics confirmed these findings, providing insights for decision-makers to plan agriculture, infrastructure, and water resource management. The study established a baseline for future research on monthly rainfall variability in South African provinces, suggesting avenues for spatial extremes and Bayesian modeling approaches.

Mashishi (2020) modelled average monthly rainfall in South Africa using extreme value theory (EVT) to mitigate risks associated with heavy rainfall and floods. Employing both block maxima and peaks-over threshold (POT) approaches, the study estimated parameters using maximum likelihood estimators. It found that distributions in the Weibull domain, Gumbel domain, and generalized Pareto distributions are suitable for modelling average monthly rainfall in South Africa. While the POT approach suggested potential increases in rainfall in certain regions, the block maxima approach yielded similar results. Average monthly rainfall is predicted to be exceeded by 120.02 mm once every 20 years, according to the return level predictions of the Weibull distribution. The Weibull distribution-based 100-year return threshold of 161.13 mm was less than the February 2000 rainfall that destroyed numerous homes in Limpoo province.

Masereka et al. (2018) employed a study on modelling the annual maximum rainfall in Nelspruit, South Africa, utilising both theoretical and empirical continuous probability distribution functions. According to their research, the area must wait roughly 10 years before experiencing another period of heavy rainfall. De Waal et al. (2017) employed 76 rainfall stations to model the 1-day rainfall patterns of the Western Cape Province using the GPD. Their findings indicated that for 48 sites, the 50-year return period of 1-day rainfall patterns would rise, while the opposite was true for the remaining stations. The results from the Western Cape are a cause of worry for people since the province receives higher rainfall and could lead to extreme weather events like floods and thunderstorms.

Diriba and Debusho (2020) conducted extreme value analysis using the GEVD for modelling rainfall data at the East London station. The distributions' parameters were estimated using both Bayesian approaches and maximum likelihood. To account for the temporal non-stationary trend in the annual maxima in the case of maximum likelihood, the GEVD was modified. The maximum likelihood estimates of the 10-, 100-, and 1000-year return levels were then calculated using the maximum likelihood estimates of the model parameters. The Markov Chain Monte Carlo (MCMC) method was also used to apply the Bayesian approach to the GEVD. When more data on the extremes is available, the block-maxima method for extreme value analysis frequently wastes data, which causes significant uncertainty in return-level predictions. It was advised that daily rainfall data might be preferable to annual maxima since they result in less data loss.

2.4 Rainfall in Limpopo province

Extreme weather events such as heavy rain, strong winds, hot temperatures and drought are common in Limpopo. According to Molautsi (2021), approximately 7,000 people were affected by the massive flooding and landslides that occurred in March 2014 in the northeastern region of South Africa, with 3,525 being forced to leave their homes. 32 deaths from flash flooding were reported by the South African government, and many are still missing (DREFO, 2014). Mpumalanga, Gauteng, North West, and Limpopo were the most hit provinces, with Limpopo being the worst affected (Molautsi, 2021).

Sikhwari et al. (2022) employed the r-largest order statistics modelling method to simulate extreme rainfall in the Thabazimbi region of the Limpopo province of South Africa. The GPD was used for a comparative study. The findings of their analysis revealed that the data adhered to assumptions and followed a GEVD. Strong proof that GEVD was a good model for the block maxima data was supplied by diagnostic plots for the chosen station. The 50-year return level was determined to be 368 mm after the best data model was selected. This means that there is a 0.02 statistical probability that the 50-year return level will exceed 368 mm in the Thabazimbi region.

2.5 Wind speed in the world

Several studies have modelled wind for many reasons, one of which is generating electricity. Wind can be of good use to countries or regions that are in a state of emergency due to intense load-shedding. Wind power has been officially included in Canadian provincial energy strategies, according to the OPA (2008), which made this announcement in 2008. The OPA has put a lot of focus on obtaining what they refer to as ”renewable and cleaner sources of electricity," like wind (OPA, 2008). Despite the fact that wind energy has been used as a source of electricity for many years now, strong winds can be very damaging.

Mehmet and Özcan (2021) determined the maximum wind speed in advance and checked to see if it is guaranteed that the turbine is locked immediately before reaching the maximum wind speed as software. According to the results of the Artificial Neural Networks (ANN) algorithm's assessment of wind speed, there will be 14 days in a year when a wind farm will be exposed to damage from high-wind days. Their findings indicate that the area is a good location for the wind farm. Four meteorological observatories in Rwanda undertook a study to classify appropriate theoretical probability density distributions of wind speed (Safari, 2022). The root-mean-square error and a mean-biased error parameter were utilised in the study to assess how well the considered distributions performed. The distributions and residuals have been compared graphically to show that they support the techniques.

Baravalle and Köhler (2018) conducted research on the calibration of structural design standards using a probabilistic description of the wind climate. In their study, the Gumbel distribution is used to represent the 10-min mean wind speed, according to analysis of the wind data using various statistical techniques. By minimising asymptotic errors and being coherent with the underlying statistic, all methods showed that the upper tail of the Gumbel distribution was accurate. Ayuketang Arreyndip et al. (2016) used the GEVD for the first time to evaluate Debuncha, South-West Cameroon's, wind energy potential and analyse how energy fluctuates throughout the year. In their study, the data is divided into minimum, maximum, and mean monthly values, and the two parameters are fitted using the maximum likelihood approach. The Fréchet distribution, as opposed to the Weibull and Gumbel distributions, was the best fit for the minimum, mean, and maximum monthly data, according to the Kolmogorov-Smirnov (K-S) test. According to their findings, the wind power density during the wet season was greater than it was during the dry season.

Sarkar et al. (2019) performed the study to address the crucial role of wind velocity data modelling in estimating wind load, wind energy potential, and fatigue failure analysis in slender structures subjected to periodic vortex shedding. While Weibull models have been commonly used, their work investigated their validity for describing extreme wind velocities in four locations on the east coast of India. It found that while Weibull models accurately predict lower wind speeds, they became inappropriate for extreme velocities beyond a certain threshold. The study proposed techniques to determine this threshold and compared various extreme value distributions for modeling wind velocities beyond it, such as Gumbel, Fréchet, and reverse Weibull distributions. Additionally, methods for calculating risk coefficients from extreme hourly mean wind speed data were introduced and validated against Indian standards. The study concluded that Weibull distribution was suitable for modeling lower wind speeds but not extremes, highlighting the importance of accurately modeling extreme wind events for structural safety and wind energy estimation. The findings offered insights into wind climate modeling and provided a methodology applicable beyond the studied locations.

Sarkar et al. (2017) employed the study underscore the importance of fitting wind speed data with appropriate statistical models, like the Weibull distribution, for various purposes such as estimating the expected number of hours per year in critical wind speed ranges for structural analysis and assessing wind energy potential. However, the wind speed data provided by the Indian Meteorological Department are subject to distortion due to conversion from knots to integer km/h, leading to biased Weibull parameters and inaccurate estimations. By determining an appropriate class width of 4 km/h, the study mitigated this bias and found the Weibull model suitable for describing wind speed distributions in most locations in India, with the exception of a few where alternative models like the Gamma distribution may be considered. Additionally, it highlighted the importance of considering uncertainty in wind climate for structural analysis, although this does not significantly affect wind energy potential estimations based on Weibull parameters. Overall, the findings emphasised the need for careful consideration of data integrity and model selection in wind speed data analysis for various applications.

2.6 Wind speed in South Africa

Although there are fewer studies for modelling strong wind and its damages in South Africa, wind remains a destructive factor, mostly in the southern part of the country. It is essential to anticipate the destructive wind and rainfall that accompany tropical cyclones in order to reduce damage (Wang et al., 2020). Diriba and Debusho (2020) carried out a study on the analysis of extreme values using the GEVD to model the data on yearly maximum wind speed at the Cape Town weather station. Frequentist and Bayesian approaches have been used to analyse the extremes for the yearly maximum wind speed recorded by automated meteorological stations in Cape Town, Western Cape, South Africa. The Bayesian approach used the Markov Chain Monte Carlo methodology using the Metropolis-Hastings algorithm, whereas the frequentist approach used maximum likelihood to estimate the GEVD parameters. According to their findings, the GEVD model with a trend in the location parameter seemed to be a more accurate model for the data on yearly maximum wind speed. The outcomes demonstrated that there would be significant amounts of severe wind speed in the future.

Diriba and Debusho (2020) simulated dependence effects on the distributions of high values using frequentist and Bayesian techniques for data on extreme wind speed from the Port Elizabeth meteorological station in South Africa. While the Bayesian approach employed the Markov Chain Monte Carlo (MCMC) technique with the Metropolis-Hastings, the frequentist approach used maximum likelihood to estimate the parameters of EVDs. In their analysis, the standard errors for the GPD parameter estimates were lower when using the threshold method to model rather than the block maxima technique. The GPD has been consistently identified as the preferred distribution to fit extreme wind speed data at Port Elizabeth Station using frequentist and Bayesian approaches.

2.7 Prevention of damages due to high rain and strong winds

The occurrence of extreme weather events cannot possibly be prevented since they happen naturally, however, their occurrence can be prevented from being disasters (Neumayer et al., 2014). A book by Zommers and Singh (2014) produced evidence that to stop harm from extreme weather occurrences, systems that are effective in providing early warnings are essential. To assist citizens in preparing for extreme weather events, the United States National Weather Service (NWS) publishes extreme weather warnings and advice (NWS, 2021). According to Wang et al. (2020), reducing damage and financial losses is the aim of tropical cyclone warning. Additionally, their research offered proof that the decrease in losses in recent years is likely due to advancements in tropical storm warning systems as well as vessel protection.

Effective land management measures can reduce the damage caused by landslides and floods (Kjekstad and Highland, 2009). Enhancing soil retention and lowering the danger of soil erosion during heavy rainfall events can be accomplished by putting into practice afforestation and reforestation measures and encouraging appropriate agricultural activities (Sultana and Tan, 2021). Drought-resistant crop varieties, precise irrigation techniques, and sustainable water management are examples of climate-adaptive agricultural practices that can reduce exposure to risks related to the climate (IPCC, 2018).

Chapter 3

Methodology

3.1 Introduction

This chapter describes methodological approaches which are used in modelling extreme values. The techniques which are used to test for stationarity are described in this chapter. Furthermore, parameter estimation using maximum likelihood estimation and techniques of model checking are discussed.

3.2 Area of study and source of data

The area of study is a village called Mara in the Limpopo province. This study uses secondary data obtained from the South African Weather Service (SAWS) which ranges from January 1990 to March 2023. The study analyses two separate datasets, one for rainfall and the other for wind. The data is on maximum monthly rainfall and wind speed. The data for both rainfall and wind speed were recorded at 08:00 AM daily and a maximum value within each month was chosen for analysis. The data is analysed using R software version 4.3.0.

3.3 Jarque-Bera (J-B) normality test

When modelling extremes, the normal distribution is not a prerequisite. In EVT data is not required to be normally distributed since the modelling of extremes is in the tail of the distribution. The study checks to see if the data is normally distributed. The Jarque-Bera test is a statistical test used to determine if sample data follows a normal distribution based on the skewness and kurtosis of the data.(Jarque and Bera, 1980, 1987). The nonnegative test statistics for J-B is given by:

$$J - B = \frac{r}{6}\left[N^2 + \frac{(P-3)^2}{4}\right], \tag{3.1}$$

where

r is the observations number,
N is the skewness of sample ,
P is the kurtosis of the sample.

A measurement that reflects the asymmetry of the probability distribution is skewness(Cisar and Cisar, 2010). If skewness is zero, the distribution is perfectly symmetrical. If skewness is positive, the distribution is said to be positively skewed, indicating a longer or fatter tail on the right side of the distribution. If skewness is negative, the distribution is negatively skewed, indicating a longer or fatter tail on the left side of the distribution. Skewness is computed as follows:

$$N = \frac{\frac{1}{r}\Bigg(\sum_{i=1}^{r}(x_i - \bar{x})^3 \Bigg)}{\Bigg[\frac{1}{r}\Bigg(\sum_{i=1}^{r}(x_i - \bar{x})^2 \Bigg)\Bigg]^{\frac{3}{2}}}, \tag{3.2}$$

Kurtosis is a statistic for determining how heavily-tailed the distribution is (Westfall, 2014). Mesokurtic distributions have an excess kurtosis of zero and kurtosis of about three (Wright and Herrington, 2011). Leptokurtic distributions have excess kurtosis of more than zero and kurtosis greater than three (Westfall, 2014). The data are heavy-tailed, which means there are too many outliers in the data, and the distribution has wider tails. Excess kurtosis of less than zero characterises a platykurtic distribution, which has a kurtosis of less than three. Excess kurtosis of less than zero characterises a platykurtic distribution, which has a kurtosis of less than three (Wright and Herrington, 2011). Kurtosis is given by the following formula:

$$P = \frac{\frac{1}{r}\Bigg(\sum_{i=1}^{r}(x_i - \bar{x})^4 \Bigg)}{\Bigg[\frac{1}{r}\Bigg(\sum_{i=1}^{r}(x_i - \bar{x})^2 \Bigg)\Bigg]^{2}}, \tag{3.3}$$

The Jarque-Bera test is performed with the following hypotheses:

$H_0 : Maximum\ monthly\ data\ at\ Mara\ is\ normally\ distributed$

$H_1 : Maximum\ monthly\ data\ at\ Mara\ is\ not\ normally\ distributed$

A high value of the test statistic J-B leads to rejecting the null hypothesis above and deciding that maximum monthly data at Mara is not normally distributed (i.e. $J - B > \chi^2_\alpha$).Furthermore, the null hypothesis is rejected when p-value is less than the level of significance.

3.4 Test for stationarity

There may be trends or seasonal patterns in the data because it is gathered over time. The statistical characteristics of the data, such as mean, variance and covariance, will change over time if the data are not stationary, which makes it challenging to construct a model that can accurately represent the variability of the data. The study utilises the Augmented Dickey-Fuller (ADF) test, Kwiatkowski-Phillips-Schmidt-Shit (KPSS) test and Phillips Perron test to check if the maximum monthly data for Mara village is stationary or not.

3.4.1 Augmented Dickey-Fuller test

The testing procedure for the ADF test is applied to the following model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t \tag{3.4}$$

where

β is the coefficient on a trend,
α is the constant,
p is the lag order of the autoregressive process,
Δ is the first difference operator,
ϵ_t is a pure white noise error term,

The test statistic is given by:

$$\tau = \frac{\gamma}{\sqrt{var(\gamma)}} \tag{3.5}$$

The following assumptions are made when conducting the ADF test:
$H_0 : \gamma = 0 \, (Maximum\, monthly\, data\, at\, Mara\, is\, non-stationary)$
$H_1 : \gamma \neq 0 \, (Maximum\, monthly\, data\, at\, Mara\, is\, stationary)$

The null hypothesis is rejected if the p-value is less than the significance level and concludes that the series is stationary. The tabulated critical value can be compared with the test statistic. When the test statistic is more negative, the stronger evidence of rejecting the null hypothesis.

3.4.2 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

When determining whether a given series is stationary around a deterministic trend, the KPSS unit root test is used. The ADF test is identical to KPSS. The assumption of the null hypothesis that the series is stationary in the KPSS test sets it apart from the ADF test in a variety of important ways.

The following presumptions are made when conducting the KPSS test:
$H_0 : \gamma = 0 \, (Maximum \, monthly \, data \, at \, Mara \, is \, stationary)$
$H_1 : \gamma \neq 0 \, (Maximum \, monthly \, data \, at \, Mara \, is \, nonstationary)$

The test statistic is given by:

$$KPSS = \frac{1}{q^2} \sum_{i=1}^{q} \frac{\hat{p_i^2}}{p^2(l)} \tag{3.6}$$

where

q is the number of observations,
$\hat{p_i}$ is the sum of the residuals,
$p^2(l)$ is the consistent estimator of the long-run variance of the errors,

The null hypothesis is rejected if the p-value is less than the significance level and concludes that the series is non-stationary. The test statistic of KPSS should be greater than the critical value to reject the null hypothesis.

3.4.3 Phillips Perron (PP) Test

Phillips Perron test is a unit root test used in time series analysis. The test is similar to the ADF test, but it incorporates an automatic correction to the Dickey Filler procedure to allow for autocorrelated residuals. The test usually gives the same conclusions as the ADF test. The Phillips-Perron test involves fitting the regression:

$$y_i = \alpha + \rho y_{i-1} + \epsilon_i$$

where
α is the constant,
ϵ_t is a pure white noise error term,
ρ is the correlation coefficient.

The following assumptions are made when conducting the PP test:
$H_0 : \gamma = 0 \, (Maximum \, monthly \, data \, at \, Mara \, is \, non - stationary)$
$H_1 : \gamma \neq 0 \, (Maximum \, monthly \, data \, at \, Mara \, is \, stationary)$

The null hypothesis is rejected if the p-value is less than the significance level and concludes that the series is stationary.

3.5 Goodness of fit test

A model's goodness of fit refers to how well it fits a given collection of observations. The differences between actual values and the values predicted by the analysed model are often summarised by the goodness of fit measures. The study uses the Kolmogorov-Smirnov test and the Anderson-Darling test to test for goodness of fit on Mara village data.

3.5.1 Kolmogorov-Smirnov (K-S) test

The K-S test is a statistical method used to determine if there is a significant difference between an empirical distribution function and a hypothesised distribution function. The K-S test is useful when testing whether a dataset follows a specific theoretical distribution or when comparing two datasets to see if they come from the same population. The K-S test statistic is given by:

$$D_n = sup_x|F(x) - F_n(x)| \tag{3.7}$$

The following presumptions are made when conducting the K-S test:

$H_0: \ Maximum\ monthly\ data\ follows\ a\ Generalised\ Pareto\ distribution$

$H_1: \ Maximum\ monthly\ data\ does\ not\ follow\ a\ Generalised\ Pareto\ distribution$

If the calculated K-S test statistic is larger than a predefined critical value at 5% level of significance, then we reject the null hypothesis

3.5.2 Anderson-Darling (A-D) test

The Anderson-Darling (A-D) test is a statistical test used to determine whether a given sample of data comes from a specified distribution. The A-D test is used to fit an observed continuous PDF (sample) to an expected continuous PDF (parent). The test statistics A-D is defined as:

$$AD = -r - \frac{t}{r}\sum_{t=1}^{r}(2t-1)\Big[lnF(x_1) + ln\Big(1 - F(x_{r-t+1})\Big)\Big] \tag{3.8}$$

where

β is the coefficient on a trend.

The following presumptions are made when conducting the A-D test:

$H_0: Maximum\ monthly\ data\ follows\ the\ generalised\ Pareto\ distribution$

$H_1 : Maximum\ monthly\ data\ does\ not\ follow\ the\ generalised\ Pareto\ distribution$

The null hypothesis is rejected when the p-value is less than the specified level of significance. Furthermore, the null hypothesis can be rejected when the test statistic A-D is greater than the critical value.

3.6 Generalised Pareto Distribution

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed random variables having a marginal distribution function F. It is normal to view the X_i sequence by X events as extreme. It follows that conditional probability describes the stochastic behaviour of extreme events.

$$P\{X > u + y | X > u\} = \frac{1 - F(u+y)}{1 - F(u)}, y > 0 \tag{3.9}$$

Theorem 1: Let $X_1 X_2, \cdots$ be a sequence of independent random variables with common distribution function F, and let

$$M_n = max\{X_1, \cdots, X_n\}$$

denote an arbitrary term in the X_i, and suppose that F fulfils Theorem 1, so that for large n,

$$P\{M_n \leq z\} \approx G(z)$$

where

$$G(z) = exp\Big\{ - \Big[1 + \xi(\frac{z-\mu}{\sigma})\Big]^{-\frac{1}{\xi}} \Big\}, \tag{3.10}$$

for some μ, ξ and $\sigma > 0$

where

σ = is a scale parameter

ξ = is a shape parameter

μ = is a location parameter

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

$$H(y) \;=\; \begin{cases} 1-\left(1+\dfrac{\xi y}{\sigma}\right)^{-\dfrac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1-exp\left(-\dfrac{y}{\sigma}\right) & \text{if } \xi = 0, \end{cases} \tag{3.11}$$

is called the **generalised Pareto distribution**

defined on

$$\left\{y : y > 0 \; and \; \left(1+\frac{\xi y}{\sigma}\right) > 0\right\} \tag{3.12}$$

The GPD has three special cases of distribution:

1. For $\xi > 0$, we have the Ordinary Pareto distribution with $\gamma = \dfrac{1}{\xi}$ (tail index).
2. For $\xi = 0$, we have the Exponential distribution.
3. For $\xi < 0$, we have Pareto Type II distribution, it is a bounded interval given by $\left[0\;;-\dfrac{\beta}{\xi}\right]$

Justification for the Generalised Pareto Model

This section present a proof for Theorem 1. Let a random variable X have distribution function F. Theorem 1 states that

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

for some parameters $\xi, \sigma > 0$ and ξ. Thus, by taking logarithmic transformation,

$$nlogF(z) \approx -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi.} \tag{3.13}$$

But for large values of z, a Taylor expansion implies that

$$logF(z) \approx -\{1-F(z)\}. \tag{3.14}$$

Substitution in (3.13), it follows that, for large value of u,

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

Similarly, for $y > 0,$

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

Therefore,

$$P\{X > u+y|X > u\} \approx \frac{n^{-1}[1+\xi(u+y-\mu)/\sigma]^{-1/\xi}}{n^{-1}[1+\xi(u-\mu)/\sigma]^{-1/\xi}}$$

$$= \left[1+\frac{\xi(u+y-\mu)/\sigma}{1+\xi(u-\mu)/\sigma}\right]^{-1/\xi}$$

$$= \left[1+\frac{\xi y}{\sigma}\right]^{-1/\xi}$$

3.7 Model selection

Model selection is the process of choosing the best model from nested models of GPD based on performance criteria. In order to determine which nested model of GPD best fits the data, this study uses the likelihood ratio test, Akaike

Information Criterion (AIC), and Bayesian Information Criterion (BIC).

3.7.1 The likelihood ratio test

To select a family of the GPD that fits the data well, the deviance statistics can be applied to the data set (Nemukula et al., 2018). Generally, the maximum likelihood estimation of the nested model leads to a simple test procedure of one model against the other. M_0 is a special case of M_1. Given that models $M_0 \subset M_1$, then the deviance statistics is defined as:

$$D = 2\{l_1(M_1) - l_0(M_0)\} \tag{3.15}$$

where $l_1(M_1)$ and $l_0(M_0)$ are maximised log-likelihood under models M_1 and M_0, respectively. Large values of D indicate that model M_1 explains significantly more of the variation in the data than M_0, and small values of D imply that the increase in model size does not bring worthwhile improvements in the model's capacity to explain the data.

The test is conducted with the following assumptions:
$H_0 : \xi = 0 \, (Exponential\ distribution)$
$H_1 : \xi \neq 0 \, (Generalised\ Pareto\ Distribution)$

The model M_0 is rejected by a test at the α-level of significance if $D > C_\alpha$, where C_α is the $(1-\alpha)$ quantile of the χ^2_k distribution, and k is the difference in the number of parameters of M_1 and M_0 (Maposa et al., 2017).

3.7.2 Akaike Information Criterion (AIC)

The Akaike information criterion (AIC) is an estimator of prediction error and thereby, the relative quality of statistical models for a given data set. AIC calculates each model's quality in relation to all other models given a collection

of models for the data. AIC therefore offers a way to choose a model. A model is thought to be more accurate when its AIC is lower.
The following formula is used to calculate the AIC:

$$AIC = 2k - 2ln(\hat{L}) \tag{3.16}$$

where $\hat{L}$ is the maximised value of the likelihood function of the model. k is the number of free parameters in the model.

3.7.3 Bayesian Information Criteria (BIC)

A criteria for choosing a model from a limited number of options is the Bayesian information criterion (BIC); models with lower BIC are often accurate. Bayesian Information Criteria (BIC) is an evaluation of the purpose of the possibility, following the model's accuracy, under a particular Bayesian structure. Therefore, a model with a lower BIC is considered to be more likely to be the accurate model. The following formula is used to calculate the BIC:

$$BIC = kln(n) - 2ln(\hat{L}) \tag{3.17}$$

where $\hat{L}$ is the maximised value of the likelihood function of the model. k is the number of free parameters in the model. n is the number of observations.

3.8 Threshold selection

High observations must be identified, and that can be done by choosing a threshold. However, there is a problem with choosing a threshold, choosing a low threshold is likely to go against the model's asymptotic foundation and result in bias. A threshold that is too high will produce few excesses that can be used to estimate the model, resulting in a large variance. According to Coles (2001), there are two methods to be used to find a threshold: the mean residual plot and the threshold choice plot.

3.8.1 Mean residual life plot

The first method is based on the mean of the GPD. A plot that consists of the mean above several thresholds with the threshold itself is used. This plot is known as the mean residual life plot. If Y has a GPD with parameters σ and ξ, then:

$$E(Y) = \frac{\sigma}{1-\xi} \tag{3.18}$$

provided $\xi < 1$. When $\xi \geq 1$ the mean is infinite.
The mean residual life plot should be roughly linear in u above a threshold u_0, at which the GPD gives a reliable approximation to the excess distribution.

3.8.2 Threshold choice plot

The second approach relies on fitting the GPD at different thresholds and assessing the stability of the parameter estimates. A plot known as the threshold choice plot is used. If the GPD is a reasonable model for excesses of a threshold u_0, then excesses of a higher threshold u should also follow the GPD. The shape parameters of the two distributions are identical. However, denoting by σ_u the value of the GPD scale parameter for a threshold of $u > u_0$, it follows:

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0) \tag{3.19}$$

Threshold selection is made to identify extreme events, and events exceeding the threshold are utilised to fit the GPD model, followed by model verification and extrapolation.

3.9 Test for independence

Extreme value models assume an underlying process consisting of a sequence of independent random variables. Extreme events are close to independent at times that are far enough apart (Coles, 2001). The Mann-Kendall (M-K) test is applied to test for independence.

3.9.1 Mann-Kendall (M-K) test

A non-parametric test called the M-K test examines data gathered over time to look for trends that are regularly increasing or decreasing. The test compares the relative magnitude of sample data rather than the data values themselves. Although there must be no auto-correlation, the data do not need to be linear or normally distributed. Data gathered seasonally and data with variables cannot be subjected to the test. Statistic S can be obtained by:

$$S = \sum_{q=1}^{n-1} \sum_{p=q+1}^{n} sgn(x_p - x_q) \tag{3.20}$$

$$sgn(x_p - x_q) \quad = \quad \begin{cases} 1 & \text{if } x_p - x_q > 0, \\ 0 & \text{if } x_p - x_q = 0, \\ -1 & \text{if } x_p - x_q < 0. \end{cases} \tag{3.21}$$

where n is the sample length, x_q and x_p are from $q = 1, 2, \cdots, n-1$ and $p = q+1, \cdots, n$. In cases when n is more than 8, statistics S comes close to a normal distribution. The variance of S can be obtained as follows:

$$Var(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_t f_t(f_t - 1)(2f_t + 5)\right] \tag{3.22}$$

where t varies over the set of tied ranks and f_t is the number of times that the rank t appears. Then the standardised test statistic Z is denoted by:

$$Z = \tau \quad = \begin{cases} \dfrac{S-1}{\sqrt{var(S)}} & \text{if } S > 0, \\ 0 & \text{if } S = 0, \\ \dfrac{S+1}{\sqrt{var(S)}} & \text{if } S < 0. \end{cases} \tag{3.23}$$

Mann-Kendall trend test is conducted with the following assumptions:

$H_0 : \tau = 0\ (There\ is\ the\ independency\ of\ maximum\ monthly\ data\ at\ Mara)$

$H_1 : \tau \neq 0\ (There\ is\ no\ independency\ of\ maximum\ monthly\ data\ at\ Mara)$

The alternative hypothesis is that there is either a positive, non-null, or negative trend.

The standardised test statistic of greater than zero indicates an increasing trend, and the standardised test statistic of less than zero indicates a decreasing trend. The null hypothesis is rejected when the p-value is less than the level of significance.

3.10 Declustering

The models of extreme value theory assume a process consisting of a sequence of independent random variables (Coles, 2001). The POT method has the disadvantage of producing dependent data. Observations above the selected threshold tend to have temporal dependence. The threshold exceedances appear to occur in clusters or groups, implying that the occurrence of one extreme weather event is likely to be followed by another extreme weather event. Fitting the GPD to dependent data leads to estimators with small standard errors. Declustering is one of the methods used to deal with the challenges that arise with clustering. Declustering filters the dependent exceedances to obtain a set of threshold excesses that are approximately independent. Declustering works by the following procedure:

1. Clusters of exceedances should be defined using an empirical rule.
2. Determine the highest excess within each cluster.
3. Assume cluster maxima to be independent, with conditional excess distribution given by the generalised Pareto distribution.
4. Fit the generalised Pareto distribution to the cluster maxima.

Extremal index (θ) is a measurement that is used to assess how strongly exceedances are clustered. The value of theta ranges between 0 and 1. Exceedances are said to be independent when θ = 1. The value of theta which is greater than 0.5 (i.e., $\theta > 0.5$) provides evidence of weak dependence of exceedances whereas the value of theta is less than 0.5 (i.e., $\theta < 0.5$) provides evidence of strong dependence of exceedances. The value of theta is computed with the following formula:

$$\theta = (limiting\ mean\ cluster\ size)^{-1} \tag{3.24}$$

where limiting is in the sense of clusters of exceedances of increasingly high thresholds.

3.11 Parameter estimation

When a threshold is determined, the generalised Pareto distribution's parameters will be estimated using maximum likelihood. Suppose $y1, \cdots, y_q$ is the q excesses of a threshold u. When $\xi \neq 0$ the log-likelihood is derived from (3.11) as follows:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\dfrac{1}{\xi}}$$

$$h(y) = \frac{1}{\sigma}\left(1 + \frac{\xi y}{\sigma}\right)^{-\dfrac{1}{\xi^{-1}}}$$

$$L(\sigma, \xi) = \prod_{i=1}^{q} \frac{1}{\sigma}\left(1 + \frac{\xi y_i}{\sigma}\right)^{-\dfrac{1}{\xi^{-1}}}$$

$$L(\sigma, \xi) = \frac{1}{\sigma^q}\prod_{i=1}^{q}\left(1 + \frac{\xi y_i}{\sigma}\right)^{-\dfrac{1}{\xi^{-1}}}$$

$$logL(\sigma, \xi) = -qlog\sigma + \sum_{i=1}^{q} log\left(1 + \frac{\xi y_i}{\sigma}\right)^{-\dfrac{1}{\xi^{-1}}}$$

$$logL(\sigma, \xi) = -qlog\sigma - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{q} log\left(1 + \xi\frac{y_i}{\sigma}\right) \tag{3.25}$$

where $\left(1 + \dfrac{\xi y_i}{\sigma}\right) > 0$ for $i = 1, 2, \cdots, q$; otherwise, $l(\sigma, \xi) = -\infty$

When $\xi = 0$ the log-likelihood is obtained from (3.11) as follows

$$H(y) = 1 - exp\left[-\frac{y}{\sigma}\right]$$

$$h(y) = \frac{1}{\sigma}exp\left[-\frac{y}{\sigma}\right]$$

$$L(\sigma, \xi) = \prod_{i=1}^{q}\frac{1}{\sigma}exp\left[-\frac{y_i}{\sigma}\right]$$

$$L(\sigma, \xi) = \frac{1}{\sigma^q}exp\sum_{i=1}^{q}\frac{-y_i}{\sigma}$$

$$logL(\sigma,\xi) = -qlog\sigma - \frac{1}{\sigma}\sum_{i=1}^{q} y_i \tag{3.26}$$

Since numerical techniques are once again necessary, care must be taken to avoid numerical instabilities, as analytical maximisation of the log-likelihood is not achievable when $\xi \approx 0$ in (3.26) and to ensure that the algorithm does not fail due to evaluation outside of the allowable parameter space (Coles, 2001).

3.12 Return levels and return periods

The return period is the estimated average time between extreme events. Return level is a measure of how often an event is likely to occur.

Assume that a generalised Pareto distribution with parameters σ and ξ is an appropriate model for exceedances of a threshold u by a variable X. That is, for $x > u,\ P\{X > x|X > u\} = \left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}$
it follows that

$$P\{X > x\} = \zeta_u\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right] \tag{3.27}$$

where $\zeta_u = P\{X > u\}$. Consequently, the level x_m that is exceeded on average once every m observations is the solution of

$$\zeta_u\left[1+\xi\left(\frac{x_m-u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{m} \tag{3.28}$$

rearranging

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] \tag{3.29}$$

assuming m is large enough to ensure that $x_m > u$, provided that $\xi \neq 0$. If ξ=0, using a similar manner, (3.11) leads to

$$x_m = u + \sigma log(m\zeta_u) \tag{3.30}$$

provided m is large enough

x_m is the m-observation return level. Giving return levels on an annual basis is frequently more practical, making the N-year return level the level that is anticipated to be exceeded once every N years. If there are n_y observations per year, this corresponds to the m-observation return level, where $m = N \times n_y$. Hence, the N–year return level is defined by:

$$z_n \quad = \begin{cases} u + \dfrac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right] & \text{if} \quad \xi \neq 0, \\ u + \sigma log(Nn_y\zeta_u) & \text{if} \quad \xi = 0 \end{cases} \tag{3.31}$$

It is necessary to replace parameter values with their estimates to estimate return levels. For σ and ξ this corresponds to substitution by the corresponding maximum likelihood estimates, but an estimate of ζ_u, the probability of an individual observation exceeding the threshold u, is also needed. This has a natural estimator of

$$\hat{\zeta}_u = \frac{k}{n} \tag{3.32}$$

the sample proportion of points exceeding u. Since the number of exceedances of u follows the binomial $Bin(n, \zeta_u)$ distribution, $\hat{\zeta}_u$ is also the maximum likelihood estimate of ζ_u.

3.13 Model Checking

According to Coles (2001), when a threshold is selected for the fitted GPD, the quality of the fitted GPD model is assessed. To evaluate the effectiveness of a fitted GPD model, probability plots, quantile plots, return level plots, and density plots will be employed.

3.13.1 Probability plot

Considering a threshold u, threshold excesses $y_{(1)} \leq \cdots \leq y_{(q)}$ and an estimated model $\hat{H}$ the probability plot consists of the pairs:

$$\left\{ \left(\frac{i}{q+1}, \hat{H} y_i \right) \; i = 1, 2, \cdots, q \right\}$$

where

$$\hat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\sigma} \right), \;\; \hat{\xi} \neq 0$$

3.13.2 Quantile plot

The quantile plot consists of the following pairs:

$$\left\{ \left(\hat{H}^{-1} (\frac{i}{q+1}), y_i \right), \;\; i = 1, 2, \cdots, q \right\}$$

where

$$\hat{H}^{-1}(y) = u + \frac{\hat{\sigma}}{\xi} \left[y^{-\hat{\xi}} - 1 \right], \;\; \hat{\xi} \neq 0$$

The probability and quantile plots should contain roughly linear points if the GPD is appropriate for representing excesses of u (Coles, 2001).

3.13.3 Return level plot

The return level plot is a plot of the level that is expected to be exceeded by the process on average once in T-years against return period T. A return level plot consists of the locus of points $(m, \hat{x}_m)$ for large values of m, where $\hat{x}_m$ is the estimated m-observation return level:

$$\hat{x}_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[(m\hat{\zeta}_u)^{\hat{\xi}} - 1\right]; \quad \hat{\xi} \neq 0$$

3.13.4 Density plot

A histogram of the threshold exceedances can be used to compare the density function of the fitted GPD (Coles, 2001).

Chapter 4

Data Analysis

4.1 Introduction

The analysis of data and discussion of results are presented in this chapter. Results are computed using R software. The Analysis of two independent datasets for Mara station is presented in this chapter. Section 4.2 presents data analysis and discussions for maximum monthly rainfall data whereas section 4.3 presents data analysis for maximum monthly wind data. The study uses a 5% level of significance.

4.2 Data Analysis for rainfall

This section presents the analysis of maximum monthly rainfall and discussions based on the results from the analysis.

4.2.1 Descriptive Statistics

To determine the centralization, and distribution of maximum monthly rainfall data, descriptive statistics analysis was carried out. The mean is used to determine the data's central tendency. The shape of the distribution for the maximum monthly rainfall data is determined using skewness and kurtosis.

Table 4.1: Descriptive statistics of maximum monthly rainfall.

Min	Q1	Median	Q3	Max	Mean	Kurtosis	Skewness
0.1	3.5	13.8	30.1	124.0	20.686	3.065	1.667

Table 4.1 shows a descriptive summary of maximum monthly rainfall data. The minimum rainfall for Mara village was 0.1 mm recorded in August 1997 and July 1998, and the maximum rainfall was 124 mm recorded in February 1997. The mean rainfall of 20.686 mm which is greater than the median of 13.8 mm, provides evidence that the data is skewed to the right. The skewness value of 1.667, which is a positive value confirms that indeed the data is skewed to the right. Kurtosis of 3.065, which is close to 3, indicates that the distribution is mesokurtic and the tails of the distribution are closer to of normal distribution. It can further be concluded that maximum monthly rainfall data might be normally distributed.

4.2.2 Exploratory data analysis

Figure 4.1 illustrates the maximum monthly rainfall in Mara from 1990 to 2023. The maximum monthly rainfall observations are distributed evenly. There is no linear relationship of maximum monthly rainfall occurrence over the years.

Figure 4.2 illustrates the histogram of maximum monthly rainfall at Mara village. It can be observed that the majority of the data is to the right of the graph's peak, and because the right tail is longer than the left, the data is skewed to the right. The skewness of data to the right is also supported by a skew value of 1.667.

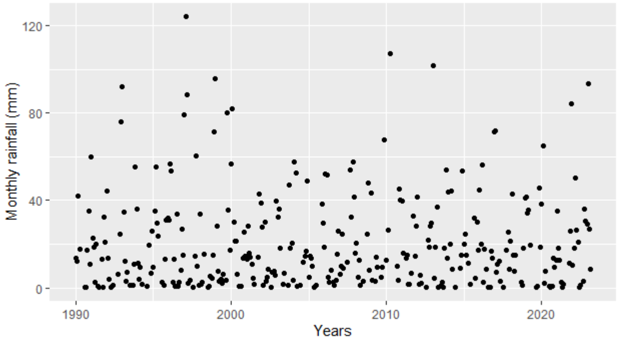

Figure 4.1: Scatter plot of maximum monthly rainfall at Mara village.

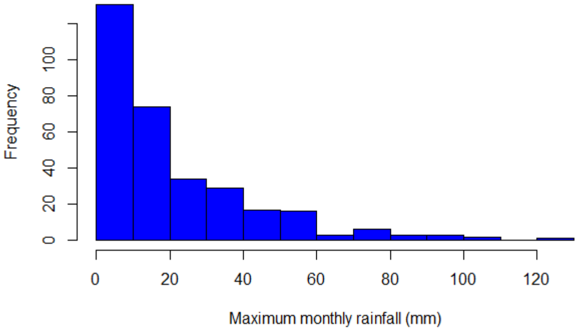

Figure 4.2: Histogram of maximum monthly rainfall at Mara village.

4.2.3 Jarque-Bera normality test

The results of the Jarque-Bera normality test are provided in Table 4.2.

The Jarque-Bera test is performed with the following hypotheses:

$H_0 : Maximum\, monthly\, data\, at\, Mara\, is\, normally\, distributed$

$H_1 : Maximum\, monthly\, data\, at\, Mara\, is\, not\, normally\, distributed$

The rejection of the null hypothesis stating that maximum monthly rainfall data is normally distributed is confirmed by results in Table 4.2. Results in

Table 4.2: Jarque-Bera normality test for maximum monthly rainfall.

Name of method	Chi-square statistic	p-value
J-B	277.1117	< 0.001

Table 4.2 provide a very small p-value of less than 0.01. Furthermore, the large chi-squares statistic value of 277.1 provides more evidence for rejecting the null hypothesis. Maximum monthly rainfall data at Mara village is therefore not normally distributed.

4.2.4 Test for stationarity

The data that has a constant mean, variance and covariance provide precise predictions and they are easier to make. A stationary data has a constant mean, variance, and covariance. This study conducted the Augmented Dickey-Fuller test, Kwiatkowski Phillips Schmidt test and Phillips Perron test to test for stationarity of maximum monthly rainfall at Mara village and the results are provided in Table 4.3.

The following assumptions are made when conducting the ADF and PP test:

$H_0 : \ Maximum\ monthly\ data\ at\ Mara\ is\ non-stationary$

$H_1 : \ Maximum\ monthly\ data\ at\ Mara\ is\ stationary$

The following assumptions are made when conducting the KPSS test:

$H_0 : \ Maximum\ monthly\ data\ at\ Mara\ is\ stationary$

$H_1 : \ Maximum\ monthly\ data\ at\ Mara\ is\ non-stationary$

Table 4.3: Stationarity test of maximum monthly rainfall data using ADF and KPSS.

Name of method	test statistic	p-value
ADF	-9.1059	0.01
KPSS	0.022139	0.1
PP	-190.17	0.01

The ADF test produced a p-value of 0.01 which is less than 0.05 level of significance, there is enough evidence to reject the null hypothesis which states that the series is non-stationary. This provides evidence that maximum monthly rainfall data at Mara is stationary. The KPSS test supported the conclusion of ADF by producing a p-value of 0.1 which is greater than the 0.05 level of significance, providing evidence to fail to reject the null hypothesis stating that the series is stationary. Furthermore, the PP test produced a p-value of 0.01 which is less than the 0.05 level of significance, providing evidence to reject the null hypothesis which states that the series is non-stationary. All the stationarity tests provided evidence that maximum monthly rainfall data is stationary.

4.2.5 Mann-Kendall independence test

The models of extreme value theory assume that a series of data is independent. Such an assumption is made in this study because extreme rainfall in a month does not imply that it will rain to the extreme in the next month. The Mann-Kendall test is used to test for independence of maximum monthly rainfall data and results are provided in Table 4.4.

Mann-Kendall trend test is conducted with the following assumptions:

$H_0 : There\, is\, the\, independency\, of\, maximum\, monthly\, data\, at\, Mara$

$H_1 : There\, is\, no\, independency\, of\, maximum\, monthly\, data\, at\, Mara$

Table 4.4: Mann-Kendall Independence test for maximum monthly rainfall.

Statistic S	Var(S)	Tau	p-value	Statistic Z
1448	3623292.667	0.0286	0.44715	0.76018

The M-K test in Table 4.4 provides a p-value of 0,44715 which is greater than the 0,05 level of significance, providing evidence to fail to reject the null hypothesis which states that maximum monthly rainfall data are independent. Furthermore, Kendall's tau value of 0.0286 is produced by the M-K test in Table 4.4 and concludes in favour of the null hypothesis. The value of the z statistic of 0.76018 provides evidence of an increasing monotonic trend. Therefore, there is enough evidence to conclude that maximum monthly rainfall observations are independent.

4.2.6 Goodness of fit test

The Anderson-Darling test and Kolmogorov-Smirnov test are employed to test if maximum monthly rainfall data for Mara village follows the generalised Pareto distribution.

The following presumptions are made when conducting the A-D and K-S test:

$H_0 : Maximum\ monthly\ data\ follows\ a\ Generalised\ Pareto\ distribution$

$H_1 : Maximum\ monthly\ data\ does\ not\ follow\ a\ Generalised\ Pareto\ distribution$

Table 4.5: Goodness of fit test for maximum monthly rainfall using A-D and K-S tests.

Name of method	test statistic	p-value
A-D	15.402	< 0.001
K-S	0.80685	< 0.001

Table 4.5 contains results for both A-D and K-S tests. Both tests have p-values less than 0.05 level of significance. They provide evidence to reject the null hypothesis and conclude that maximum monthly rainfall data does not follow the generalised Pareto distribution.

4.2.7 Threshold selection

It is important to select a threshold that does not lead to biasedness or high variance of the estimates. The mean residual life plot in Figure 4.3 appears to curve from threshold $u = 0$ to u≈ 25. From u≈ 25 to u≈ 50 the graph looks like there is stability. At $u = 50$ to $u = 60$, there is a linear rise of the graph followed by a gradual decay which suggests that we select the threshold to be 60. There are few observations when $u = 60$. It is a better choice to work with a threshold set at u≈ 50. Figure 4.4 shows how the pattern changes for very high thresholds that are seen in the mean residual life plot, but the changes are now found to be minor in comparison to sampling errors. It can be seen that a threshold of $u \approx 50$ appears reasonable.

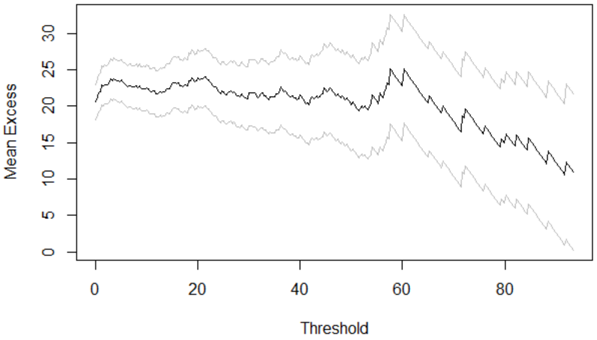

Figure 4.3: Mean residual life plot for maximum monthly rainfall data.

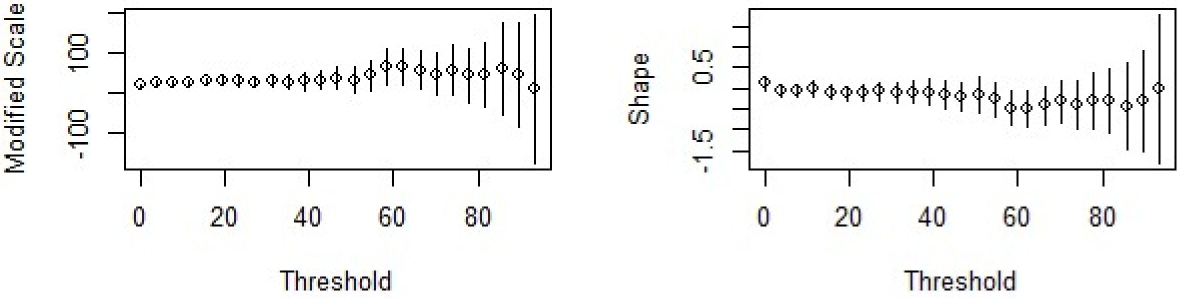

Figure 4.4: Threshold choice plots for maximum monthly rainfall

4.2.8 Parameter estimation and model fitting

The estimated parameters of the generalised Pareto distribution are presented in Table 4.6. The parameter estimation method utilised is maximum likelihood estimation. The results in Table 4.6 are the estimates with a 95% confidence interval. The GPD is fitted to all exceedances.

Table 4.6: Parameter estimates of GPD exceedances using MLE for maximum monthly rainfall data.

	Exceedances	
Parameter	Estimate (SE)	95% CI
ξ	-0.162 (0.2101)	(-0.574; 0.250)
σ	23.619 (6.373)	(11.128; 36.109)
$\hat{\theta}$	0.9785	

The run length of 1 with extremal index estimated as $\hat{\theta}$ =0.978 suggests the independence of maximum monthly rainfall observations, which support results from Mann Kendall test. An extremal index below 0.82 validates the dependency of the data (Diriba et al., 2015). The value of the shape parameter is less than zero ($\xi < 0$) which suggests that the fitted generalised Pareto distribution in Table 4.6 provides evidence that the maximum monthly rainfall events are in the Pareto Type II family. However, the 95% confidence interval for the shape parameter can be zero or positive. This provides more concerns about which distribution fits the maximum monthly rainfall data well. Likelihood ratio test using the deviance statistics is used to choose a distribution that fits the data well.

The profile likelihood is further used to provide more accurate confidence intervals for the shape parameter. The confidence interval of the shape parameter using the profile likelihood is (-0.3105; 0.4145) and it can be seen that zero is included in the interval. The inclusion of zero provides evidence that the shape parameter is in the Exponential family.

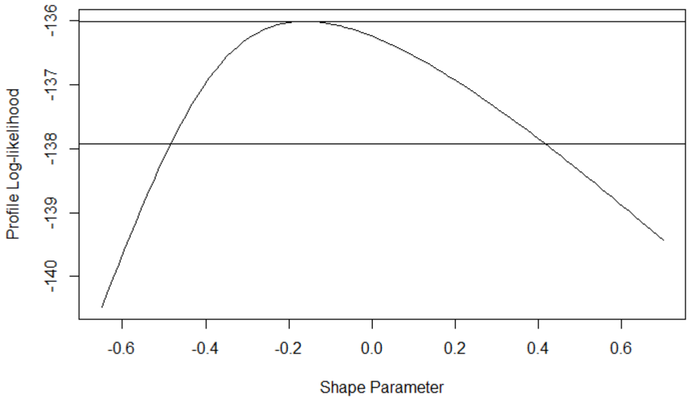

Figure 4.5: Profile likelihood of maximum monthly rainfall exceedances.

4.2.9 Model selection

Now that we have fitted the GPD, we need to find the nested model of GPD that is the best fit for the data. The study uses the likelihood ratio test, AIC and BIC to select the best nested model of the GPD for the data.

The likelihood ratio test

The likelihood ratio test is performed to select the best distribution that fits the data well.

The test is conducted with the following assumptions:
$H_0 : \xi = 0 \, (Exponential\ distribution)$
$H_1 : \xi \neq 0 \, (Generalised\ Pareto\ Distribution)$

Table 4.7: Likelihood ratio test for maximum monthly rainfall exceedances.

	Test statistic	Chi-square critical value	p-value
Exceedances	0.4648	3.8415	0.4954

The Likelihood ratio test in Table 4.7 produces a p-value of 0.495, which is above the 0.05 level of significance. This provides evidence to fail to reject the null hypothesis which states that all exceedances can be fitted well by the Exponential distribution. Another evidence in favour of failing to reject the null hypothesis is provided by the likelihood test statistic of 0.465 which is less than the critical value of 3.8415. We then conclude that all maximum monthly rainfall exceedances for Mara village can be best fitted by Exponential distribution.

Information Criteria

Akaike Information Criterion and Bayesian Information Criteria are also utilised to select the best nested model of the GPD that fits the data well. The results of AIC and BIC are produced in Table 4.8.

Table 4.8: AIC and BIC model selection for maximum monthly rainfall.

	AIC	BIC
GPD	276.0206	279.0733
Exponential	274.4854	276.0117

Exponential distribution has small AIC and BIC, this provides evidence that the exponential distribution is the best fit as compared to the GPD for maximum monthly rainfall data.

4.2.10 Model checking (Diagnostic plots)

The diagnostic plots for assessing the accuracy of the fitted GPD model are displayed in Figure 4.6. The probability plot, density plot, return level plot and density plot are used in this study to check the good fit of the generalised Pareto distribution. The probability plot illustrated in Figure 4.6 suggests a lack of fit at the centre of the generalised Pareto distribution. The data points do not fall close to the straight line at the centre which indicates that GPD is not a good fit. The quantile plot shows that the GPD does not fit the data well at the tail.

The data points at the tail do not follow the fitted distribution line. It indicates that the GPD is not a good fit for maximum monthly rainfall exceedances. The plot of return periods in years against the return level shows that the selected threshold is reasonable since the points are not outside the estimated confidence intervals represented by blue lines in Figure 4.6. The density plot shows that the modelled distribution does not align with the empirical density, which indicates that the GPD is not a good fit for maximum monthly rainfall data.

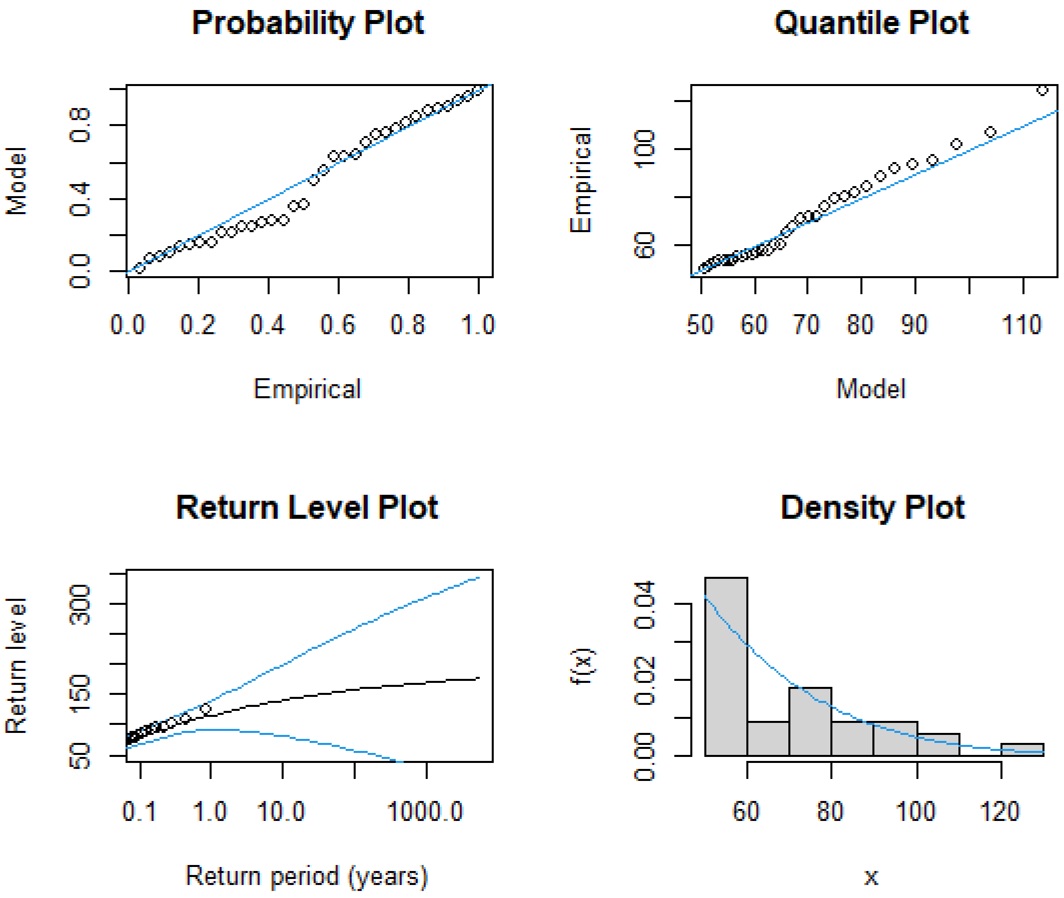

Figure 4.6: GPD diagnostic plots for maximum monthly rainfall exceedances.

4.2.11 Return level estimates

Table 4.9 presents the estimated return levels with a 95% confidence interval and standard errors corresponding to their return levels. The return level corresponds to the return periods of 10, 20, 40, 50, and 100 years. Results in Table 4.9 show that the return levels for exceedances increase slightly for higher return periods. The confidence intervals are increasingly wider as the return period is increasing. It is suggested from Table 4.9 that the maximum monthly rainfall will exceed 157.6821 mm on average once every 100 years.

Table 4.9: GPD return level estimates of maximum monthly rainfall exceedances.

Return Periods	Estimate	95% CI
10-years	140.3377	(78.5899; 202.1643)
20-years	146.2795	(72.1076; 220.4514)
40-years	151.5559	(64.6536;238.4582)
50-years	153.1325	(62.0934; 244.1716)
100-years	157.6821	(53.7582; 261.6061)

4.3 Data analysis for wind speed

This section presents the analysis of maximum monthly wind speed data and a discussion of its results. Statistical tests are performed to reveal hidden patterns in wind data.

4.3.1 Descriptive statistics

The summary statistic for maximum monthly wind speed data is provided in Table 4.10. Mara village maximum monthly wind speed data ranges from 1.5 m/s to 10 m/s. The median wind speed is 5 m/s. The mean wind speed of 4.974 m/s, being less than the median of 5 m/s, shows that the data is skewed to the left. Kurtosis of -0.1434, indicates that the distribution is platykurtic. The tails of the distribution are thinner than of normal distribution and there is a lack of outliers in the data.

Table 4.10: Descriptive statistics for maximum monthly wind speed data.

Min	Q1	Median	Q3	Max	Mean	Kurtosis	Skewness
1.5	3.7	5	6	10	4.97396	0.4609	-0.1434

4.3.2 Exploratory data analysis

The scatter plot of maximum monthly wind speed data in Figure 4.7 shows that in the past three decades, the occurrence of strong wind has been decreasing.

The histogram of maximum monthly wind speed data illustrated by Figure 4.8 shows that the tail extends much further out to the right.

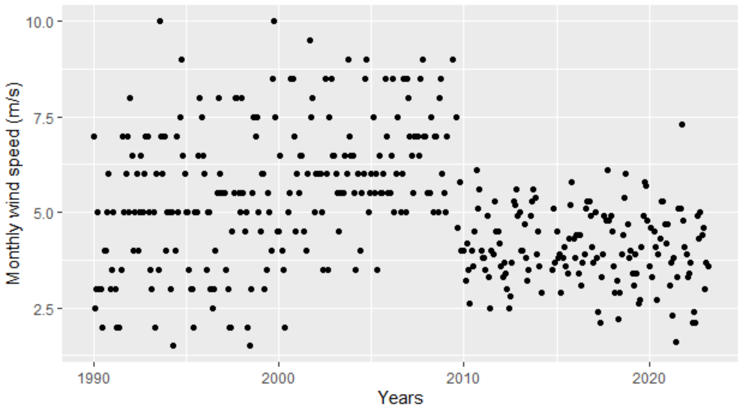

Figure 4.7: Scatter plot of maximum monthly wind speed data at Mara village.

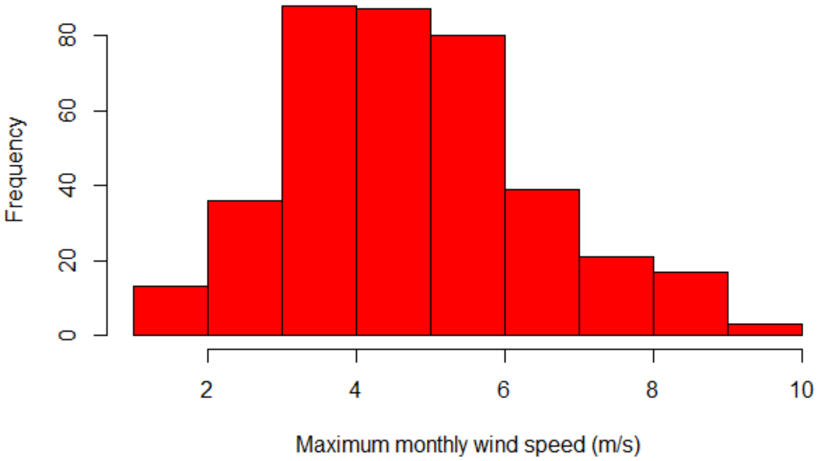

Figure 4.8: Histogram of maximum monthly wind speed at Mara village.

4.3.3 Jarque-Bera normality test

To model extreme value events like a strong wind, it is not a requirement that the data must not follow the normal distribution since the focus is on the tails of the data. The results of the Jarque-Bera normality test are provided in Table 4.11.

The Jarque-Bera test is performed with the following hypotheses:

$H_0 : Maximum\ monthly\ data\ at\ Mara\ is\ normally\ distributed$

$H_1 : Maximum\ monthly\ data\ at\ Mara\ is\ not\ normally\ distributed$

Table 4.11: Jarque-Bera normality test for maximum monthly wind speed data.

Normality test method	Chi-squared test statistic	p-value
J-B	13.9664	< 0.001

The results in Table 4.11 provide a very small p-value less than 0.05 level of significance, and a chi-squares statistic value of 13.9664. These values provide evidence of rejecting the null hypothesis stating that maximum monthly wind speed data at Mara is normally distributed. Maximum monthly wind speed data at Mara village is therefore not normally distributed.

4.3.4 Test for stationarity

To produce precise predictions, we need to know if the mean, variance, and covariance of maximum monthly wind speed data are not changing over time.

The following assumptions are made when conducting the ADF and PP test:

$H_0 :\ Maximum\ monthly\ data\ at\ Mara\ is\ non-stationary$

$H_1 :\ Maximum\ monthly\ data\ at\ Mara\ is\ stationary$

Table 4.12: Stationarity test for maximum monthly wind speed data using ADF and KPSS.

Stationarity test method	Chi-squared test statistic	p-value
ADF	-5.5326	0.01
PP	-179.17	0.01

The ADF test in Table 4.12 produced a p-value of 0.01 which is less than 0.05 level of significance, there is enough evidence to reject the null hypothesis which states that the series is not stationary. The PP test produced a p-value of 0.01 which is less than 0.05 level of significance, there is enough evidence to reject the null hypothesis which states that the series is not stationary. All stationarity tests provide evidence that maximum monthly wind speed data is stationary.

4.3.5 Mann-Kendall (M-K) independence test

Extreme weather events are rare events, thus it is unlikely for an extremely windy month to be followed by another extremely windy month. Observations are likely to appear clustered, which induces temporal dependency on data. The models of extreme value theory require a series of data to be independent. The Mann-Kendall test is applied to check the dependency of data and the results are illustrated in Table 4.13.

Mann-Kendall trend test is conducted with the following assumptions:

$H_0 : There\ is\ the\ independency\ of\ maximum\ monthly\ data\ at\ Mara$

$H_1 : There\ is\ no\ independency\ of\ maximum\ monthly\ data\ at\ Mara$

Table 4.13: Mann-Kendall independence test for maximum monthly wind speed.

Statistic S	Var(S)	Tau	p-value	Statistic Z
-12808	6299504.667	-0.1776	< 0.001	-5.1026

The p-value is less than 0,05 level of significance, providing evidence to reject the null hypothesis which states that maximum monthly wind speed data are independent. Furthermore, the negative value of the standardised test statistic z indicates that there is a decreasing monotonic trend in the data and maximum monthly wind speed data of Mara village is dependent.

4.3.6 Goodness of fit test

The goodness of fit test is performed to measure how well the generalised Pareto distribution fits the wind speed data of Mara village. A-D test and K-S test are employed to test if maximum monthly wind speed data for Mara village follows the generalised Pareto distribution and results are illustrated in Table 4.14.

The following presumptions are made when conducting the A-D and K-S test:

$H_0 : Maximum\, monthly\, data\, follows\, a\, Generalised\, Paretodistribution$

$H_1 : Maximum\, monthly\, data\, does\, not\, follow\, a\, Generalised\, Pareto\, distribution$

Table 4.14: Goodness of fit test using A-D and K-S for maximum monthly wind speed.

Name of method	test statistic	p-value
A-D	2.4254	< 0.001
K-S	0.8695	< 0.001

Both tests have p-values less than 0.05 level of significance. They provide evidence to reject the null hypothesis and conclude that maximum monthly wind speed data does not follow the generalised Pareto distribution.

4.3.7 Threshold selection

Selecting a threshold comes with two challenges: too high a threshold generates few outliers, leading to high variance and too low a threshold violates the asymptotic basis of the model, leading to bias. The mean residual life plot and threshold choice plot are illustrated in Figure 4.9 and Figure 4.10. It is not easier to interpret the mean residual life plot, according to Figure 4.10 it is reasonable to choose a threshold of 7.1 m/s.

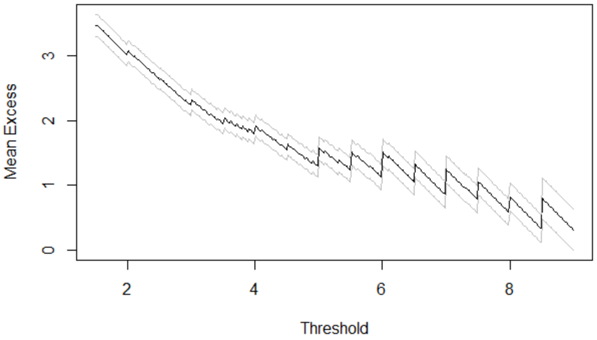

Figure 4.9: Extremal Index for declustering of maximum monthly wind speed data.

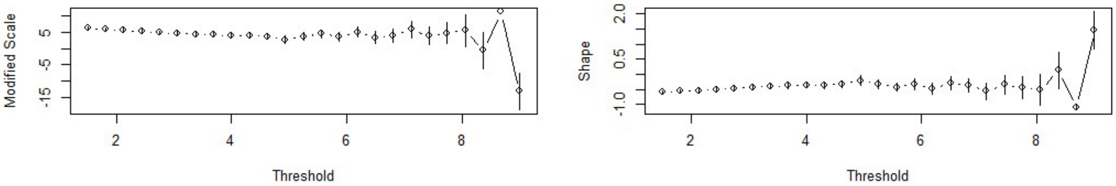

Figure 4.10: Threshold choice plots for maximum monthly wind speed.

4.3.8 Declustering

The Mann-Kendall results in Table 4.13 prove that wind observations are dependent which induces clustering. Estimating the parameters of the GPD while ignoring the dependency will yield estimators with small standard errors (Kearns and Pagans, 1997). The results in Table 4.15 conclude that due to an extremal index of 0.65854 at a run length of 1, therefore, wind observations are clustered. A run length of 1 was chosen during the analysis because as run length increases the extremal index does not go beyond 0.82 but rather useful data is lost.

Red coloured observations above a threshold of 7.1 m/s in Figure 4.11 induce

Table 4.15: Extremal Index for declustering of maximum monthly wind speed data.

$\hat{\theta}$	Number of clusters	Run length
0.65854	27	1

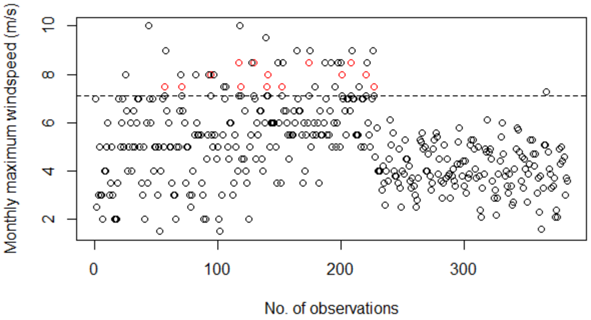

Figure 4.11: Declustered maximum monthly wind speed exceedances.

clustering leading to dependence on data. They are replaced by the value of the threshold and those observations are not used to estimate the parameters of the generalised Pareto distribution.

4.3.9 Parameter estimation and model fitting

Maximum likelihood estimates along with a 95% confidence interval of the generalised Pareto distribution parameters are presented in Table 4.16. The GPD is fitted to both the declustered data and all exceedances.

Table 4.16: Parameter estimates of GPD exceedances using MLE for maximum monthly wind speed data.

	All exceedances		Declustered data	
Parameter	Estimate (SE)	95% CI	Estimate (SE)	95% CI
ξ	-0.576 (0.143)	(-0.855; -0.296)	-0.864 (0.254)	(-1.252; -0.256)
σ	1.784 (0.342)	(1.114; 2.4541)	2.259 (0.932)	(0.967; 3.393)
$\hat{\theta}$	0.65854		1	

For both the declustered data and all exceedances, the value of the shape parameter is less than zero ($\xi < 0$) which suggests that the fitted generalised

Pareto distribution in Table 4.16 provide evidence that the extreme monthly wind speed events are in the Pareto Type II family. Furthermore, 95% CI for the shape parameter confirms that zero is not included in the interval and the intervals are negative, which supports the Pareto Type II family. It can further be seen from Table 4.16 that while ignoring the dependence of observations, the estimates of all exceedances have small standard errors as compared to the declustered data.

4.3.10 Model selection

Now that we have fitted the generalised Pareto distribution, we perform deviance statistics to check which nested distribution of the GPD represents wind speed data well. The results of the likelihood ratio test, AIC and BIC are as follows.

The likelihood ratio test

The test is conducted with the following assumptions:

$H_0 : \xi = 0 \; (Exponential \; distribution)$

$H_1 : \xi \neq 0 \; (Generalised \; Pareto \; Distribution)$

Table 4.17: Likelihood ratio test for the declustered data and all exceedances of maximum monthly wind speed data.

	test statistic	critical value	p-value
Exceedances	11.267	3.8415	< 0.001
Declustered	11.26	3.8415	< 0.001

For both declustered data and all exceedances, the Likelihood ratio test in Table 4.16 produces a very small p-value which is less than, the 0.05 level of significance and likelihood ratio test statistics are greater than the critical value of 3.8415 for both datasets. This provides evidence to reject the null hypothesis which states that all exceedances can be fitted well by the Exponential distribution. We then conclude that both declustered wind speed data and all wind speed exceedances for Mara village are in the GPD family.

Information Criteria

The results of AIC and BIC for model selection are provided in Table 4.18

Table 4.18: AIC and BIC model selection for maximum monthly wind speed.

	All exceedances		Declustered data	
	AIC)	BIC	AIC	BIC
GPD	86.2803	89.7075	61.6713	64.2630
Exponential	95.5474	97.2610	70.9315	72.2273

For both all exceedances and declustered data, GPD is the best fit for maximum monthly wind speed due to small values of AIC and BIC compared to Exponential distribution.

4.3.11 Model checking (Diagnostic plots)

Diagnostic plots are used to check the good fit of the GPD. The diagnostic plots for all exceedances are illustrated by Figure 4.14 whereas the diagnostic plots for the declustered data are shown by Figure 4.15. The data points in Figure 4.14 and Figure 4.15 do not fall closely along the straight line which indicates that GPD is not a good fit. The plots of empirical data quantiles plotted against the quantiles derived from the GPD also show that the data points do not follow the fitted distribution line. It indicates that the GPD is not a good fit for both the declustered wind speed data and all exceedances. The plots of return periods in years against return level show that the selected threshold is reasonable since the points are not deviating away from confidence intervals represented by blue lines. The density plots show that the modelled distribution does not align with the empirical density, which indicates that the GPD is not a good fit for both the declustered wind speed data and all exceedances of wind speed data.

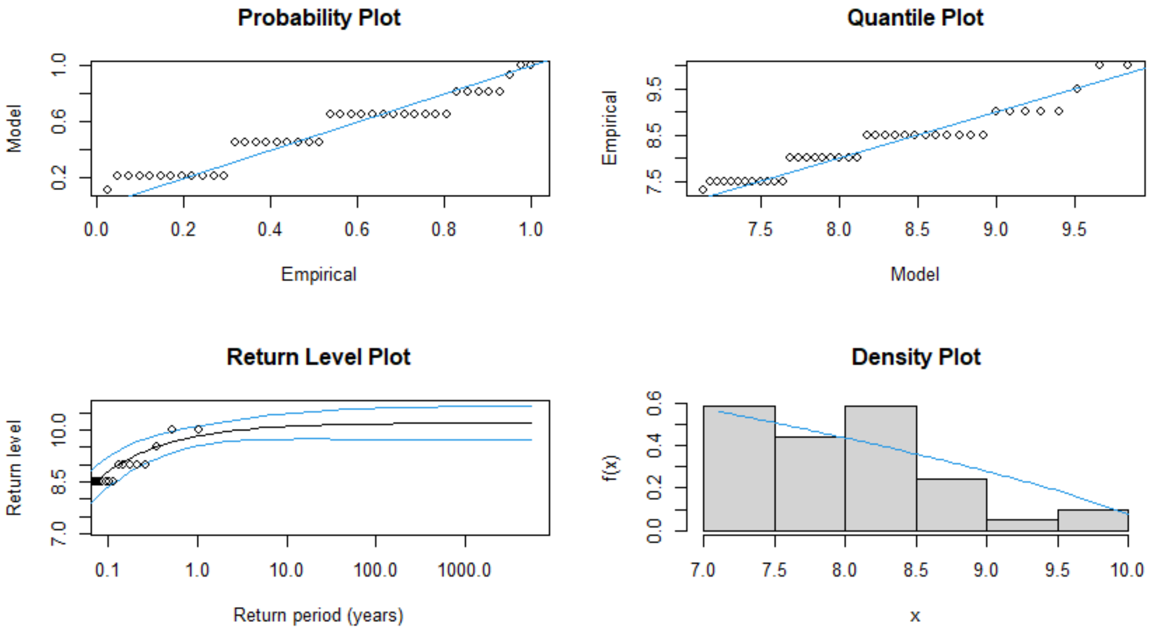

Figure 4.12: GPD diagnostic plots for maximum monthly wind speed exceedances.

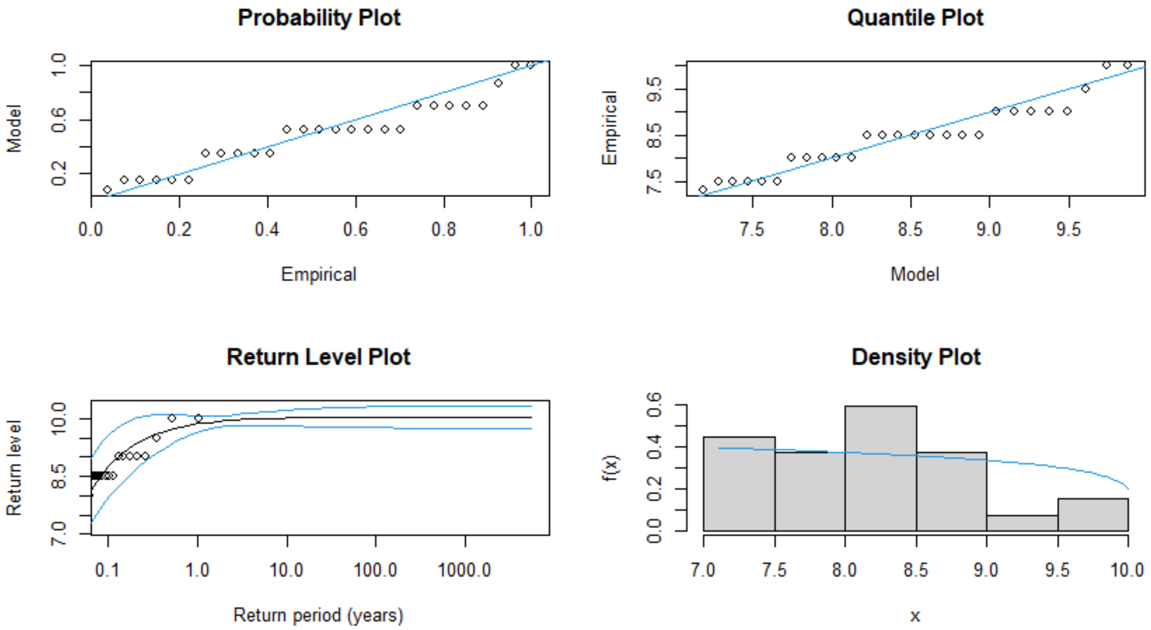

Figure 4.13: GPD diagnostic plots for declustered maximum monthly wind speed data.

4.3.12 Return level estimates

The estimated return levels with a 95% confidence interval corresponding to their return levels are presented in Table 4.17. The results include the return level estimates for all exceedances when ignoring the dependence of data and the return level estimates for the declustered data. The return levels are corresponding to the return periods of 10, 20, 40, 50, and 100 years. The analysis based on declustering resulted in a decrease in the return level estimates relative to the estimates from all exceedances. Fawcett and Walshaw (2006) suggest that the data with return levels increasing slightly compared to the other data is precise and their finding supports the current result that the declustered data is worth making a conclusion with. An estimate of 10.036 m/s suggests the maximum monthly wind speed that will be exceeded once every 100 years.

Table 4.19: Estimated GPD return levels for all exceedances and the declustered data for maximum monthly wind speed.

	All exceedances	Declustered
Return Periods	Estimate (95% CI)	Estimate (95% CI)
10-years	10.100 (-5.8906; 26.0908)	10.015 (-101.4905; 121.5205)
20-years	10.133 (-13.6924; 33.9585)	10.026 (-192.9171; 212.9689)
40-years	10.155 (-25.3458; 45.6561)	10.032 (-359.3300; 379.3939)
50-years	10.161 (-30.2040; 50.5251)	10.033 (-437.8642; 457.9307)
100-years	10.174 (-49.9736; 70.3208)	10.036 (-805.1503; 825.2223)

Chapter 5

Conclusion

5.1 Introduction

This chapter summarises the conclusions of the statistical analysis on the maximum monthly rainfall and wind speed in Mara village in the province of Limpopo, along with recommendations. The conclusion is based on both maximum monthly rainfall and wind speed. Finally, recommendations from the study are discussed towards the end of the chapter.

5.2 Conclusion

The study modelled the maximum monthly rainfall and wind speed at Mara village of the Limpopo province from January 1990 to March 2023. To estimate the frequency of extreme monthly rainfall and wind speed, the generalised Pareto distribution (GPD) is applied to maximum monthly rainfall data and wind speed data at Mara station for 32 years (1990–2023). The secondary data on maximum monthly rainfall and wind speed from January 1990 to March 2023 was obtained from the South African Weather Service (SAWS) and analysed.

The maximum monthly rainfall data for Mara village was found to be stationary according to the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. The Mann-Kendall (M-K) revealed data independence with an increasing monotonic trend. The diagnostic plot for maximum monthly rainfall data provided evidence that the GPD is not a good fit for the data. The comparison of the Exponential distribution to the GPD model using the likelihood ratio test showed that the Exponential distribution fit the maximum monthly rainfall data better compared to GPD. When the suitable model was chosen, the 50-year return level expected to be exceeded once every 50 years was estimated as 153.1325 mm. The maximum monthly rainfall expected to be exceeded once every 100 years was 157.6821 mm.

The Augmented Dickey-Fuller (ADF) test indicated that the maximum monthly wind speed data for Mara village is stationary. A Mann-Kendall Independence test revealed the wind speed data dependence, which induced clustering. The comparison of the GPD and the Exponential distribution using the likelihood ratio test revealed that the GPD fit the data better than the Exponential model. The GPD fit exceedances better than the Exponential distribution for all the exceedances. The return level estimates of the GPD suggested that 10.133 m/s is the maximum monthly wind speed expected to be exceeded once every 20 years. The maximum monthly wind speed of all exceedances is expected to exceed 10.174 m/s once every 100 years. For the declustered data, the return level estimates of the GPD suggested that 10.026 m/s is the maximum monthly wind speed expected to be exceeded once every 20 years. The maximum monthly wind speed of the declustered data is expected to exceed 10.036

m/s once every 100 years.

5.3 Discussion

This study modelled the maximum monthly rainfall and wind speed using the generalised Pareto distribution. The analysis of rainfall indicated that the exponential distribution was the best fit for the data as compared to GPD and it showed a slight increase in maximum monthly rainfall over the years. It is expected that the return level of 157.6821 mm for maximum monthly rainfall will be exceeded once in 100 years. These results are similar to the results of Mashishi (2020) which revealed a slight increase in the average monthly rainfall in South Africa with Weibull distribution as the best fit for the data. Their study revealed that it is expected that the return level of 168.72 mm for maximum rainfall will be exceeded once in 100 years based on the GPD.

The analysis of wind speed indicated a slight increase in wind speed and it showed that Pareto Type II was the best fit for the data. It is expected that the return level of 10.036 m/s for maximum monthly rainfall will be exceeded once in 100 years. The results of the analysis are similar to the results of Diriba and Debusho (2020) which also revealed a slight increase in maximum wind speed when modelling dependency effect to extreme value distributions with application to extreme wind speed at Port Elizabeth, South Africa. From their study, it is expected that the return level of 38.490 km/hr for maximum wind speed will be exceeded once in 100 years.

5.4 Limitations of the study

The choice of the threshold is crucial in GPD analysis, and selecting an appropriate threshold can be subjective. The results of the analysis can be sensitive to the chosen threshold, and there may not be a clear, objective criterion for selecting it. The GPD approximation might not hold and bias could arise if the threshold is set too low. The lower sample size raises the variance of parameter estimations if the threshold is set too high.

5.5 Recommendations

The study makes the following recommendations :

1. The researcher was aware of the missing daily data, which makes the maximum monthly data to be not precise. In order for the analysis to be more accurate, SAWS should try to supply researchers with data that has few missing values.

2. The return level analyses suggested a slight increase in maximum monthly rainfall and wind speed. Engineers should build safe structures that can resist the anticipated slight increase in maximum monthly rainfall and maximum wind speed.

3. To investigate non-stationary extremes and apply extreme value theory to climate data, it is advised that future researchers combine time series with the extreme value theory.

4. It is advised that interested researchers use a Bayesian technique or extreme quantiles to explore better how extreme weather conditions behave in Mara village.

5. The use of multivariate and spatial modelling of wind speed and rainfall to incorporate expert knowledge into estimation to potentially improve the findings.

References

ABDULLAH, R., PRATIWI, E., ABDULLAH, L., DJA'WA, A., TENRIAWARU, A., AND FUJAJA, L. (2019). The role of economic in natural resources development in the City of Baubau. *In IOP Conference Series: Earth and Environmental Science*, volume 235. IOP Publishing, p. 012002.

ACERO, F., GARCÍA, J. A., GALLEGO, M. C., PAREY, S., AND DACUNHA-CASTELLE, D. (2014). Trends in summer extreme temperatures over the Iberian Peninsula using nonurban station data. *Journal of Geophysical Research: Atmospheres*, **119** (1), 39–53.

ALABI, M., TELUKDARIE, A., AND JANSEN, N. V. R. (2019). Industry 4.0 and water industry: A South African perspective and readiness. *In Proceedings of the International Annual Conference of the American Society for Engineering Management*. American Society for Engineering Management (ASEM), pp. 1–11.

ASTAIZA-GÓMEZ, J. (2020). Lagrange multiplier tests in applied research. *Journal de Ciencia e Ingeniería*, **12** (1).

AYUKETANG ARREYNDIP, N., JOSEPH, E., ET AL. (2016). Generalized extreme value distribution models for the assessment of seasonal wind energy potential of Debuncha, Cameroon. *Journal of Renewable Energy*, **2016**.

BAKKER, F. P. (2021). *Characterisation of the South African extreme wind environment relevant to standardisation*. Ph.D. thesis, Stellenbosch: Stellenbosch University.

BARAVALLE, M. AND KÖHLER, J. (2018). On the probabilistic representation of the wind climate for calibration of structural design standards. *Structural safety*, **70**, 115–127.

BHAGWANDIN, L. (2017). *Multivariate extreme value theory with an application to climate data in the Western Cape province*. Master's thesis, University of Cape Town.

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.

BOURQUE, L. B., SIEGEL, J. M., KANO, M., AND WOOD, M. M. (2006). Weathering the storm: The impact of hurricanes on physical and mental health. *The Annals of the American Academy of Political and Social Science*, **604** (1), 129–151.

CASTILLO, E. AND HADI, A. S. (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, **92** (440), 1609–1620.

CHEN, H., BIRKELUND, Y., ANFINSEN, S. N., STAUPE-DELGADO, R., AND YUAN, F. (2021). Assessing probabilistic modelling for wind speed from numerical weather prediction model and observation in the Arctic. *Scientific Reports*, **11** (1), 7613.

CHRISTIE, F. AND HANLON, J. (2001). *Mozambique & the great flood of 2000*. Indiana University Press.

CHUANGCHID, K., SRIBOONCHITTA, S., RAHMAN, S., AND WIBOONPONGSE, A. (2013). Predicting Malaysian palm oil price using extreme value theory. *International Journal of Agricultural Management*, **2** (2), 91–99.

CISAR, P. AND CISAR, S. M. (2010). Skewness and kurtosis in function of selection of network traffic distribution. *Acta Polytechnica Hungarica*, **7** (2), 95–106.

COLES, S. (2001). Classical extreme value theory and models. *An introduction to statistical modeling of extreme values*, 45–73.

COOLEY, D., NYCHKA, D., AND NAVEAU, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, **102** (479), 824–840.

DASH, P. AND PUNIA, M. (2019). Governance and disaster: Analysis of land use policy with reference to Uttarakhand flood 2013, india. *International Journal of Disaster Risk Reduction*, **36**, 101090.

DE OLIVEIRA, M. M. F., EBECKEN, N. F. F., DE OLIVEIRA, J. L. F., AND GILLELAND, E. (2011). Generalized extreme wind speed distributions in South America over the Atlantic Ocean region. *Theoretical and applied climatology*, **104**, 377–385.

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.

DE ZEA BERMUDEZ, P. AND KOTZ, S. (2010). Parameter estimation of the generalized Pareto distribution—Part I. *Journal of Statistical Planning and Inference*, **140** (6), 1353–1373.

DEBUSHO, L. K. AND DIRIBA, T. A. (2016). Bayesian modelling of summer daily maximum temperature data. *In Proceedings of the 4th International Conference on Mathematical, Computational and Statistical Sciences (MCSS'16), Barcelona*. pp. 126–133.

DIRIBA, T. A. AND DEBUSHO, L. K. (2020). Modelling dependency effect to extreme value distributions with application to extreme wind speed at Port Elizabeth, South Africa: a frequentist and Bayesian approaches. *Computational Statistics*, **35** (3), 1449–1479.

DIRIBA, T. A. AND DEBUSHO, L. K. (2021). Statistical Modelling of Extreme Rainfall Indices using Multivariate Extreme Value Distributions. *Environmental Modeling & Assessment*, **26** (4), 543–563.
URL: *https://doi.org/10.1007/s10666-021-09766-6*

DIRIBA, T. A., DEBUSHO, L. K., AND BOTAI, J. (2015). Modelling extreme daily temperature using generalized Pareto distribution at Port Elizabeth, South Africa. *In Annual Proceedings of the South African Statistical Association Conference*, con-1. South African Statistical Association (SASA), pp. 41–48.

DIRIBA, T. A., DEBUSHO, L. K., BOTAI, J., AND HASSEN, A. (2014). Analysis of extreme rainfall at east london, south africa. *In Annual Proceedings of the South African Statistical Association Conference*, con-1. South African Statistical Association (SASA), pp. 25–32.

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

DOSIO, A. (2017). Projection of temperature and heat waves for Africa with an ensemble of CORDEX Regional Climate Models. *Climate Dynamics*, **49** (1-2), 493–519.

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

eNCA (2019). eNCA: South African News.

URL: *https://theworldnews.net/za-news/enca-csir-pinpoints-cause-of-devastating-knysna-fires*

FAWCETT, L. AND WALSHAW, D. (2006). A hierarchical model for extreme wind speeds. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **55** (5), 631–646.

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

FISCHER, E. M. AND KNUTTI, R. (2016). Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, **6** (11), 986–991.

GÁLFI, V. M., BÓDAI, T., LUCARINI, V., ET AL. (2017). Convergence of extreme value statistics in a two-layer quasi-geostrophic atmospheric model. *Complexity*, **2017**.

GILLI, M. AND KËLLEZI, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, **27**, 207–228.

GOUDENHOOFDT, E., DELOBBE, L., AND WILLEMS, P. (2017). Regional frequency analysis of extreme rainfall in Belgium based on radar estimates. *Hydrology and Earth System Sciences*, **21** (10), 5385–5399.

HARRIS, I. (2005). Generalised Pareto methods for wind extremes. Useful tool or mathematical mirage? *Journal of Wind Engineering and Industrial Aerodynamics*, **93** (5), 341–360.

HASAN, H., RADI, N. A., AND KASSIM, S. (2012). Modeling of extreme temperature using generalized extreme value (GEV) distribution: A case study of Penang. *In Proceedings of the world congress on engineering*, volume 1. pp. 181–186.

HOSSAIN, I., KHASTAGIR, A., AKTAR, M., IMTEAZ, M., HUDA, D., AND RASEL, H. (2021). Comparison of estimation techniques for generalised extreme value (GEV) distribution parameters: a case study with Tasmanian rainfall. *International Journal of Environmental Science and Technology*, 1–14.

IPCC (2018). Summary for policymakers, global warming of 1.5° c.

IPCC (2019). Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems.
URL: *https://www.ipcc.ch/srcc*

JARQUE, C. M. AND BERA, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, **6** (3), 255–259.

JARQUE, C. M. AND BERA, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 163–172.

JENTSCH, A. AND BEIERKUHNLEIN, C. (2008). Research frontiers in climate change: effects of extreme meteorological events on ecosystems. *Comptes Rendus Geoscience*, **340** (9-10), 621–628.

KEEF, C., PAPASTATHOPOULOS, I., AND TAWN, J. A. (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the heffernan and tawn model. *Journal of Multivariate Analysis*, **115**, 396–404.

KHAN, A. N. (2011). Analysis of flood causes and associated socio-economic damages in the hindukush region. *Natural hazards*, **59** (3), 1239–1260.

KHAVARIAN-GARMSIR, A. R., POURAHMAD, A., HATAMINEJAD, H., AND FARHOODI, R. (2019). Climate change and environmental degradation and the drivers of migration in the context of shrinking cities: A case study of Khuzestan province, Iran. *Sustainable Cities and Society*, **47**, 101480.

KHUMBANE, T. (2004). Food security: traditional knowledge and permaculture: application of indigenous knowledge systems. *South Africa Rural Development Quarterly*, **2** (4), 44–49.

KJEKSTAD, O. AND HIGHLAND, L. (2009). Economic and social impacts of landslides. *Landslides–disaster risk reduction*, 573–587.

LI, Y., CAI, W., AND CAMPBELL, E. (2005). Statistical modeling of extreme rainfall in southwest Western Australia. *Journal of climate*, **18** (6), 852–863.

MAGNOU, G. (2017). An application of extreme value theory for measuring financial risk in the uruguayan pension fund. *Compendium: Cuadernos de Economía y Administración*, **4** (7), 1–19.

MAPOSA, D., COCHRAN, J. J., ET AL. (2017). Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution. *Statistics and its Interface*, **10** (1), 131–144.

MAPOSA, D., SEIMELA, A. M., SIGAUKE, C., AND COCHRAN, J. J. (2021). Modelling temperature extremes in the limpopo province: Bivariate time-varying threshold excess approach. *Natural Hazards*, **107**, 2227–2246.

MARTINS, A. L. A., LISKA, G. R., BEIJO, L. A., MENEZES, F. S. D., AND CIRILLO, M. Â. (2020). Generalized Pareto distribution applied to the analysis of maximum rainfall events in Uruguaiana, RS, Brazil. *SN Applied Sciences*, **2** (9), 1479.

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.

MASHISHI, D. (2020). *Modeling average monthly rainfall for South Africa using extreme value theory*. Ph.D. thesis, University of Limpopo.

MASINGI, V. N. AND MAPOSA, D. (2021). Modelling long-term monthly rainfall variability in selected provinces of South Africa: Trend and extreme value analysis approaches. *Hydrology*, **8** (2), 70.

MCNEIL, A. J. AND SALADIN, T. (1997). The peaks over thresholds method for estimating high quantiles of loss distributions. *In Proceedings of 28th international ASTIN Colloquium*, volume 23. p. 43.

MEHMET, Ş. AND ÖZCAN, M. (2021). Maximum wind speed forecasting using historical data and artificial neural networks modelling. *International Journal of Energy Applications and Technologies*, **8** (1), 6–11.

MOLAUTSI, S. V. (2021). *Modelling the sporadic behaviour of rainfall in the Limpopo Province, South Africa*. Ph.D. thesis, University of Limpopo.

MULLER, C.-L. (2022). The impact of desertification on agriculture in South Africa. *Stockfarm*, **12** (1), 12–13.

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

NEMUKULA, M. M., SIGAUKE, C., AND MAPOSA, D. (2018). Bivariate threshold excess models with application to extreme high temperatures in Limpopo province of South Africa. *In Annual Proceedings of the South African Statistical Association Conference*, Congress 1. South African Statistical Association (SASA), pp. 33–40.

NEUMAYER, E., PLÜMPER, T., AND BARTHEL, F. (2014). The political economy of natural disaster damage. *Global Environmental Change*, **24**, 8–19.

NKRUMAH, S., MINKAH, R., AND DOKU-AMPONSAH, K. (2017). Extreme value analysis of temperature and rainfall: Case study of some selected regions in Ghana. *University of Ghana space*.

NWS (2021). NWS Enterprise Resource.

OLIVER, U. AND MUNG'ATU, J. (2018). Modelling extreme maximum rainfall using generalized extreme value distribution: case study Kigali City. *International Journal of Science and Research*, **7** (6), 121–126.

OPA (2008). On the path to a sustainable Electricity future.

OUTTEN, S. AND SOBOLOWSKI, S. (2021). Extreme wind projections over Europe from the Euro-CORDEX regional climate models. *Weather and Climate Extremes*, **33**, 100363.

PHILIPPON, N., ROUAULT, M., RICHARD, Y., AND FAVRE, A. (2012). The influence of ENSO on winter rainfall in South Africa. *International Journal of Climatology*, **32** (15), 2333–2347.

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

RAJABI, M. AND MODARRES, R. (2008). Extreme value frequency analysis of wind data from Isfahan, iran. *Journal of Wind Engineering and Industrial Aerodynamics*, **96** (1), 78–82.

RONCOLI, C., INGRAM, K., AND KIRSHEN, P. (2002). Reading the rains: Local knowledge and rainfall forecasting in burkina faso. *Society & Natural Resources*, **15** (5), 409–427.

SAFARI, B. (2022). Modelling extreme rainfall with block maxima and peak-over threshold methods in rwanda. *Research Square*.
URL: *https://doi.org/10.21203/rs.3.rs-1764882/v1*

SARKAR, A., DEEP, S., DATTA, D., VIJAYWARGIYA, A., ROY, R., AND PHANIKANTH, V. (2019). Weibull and generalized extreme value distributions for wind speed data analysis of some locations in India. *KSCE Journal of Civil Engineering*, **23**, 3476–3492.

SARKAR, A., GUGLIANI, G., AND DEEP, S. (2017). Weibull model for wind speed data analysis of different locations in India. *KSCE Journal of Civil Engineering*, **21**, 2764–2776.

SCARROTT, C. AND MACDONALD, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, **10** (1), 33–60.

SIKHWARI, T., NETHENGWE, N., SIGAUKE, C., AND CHIKOORE, H. (2022). Modelling of extremely high rainfall in Limpopo Province of South Africa. *Climate*, **10** (3), 33.

STEPHENSON, D. B., DIAZ, H., AND MURNANE, R. (2008). Definition, diagnosis, and origin of extreme weather and climate events. *Climate extremes and society*, **340**, 11–23.

SULTANA, N. AND TAN, S. (2021). Landslide mitigation strategies in southeast Bangladesh: Lessons learned from the institutional responses. *International Journal of Disaster Risk Reduction*, **62**, 102402.

TIMESLIVE (2019). Search for bodies continues as the death toll approaches 70.
URL: *https://www.timeslive.co.za/outh-africa/2019-04-25-watch-kzn-floods-search-for-bodies-continues-as-death-toll-rises-to-around-70/*

VINING, K. C. (1990). Effects of weather on agricultural crops and livestock: an overview. *International journal of environmental studies*, **36** (1-2), 27–39.

WANG, L., ZHOU, Y., LEI, X., ZHOU, Y., BI, H., AND MAO, X.-Z. (2020). Predominant factors of disaster caused by tropical cyclones in South China coast and implications for early warning systems. *Science of The Total Environment*, **726**, 138556.

WESTFALL, P. H. (2014). Kurtosis as peakedness, 1905–2014. RIP. *The American Statistician*, **68** (3), 191–195.

WESTRA, S., FOWLER, H. J., EVANS, J. P., ALEXANDER, L. V., BERG, P., JOHNSON, F., KENDON, E. J., LENDERINK, G., AND ROBERTS, N. (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics*, **52** (3), 522–555.

WILLIAMS, C. J. R., KNIVETON, D., AND LAYBERRY, R. (2008). Influence of south atlantic sea surface temperatures on rainfall variability and extremes over southern Africa. *Journal of Climate*, **21** (24), 6498–6520.

WRIGHT, D. B. AND HERRINGTON, J. A. (2011). Problematic standard errors and confidence intervals for skewness and kurtosis. *Behavior research methods*, **43**, 8–17.

WU, L. (2014). Extreme event modelling. *Journal of Climate*.

WU, L., ZHANG, J., LU, Q., AND RAHMAN, A. S. (2017). Tourist adaptation behavior in response to climate disasters in Bangladesh. *Journal of Sustainable Tourism*, **25** (2), 217–233.

ZOMMERS, Z. AND SINGH, A. (2014). *Reducing disaster: Early warning systems for climate change*. Springer.

R Codes

```
#Install Packages
install.packages(”evd”)
library(evd)
install.packages(”extRemes”)
library(extRemes)
install.packages(”ggpubr”)
library(ggpubr)
install.packages(”trend”)
library(trend)
install.packages(”tseries”)
library(tseries)
install.packages(”fBasics”)
library(fBasics)
install.packages(”nortest”)
library(nortest)
install.packages(”moments”)
library(moments)
install.packages(”ggplot2”)
library(ggplot2)
install.packages(”dplyr”)
library(dplyr)
install.packages(”Kendall”)
library(Kendall)
```

```
install.packages(”installr”)
library(installr)
install.packages(”POT”)
library(POT)
install.packages(”ismev”)
library(ismev)
install.packages(”e1071”)
library(e1071)
install.packages(”eva”)
library(eva)
install.packages(”extremefit”)
library(extremefit)
install.packages(”xts”)
library(xts)

#Read rainfall dataset
library(readxl)
rainfall < − read-excel(”C:/Users/MR ML SEBOLA/Desktop/MAX_RAINFALL1.xlsx”)
View(rainfall)
attach(rainfall)

#Convert exponential values to decimals
options(Scipen = 999)
#Descriptive Statistics
summary(rainfall$Observations)
basicStats(rainfall$Observations)

# Exploratory data analysis
plot(rainfall$Observations, ylab = ”Monthly maximum rainfall (mm)”, xlab =
”Time(Years)”)
hist(rainfall$Observations, main = ”Monthly maximum rainfall (mm)”, xlab =
”Daily rainfall(mm)”, col = ”blue”)

#Test for normality
jarqueberaTest(rainfall$Observations)
#Test for stationarity
adf.test(rainfall$Observations)
kpss.test(rainfall$Observations)
pp.test(wind$Observations)
#Test for independence
MannKendall(rainfall$Observations)
mk.test(rainfall$Observations)
#threshold selection
#mean residual life plot
mrlplot(rainfall$Observations, main = ”Mean Residual Life Plot”)
#threshold choice plot
tcplot(rainfall$Observations)

#fitting the gpd to the all exceedances
exceedances < −fevd(rainfall$Observations, threshold = 50, type = ”GP”)
print(exceedances)
ci(exceedances, type = ”parameter”)
# Fitting the model GPD
excee < − gpd.fit(rainfall$Observations, 50)
# Model diagnostic
gpd.diag(excee)
#Profile log likelihood
gpd.profxi(excee, xlow = -0.5, xup = 0.3)
```

```
#fitting exponential model to all exceedances
exp-exceedances < − fevd(rainfall$Observations, threshold = 50, type = ”Exponential”)
print(exp_exceedances)
#likelihood ratio test
lr.test(exceedances,exp_exceedances)
#return levels for all exceedances
return.level(exceedances, return.period=c(10,20,40,50,100))
#confidence interval for return level estimates
ci(exceedances, return.period=c(10,20,40,50,100))

#extremal index
extremalindex(rainfall$Observations, threshold = 50, run.length = 2)
extremalindex(rainfall$Observations, threshold = 50, run.length = 3)
extremalindex(rainfall$Observations, threshold = 50, run.length = 4)
extremalindex(rainfall$Observations, threshold = 50, run.length = 5)
extremalindex(rainfall$Observations, threshold = 50, run.length = 6)

#Read wind speed dataset
library(readxl)
wind < − read_excel(”C:/Users/MR ML SEBOLA/Desktop/MAX_WINDSPEED1.xlsx”)
View(wind)
attach(wind)
plot.ts(wind$Observations)

#Convert exponential values to decimals
options(scipen = 999)
#Descriptive Statistics
summary(wind$Observations)
```

```
basicStats(wind$Observations)

#threshold selection
#mean residual life plot
mrlplot(wind$Observations)
#threshold choice plot
tcplot(wind$Observations)
#combine threshold choice plots into one figure
init < − par(no.readonly = TRUE)
par(mfrow = c(2, 2))
tcplot(wind$Observations)
par(init)

#extremal index before declustering
extremalindex(wind$Observations,7.1, method = ”run”, run.length=1)
#declustering
block< −wind$Years
dataset< −decluster(wind$Observations, threshold=7.1, method= ”run”, group=block)
dataset
plot(dataset, col=”red”, main =”Declustered data”, ylab = ”Monthly maximum
windspeed (m/s)”, xlab = ”No. of observations”)
#extremal index after declustering
extremalindex(dataset,7.1, method = ”run”, run.length=1)

#fitting the gpd to the all exceedances
exceedances< −fevd(wind$Observations, threshold = 7.1, type = ”GP”)
print(exceedances)
plot(exceedances)
ci(exceedances, type = ”parameter”)
```

```
# Fitting the model GPD
excee < − gpd.fit(wind$Observations, 7.1)
# Model diagnostic
gpd.diag(excee)
#Profile log likelihood
gpd.profxi(excee, xlow = -0.881, xup = -0.2, conf = 0.95)
#fitting exponential model to all exceedances
exp_exceedances< −fevd(wind$Observations, threshold = 7.1, type = ”Exponential”)
print(exp_exceedances)
#likelihood ratio test
lr.test(exceedances,exp_exceedances)
#return levels for all exceedances
return.level(exceedances, return.period=c(10,20,40,50,100))
#confidence interval for return level estimates
ci(exceedances, return.period=c(2,5,10,20,40,50,100))

#fitting the gpd to the declustered data
declu_data< −fevd(dataset, threshold = 7.1, type = ”GP”)
print(declu_data)
plot(declu_data)
ci(declu_data, type = ”parameter”)
# Fitting the model GPD
excee1 < − gpd.fit(dataset, 7.1)
# Model diagnostic
gpd.diag(excee1)
#Profile log likelihood
gpd.profxi(excee, xlow = -0.9296 , xup = -0.3, conf = 0.95)
#fitting exponential model to the declustered data
```

```
exp_declu< −fevd(dataset, threshold = 7.1, type = ”Exponential”)
print(exp_declu)
#likelihood ratio test
lr.test(declu_data,exp_declu)
#return levels for all exceedances
return.level(declu_data, return.period=c(10,20,40,50,100))
#confidence interval for return level estimates
ci(declu_data, return.period=c(10,20,40,50,100))
```