UNIVERSITY OF GHANA

COLLEGE OF BASIC AND APPLIED SCIENCES

DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

REVISITING THE APPLICATION OF EXTREME VALUE THEORY TO THE MANAGEMENT OF A HYDROELECTRIC DAM

BY
EMMANUEL ANNOR
(10877122)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL IN STATISTICS DEGREE

JUNE, 2023

DECLARATION

I, Emmanuel Annor, declare that the work provided in this thesis, "Revisiting the application of extreme value theory to the management of a hydroelectric dam" is mine. I certify that:

- I completed entirely or primarily while pursuing a research degree at this institution.
- When I consult other people's published work, I always give credit where it's due.
- I always cite the source when I quote from other people's work. This thesis is totally my work, with the exception of such quotes.
- Where other people's contributions are involved, every attempt is taken to indicate this clearly, with proper citations.
- Where the thesis is based on cooperative work by myself and others, I have specified who did what and how much I contributed.

Signature:

Date: 14th June, 2023

I certify that the foregoing claims are true to the best of my knowledge as the supervisor of this candidate's thesis.

Supervisor: Dr. Richard Minkah

Signature:

Date: 14th June, 2023

Co-supervisor: Dr. Eric Ocran

Signature:

Date: 14th June, 2023

ABSTRACT

The Akosombo hydroelectric dam accounts for over a third of the electricity generation in Ghana. The amount of electricity produced depends on the amount of water in the dam. Therefore, studying the tail behaviour of the dam's water level is crucial given the country's rising demand for energy and the strain that this increased demand places on the dam. For engineers and coastal development planners, determining the likelihood that the water level of the Akosombo dam may rise due to heavy rains is crucial as it can lead to flooding. In this study, Extreme Value Theory was to model the tail behaviour of the water levels of the Akosombo dam. Truncation which is introduced naturally by the height of the dam was incorporated. The possibility of exceeding high-water levels that could cause flooding and its effects, as well as their associated return periods were also estimated. An evaluation of the dam water level data's domain of attraction served as the study's starting point. The data were fitted using the Generalized Extreme Value distribution (GEV) and the Generalized Pareto distribution (GPD). To account for potential truncation at very high-water levels, the Right-Truncated Peaks-Over-Threshold (RT-POT) Distribution was fitted to the data. The parameters of the GEV and GPD distributions were estimated using the Maximum Likelihood (ML) and Bayesian estimation methods. The parameters of the RT POT distribution were also estimated using the Maximum Likelihood (ML) and Hill estimation methods. The results show that Akosombo dam water level data tail distribution has a negative shape parameter ($\gamma < 0$), which places it in the Weibull domain of attraction. Both estimation methods yielded remarkably similar estimates. Several exceedance probabilities for various levels of the dam are also estimated. The results show that it is not conceivable for the dam's water level to rise over its 278-foot maximum operating water level. Therefore, it is very unlikely for the water level to rise above the crest of the dam under the prevailing operating conditions.

DEDICATION

I am pleased today to live to see this work completed. This work is affectionately dedicated

To: GOD be all the Glory-without Him I would not have reached this far

And to: Dr. Richard Minkah and Dr. Eric Ocran, for their supervision.

My parents, for their love and support.

ACKNOWLEDGEMENT

This project work was made possible by the cooperation and assistance of a number of people, who allowed me to gain far more than the scholastic components of the project work could have provided. I am grateful to all of them, and I especially appreciate the following individuals:

The Almighty God, for your mercy, care, and support in allowing me to pursue this rewarding study. I am grateful, and may His holy name be exalted forever. Dr. Richard Minkah and Dr. Eric Ocran, who supervised and taught me the principles of assessment in order to determine project efficiency and effectiveness, deserve my heartfelt gratitude. I owe them a debt of gratitude for going to such tremendous lengths to keep me on track. Their support and cooperation were vital to the project's success.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riteria
AMS	**A**nnual **M**aximum **S**eries
BIC	**B**ayesian **I**nformation **C**riteria
CLT	**C**entral **L**imit **T**heorem
DIDR	**D**evelopment **I**nduced **D**isplacement and **R**esettlement
EVI	**E**xtreme **V**alue **I**ndex
EVT	**E**xtreme **V**alue **T**heory
GEV	**G**eneralised **E**xtreme **V**alue
LRT	**L**ikelihood **R**atio **T**est
MCMC	**M**arkov **C**hain **M**onte **C**arlo
ME	**M**ean **E**xcess
MLE	**M**aximum **L**ikelihood **E**stimation
PSP	**P**arameter **S**tability **P**lot
PWM	**P**robability **W**eighted **M**oments
RT-POT	**R**ight **T**runcated **P**eaks **O**ver **T**hreshold
RT-TDGPD	**R**ight **T**runcated **T**runcated **D**ouble **G**eneralized **P**areto **D**istribution
TGEVD	**T**runcated **G**eneralized **E**xtreme **V**alue **D**istribution
TGPD	**T**runcated **G**eneralized **P**areto **D**istribution

Chapter 1

Introduction

This thesis seeks to revisit the application of extreme value theory in the management of a hydroelectric dam. Extreme Value Theory is a branch of statistics concerned with the extreme deviations from the median of probability distributions. A dam is a large infrastructure built on a river or stream to impound water to form a reservoir. This reservoir can be used to control flooding, generate electricity, provide water for irrigation, etc.

This chapter is organized into five core sections. The first section provides a comprehensive background. The second section describes hydroelectric dams and how they operate, as well as electricity generation in Ghana. Subsequent sections describe the problem, along with suitable justifications, research significance, scope and contributions of the study and conclude with the source of data.

1.1 Background and Problem Statement

Extreme Value Theory (EVT) has over the past five decades developed into one of the most essential statistical tools applied in disciplines such as meteorology, finance, insurance and geology.

Minkah (2016) defines EVT as "a branch of statistics that examines the extreme deviations from the median of probability distributions". Conventional statistical analysis techniques heavily make use of measures of central tendency (i.e. mean, mode and median). However, these measures are unable to comprehensively describe the tail behaviours of extreme observations, which is vital in modelling the extremal behaviour of observed rare events such as floods, and earthquakes, among others.

Hence, EVT is needed to study the likelihood of rare events in a given sample of a random variable.

A practical example of the use of EVT is the study of river levels over periods. Here, our interest is mainly in the levels that are either too high or too low. Too high levels could lead to flooding, while too low levels could lead to the dry-out of the river. Similarly, EVT can be used in the management of a hydroelectric dam. Too much water could lead to flooding of neighbouring towns. On the other hand, too low water levels could lead to the water drying out. Hence in both situations, management must be able to track and control the water levels in other to ensure the smooth run of the dam.

Extreme events are rare and often accompanied by many unforeseen outcomes. This has instigated the rising need to model the occurrence of these extreme events. Many methods of estimation using extreme quantiles and corresponding return periods have emerged over the years to model the occurrence of these extreme events.

1.1.1 Electricity Generation in Ghana

Electricity serves as the cornerstone for many modern societies. It is generated by natural gas, coal, and nuclear energy. Other renewable sources of electricity include hydropower, biomass, wind, geothermal, and solar power. Most developing countries depend heavily on hydropower.

Hydroelectric dams have been one of the highly distinguished renewable methods of electricity generation for decades. Ghana's Akosombo dam prides itself on providing 32.14% of the country's electricity (Energy Commission, 2016). The dam is equipped with six 170 megawatts (230,000 hp) turbines, producing 1,020 megawatts. It is a colossal and strong rock-filled embankment dam, built on the Volta River, which routes its course from northern Ghana right through to the south. The dam was built in such a way that flooded parts of the Volta River formed a lake (i.e. Volta lake) behind the dam.

After claiming independence, Dr. Kwame Nkrumah embarked on several industrial projects. This heavy industrialization phase of Ghana required a constant electricity supply. Paramount among these industries was the aluminium industry. In 1962, the government of Ghana decided to use the Akosombo gorge to construct a hydroelectric dam to serve Ghana and its neighbouring countries with electricity.

Although the construction of the dam had many advantages, it must be stated that it had several drawbacks, many of which the project did not envisage. The river that had served the communities along its banks for years was going to become a threat to their existence. Worldwide estimates show that about 80 million people are displaced by dam projects (IDMC, 2017). Ghana is no exception. Although environmentally benign, the development of the Akosombo dam caused Development-induced displacement and resettlement (DIDR), which lead to the inundation of 15,000 homes and 740 villages and the resettlement of 80,000 people, along with their livelihoods (Mettle, 2011). However, the dam had other good prospects. The dam improved fishing, river transportation, farming activities and drinking water. Thus, alleviating poverty and fostering sustainable development. Due to such displacements, regular evaluations must be made on the water levels in the Akosombo dam to control spillage from the dam.

Displacements, along with other issues such as drying out of the dam are growing concerns of many authors. Minkah (2016) demonstrated using EVT that, increasing the capacity of the dam would increase the maximum operating water level. Thus, increasing the average time period between floods. Ocran et al. (2017) also fitted the generalized extreme value (GEV) distribution to estimate the bounds (upper and lower) of the water levels in the Akosombo dam. They also computed the exceedance probabilities for some very low, or very high water levels. Their results indicated that the water level of the Akosombo dam cannot go above the maximum operating water level (i.e. 278 feet). In addition, they demonstrated that the probability of the dam reaching its minimum water level is rare.

How dams Operate

A dam is made up of two sides (the inlet and the outlet). Between these two sides is a continuous flow of water which is significantly above sea level. There are two streams of water located at the head (upstream) which contains the headwater and the tail (downstream) which contains the tailwater of the dam. The dam serves as a reservoir for storing river water. To generate electricity, water from the reservoir flows through turbines, rotating their blades. Water is released depending on electricity demand. The movement of the blades converts potential energy into mechanical energy, which electrical generators use to produce electricity.

1.1.2 Truncation

Truncation in data occurs when observations below and/or above a specific threshold are not included in the data. Probability tails are sometimes bounded naturally, such as in the cases of the Maximum Possible Loss in insurance treaties and the Probable Maximum Precipitation (PMP) in meteorology. For other instances of truncation occurring naturally, see Aban et al. (2006). In our case, the crest of the dam serves as a natural truncation point. The truncated generalized Pareto distribution (TGPD) is a probability distribution that is often used to model the behaviour of extreme values (e.g., maximum or minimum values) in a dataset. It is a generalization of the Pareto distribution, which is commonly used to model the distribution of income or wealth. The TGPD is characterized by three parameters: the shape parameter, the scale parameter, and the lower truncation point. The shape parameter determines the shape of the distribution, the scale parameter determines the scale of the distribution, and the lower truncation point determines the minimum value that the distribution can take on.

In terms of applications, the truncated generalized Pareto distribution has been used in a variety of contexts, including insurance, finance, and hydrology. For example, it has been used to model the distribution of flood magnitudes, wind speeds, and other environmental variables. It has also been used in finance, where it has been applied to model the distribution of extreme stock returns and other financial variables.

There are several extensions of the TGPD, including the truncated double generalized Pareto distribution (TDGPD) and the truncated generalized extreme value distribution (TGEV). These extensions have been used to model more complex datasets and to provide a better fit to data with multiple modes or skewness.

There are several limitations to the use of the Truncated Generalized Pareto Distribution. One limitation is the assumption of independence between the data points and the assumption of a single mode in the distribution. These assumptions may not always hold in real-world datasets, which can impact the accuracy of the model. Another challenge in using the TGPD is dealing with censored data, where the value of some observations is unknown. This can occur, for example, when the maximum value that can be observed is capped at a certain level. In these cases, special techniques must be used to estimate the parameters of the TGPD. In addition, the TGPD may be sensitive to the choice of parameters, and the estimates of these parameters may be subject to uncertainty. Finally, the TGPD may not accurately model the distribution of extreme values in certain types of datasets, and other distributions may be more appropriate in certain cases.

1.1.3 Research Problem and Justification

Due to the lack of substantial data concerning extreme events, classical statistical techniques are not appropriate, and hence, EVT which focuses on the tails of underlying distributions comes in handy. In Ghana, there have been a number of floods arising from the spillage of water from the Akosombo dam and a shortage of electricity due to low water levels. Many of these could be predicted using EVT. In addition, the failures of some of the world's largest dams such as the Oroville, Sanford and Edenville dams instigate the need to study the performance of the Akosombo dam, using techniques that take truncation at high levels into consideration.

Several authors have attempted to model the water level of the Akosombo dam. However, none considered the truncation introduced naturally by the dam level. The recorded data may be truncated naturally because it is constrained within a specific range set by the maximum levels of the dam. If the dam has a maximum capacity, the collected data will only include observations within that capacity. Any measurements exceeding the maximum level will not be included in the data. The study aims to apply extreme value theory for better management of a hydroelectric dam, with a focus on estimating the maximum water level that could cause overflow of the Akosombo dam, along with other parameters of interest.

1.2 Research Objectives

In summary, the main objective of the study in its broadest terms is to provide a model that would describe the extreme elevation of the Akosombo dam, taking into account truncation. Specifically, this study seeks to:

i Compare the proposed model to existing distributions (i.e. GEV and GPD distributions)

ii Estimate high quantiles (truncated and underlying non-truncated), exceedance probabilities and return periods.

1.3 Significance

Earlier research on the Akosombo dam considered the maximum and minimum water levels required for the smooth functioning of the dam. However, no research has considered water levels in the reservoir that would rise above the dam's crest, breaching the integrity of the dam. This research aims to determine the water levels that would overflow the parapet walls of the dam. Consequently, this overflow would cause overtopping of the dam, causing dam failure and massive erosion downstream.

1.4 Scope and Contributions of the Study

This project responds to the need of Ghanaian researchers that have embarked on studies on the elevation level of the Volta lake of the Akosombo dam for management and socio-economic purposes, yet lacked the time and resources to look into alternative perspectives on the issue. The proposed models will enrich literature, as well as the management of the dam with deeper knowledge on how to ensure the dam runs smoothly.

1.5 Data Source

Data for the study comprises of water levels of the Akosombo dam, between the periods January 1965 and December 2013.

Chapter 2

Literature Review

This chapter is tailored towards a brief review of extreme value theory and an intuitive discussion of its application to the management of a hydroelectric dam. The surges in economic growth in many countries have drastically increased electricity demand. This increase is particularly observed in developing countries like Ghana. Among the many renewable methods of electricity generation, hydroelectric power is mostly used in developing countries. Ghana's Akosombo dam provides electricity for the country and trades its surplus to neighbouring countries. The smooth functions of dams necessitate the need for management and independent researchers to investigate possible extreme events.

2.1 Extreme Value Theory

To extrapolate tail distributions and subsequently estimate extreme quantiles, extreme value analysis (EVA) offers many plausible statistical concepts (Blanchet et al., 2020). Extreme value theory antedates the Swiss Mathematician Nicolaus I Bernoulli's work in the eighteenth century. Bernoulli studied the mean maximum distance from the origin of a normal distribution in an attempt to estimate the lifetime of the last survivor among a number of men (Kotz & Nadarajah, 2000).

Two distinct but complementary works by Von Bortkiewicz (1922) appear to have sparked a vigorous advancement of the EVT. These studies broadened Bernoulli's work on determining the expected value of the maximum of a set of variables with independent but identical distributions by estimating the range and the ordered absolute errors. Von Bortkiewicz (1922) established the concept of greatest values from a normal distribution in his article.

Von Mises (1923) and Dodd (1923) extended the results of Von Bortkiewicz (1922) by assessing the expected value and median of the said distribution respectively. Additionally, they diligently acquired numerous legitimately asymptotic results.

The technique sparked the interest of many astronomers since they needed a method to decide whether or not to dismiss extreme values.

Fisher and Tippett (1928) are considered to be the pioneers of EVT by identifying all possible extreme value distributions(i.e. Gumbel, Fretchet and Weibull distributions). However, Gnedenko (1943) is credited for formerly unifying the types of theorems proposed by Fisher and Tippett (1928). The three distributions are jointly referred to as the ”extreme value distribution”.

Frank (1954) studied EVA, establishing the Von Mises conditions on the hazard rate to produce situations under which extreme value analysis behaviour can be observed (Kotz & Nadarajah, 2000).

The EVT paradigm was mainly conceptual, until the 1950s. Gumbel (1958) used theories established by Gnedenko to model the extremal behaviour of real physical phenomena. The Fisher–Tippett–Gnedenko theorem(EVT), popularly known as the first theorem in extreme value theory considered a full range of data. Works by Balkema and De Haan (1974) developed existing theories into the second theory in extreme value theory. The Pickands–Balkema–De Haan theorem focuses on data values that exceed a certain threshold, instead of the full range of observations used in the first theorem.

Another significant contribution to the development of EVA is Haan's paper in 1970, on “Regular Variation and its Application to the Weak Convergence of Sample Extremes” (Beirlant et al., 2004). Extreme value analysis is widespread in many disciplines including economics (Nolde & Zhou, 2021), environmental science (Smith, 1989), and in engineering (Castillo, 2012) among others. Several methods of application and theoretical developments have been introduced since then, some of which will be discussed.

Smith (1989) used extreme value analysis for environmental studies of the ozone layer. Nolde and Zhou (2021) applied the theories of EVT in the evaluation of risk in finance. In the field of energy generation, there have been profound applications of EVT. In the areas of rainfall and temperature in Ghana, Nkrumah (2017) modelled the tail distributions of temperature and rainfall in Ghana based on the Generalized Pareto distribution. In the management of a hydroelectric dam, Minkah (2016) applied extreme value models based on the generalized Pareto (GP) distribution to show that an expansion of the Akosombo dam can reduce the frequency of floods happening. Similarly, Ocran et al. (2017) also fitted the generalized extreme value (GEV) distribution to model the monthly maximum and minimum water levels of the Akosombo dam.

2.1.1 Fields of Applications of EVT (Hydrology)

In hydrology, extreme value theory is often used to analyze extreme precipitation, floods, and droughts, and to estimate the likelihood of their occurrence. Here are some specific examples of how extreme value theory has been applied in hydrology:

- Flood frequency analysis: Extreme value theory is used to estimate the frequency and magnitude of extreme floods, which are defined as floods that have a very low probability of occurring. This information is used to design and assess the reliability of flood protection structures, such as levees and dams. Leščešen and Dolinaj (2019) used the Annual Maximum Series (AMS), a prominent example of the block maxima approach of the extreme value theorem, to provide more accurate flood forecasts for the design and upgrade of flood defence structures in the Pannonian Basin.

- Drought assessment: Extreme value theory provides tools for assessing the probability and return period of droughts, which are defined as periods of low precipitation or low streamflow. This information can be used to plan for water resources management during drought conditions. Skakun et al. (2014) utilized the Poisson-GP (Generalized Pareto) model to estimate and plot return periods for various categories of drought severity. With the help of this model, they were able to map the drought risk using a time series of vegetation health index (VHI) data that was gathered from National Oceanic and Atmospheric Administration (NOAA) satellites. Furthermore, EVT is used to design and size water resources infrastructure such as reservoirs, to ensure they can withstand extreme precipitation events.

- Hydrological risk assessment: In the context of hydrological risk assessment, EVT can estimate the likelihood of extreme hydrological events, such as flash floods, landslides, and debris flows, and assess the potential consequences of these events. This can help in identifying high-risk areas and in developing mitigation and management strategies. By analyzing historical data on flood or drought events, EVT can estimate the parameters of the probability distributions that govern these events and use these parameters to make predictions about the likelihood of extreme events in the future. Researchers have used EVT to estimate the probability of extreme floods and to analyze the spatial and temporal variability of flood risk. For example, Vitanov et al. (n.d.) used EVT to analyze the probability of severe floods in the Mississippi River in the United States and to estimate the likelihood of future floods in these areas. It's important to note that it's often useful to combine EVT with other approaches such as Time Series Analysis, downscaling climate models, etc. In the context of flood forecasting and warning systems, EVT can be used to improve probabilistic flood forecasting, this can be of great use in flash flood warning systems, and it can also help set the warning thresholds and criteria. Furthermore, a plethora of hydraulic infrastructures, such as dams and levees, are designed using the Extreme Value Theory. The theory allows for the estimation of the extreme loads that the structure may be subjected to and the design of the structure accordingly.

- Climate change: Extreme value theory can be used to analyze and project changes in the frequency and amplitude of severe hydrological events in response to climate change This can inform the adaptation strategies for infrastructure and communities and the design of resilience strategy. Also, in regional and global analysis, extreme value theory is used to study the regional distribution of extremes and to detect spatial patterns of the variability of extreme events across a region, thus allowing the understanding of the relationship between climate, land use, and the occurrence of extreme events. EVT has also been used to analyze the potential impacts of climate change on other types of extreme events, such as heat waves and storms. For example, Dosio et al. (2018) used

EVT to estimate the probability of extreme heat events under future climate scenarios and to analyze the potential changes in the frequency and intensity of heat waves. Studies such as MacAfee and Wong (2007) have also used EVT to analyze the potential impacts of climate change on the frequency and intensity of storms, such as hurricanes and typhoons.

These examples illustrate the wide range of applications of extreme value theory in hydrology and how it can inform water resources management and decision-making process. EVT is a helpful tool for managing water resources since it may be used to determine how likely uncommon phenomena are to occur and to evaluate their possible effects. While EVT is good at modelling the extremes it is important to note that EVT alone might not be able to capture all the complexities and factors that might affect hydrological events, so it's often useful to combine it with other techniques and approaches to get a more comprehensive understanding.

2.1.2 Applications in Other Fields

In general, EVT can be applied to any field where data on rare events is available and where the analysis of extreme events can provide valuable insights or information.

Financial risk management: In the context of financial risk management, EVT is used to estimate the likelihood and potential impact of rare but severe events, such as stock market crashes or extreme changes in interest rates, and to analyze the risk of portfolio losses. This information can be used by financial institutions and investors to make better-informed decisions about managing risk and allocating capital. Additionally, EVT techniques can be used to create more precise versions of well-known risk indicators like Value-at-Risk (VaR) and Expected Shortfall (ES), which are frequently employed in financial risk management. EVT can be used to estimate the tail of the probability distribution, which is the extreme events, with this information it can be used to estimate risk measures like VaR, CVA and stress testing. Several studies have used extreme value theory to estimate the probabilities of significant losses in financial markets and to evaluate the performance of risk management strategies. A paper by Demirer et al. (2019) demonstrated using EVT that high levels of political uncertainty do not necessarily cause extreme stock market outcomes, but low levels of political uncertainty are more likely to result in extreme positive stock market returns than high levels of political uncertainty are to result in extreme negative stock market returns.

Engineering and infrastructure: Extreme value theory has been applied in literature to design and assess the reliability of structures and infrastructure, such as bridges, dams, and offshore platforms, that are subjected to extreme loads. The authors of Li and Jones (2019) proposed an EVT model for forecasting substation MD (maximum demand) by utilizing the stability and explanatory power of three common energy factors. The model is flexible enough to adapt to changes in a network configuration as long as they are included in the explanatory variables.

Insurance: A lot of studies have used extreme value theory to model extreme claims, such as those resulting from natural disasters, and to set premiums for insurance policies. The application of the theory aims for a better understanding of the risk and mitigation strategy. Muela et al. (2017) studied the use of conditional extreme value theory (CEVT) to estimate price and liquidity risk and found it to be more effective than standard methods, specifically in the estimation of Value-at-Risk (VaR).

Environmental Science: Extreme value theory has been used to examine the effects of infrequent occurrences like wildfires, earthquakes, hurricanes, and heat waves in order to improve management and decision-making with the goal of minimizing the impact on the environment. Beirlant et al. (2017) utilized EVT's modelling capabilities to simulate the distribution of seismic event magnitudes in the Groningen gas field.

Sports Science: EVT is used to study the extreme performance of athletes and teams in various sports. Vicente (2012) applied concepts of EVT to demonstrate that a 100-meter runner can speed up in the current conditions and possibly lower Usain Bolt's current record by under 9 seconds. Also, Spearing et al. (2019) modelled the swim times of elite swimmers using the concept of EVT.

2.2 Truncation of Dams

As explained in the previous chapter, truncation in data occurs when observations below and/or above a specific threshold are not included in the data. Truncation levels are usually considered rare levels in the field of hydrology and hydroelectric power generation. They require immediate attention to curb any unprecedented damaging events. Dam failures can be caused by a variety of factors, but the most prevalent include overtopping, internal erosion, and instability. When there is not enough spillway capacity to drain surplus water during severe floods, overtopping occurs (Lewin et al., 2003). As a result, the water level of the dam's reservoir would incessantly increase, climbing above the parpet walls of the dam's crest. Without the intervention of harsh emergency measures, the dam would be eventually overtopped, leading to imminent damage and failure.

Spillways are generally considered to be the most expensive components of dams due to their high operations and regular maintenance costs. Albeit spillways go through extensive inspections and maintenance, they are susceptible to a variety of damages. History is littered with many cases of dam failures. Dam failures owing to faulty spillways have occurred at the Oroville, Edenville, and Sanford dams, to name a few.

2.2.1 The failure of the Oroville dam

The Oroville dam is a massive zone earth-filled dam built in 1968, with an impressive height of 770 feet (235 metres). Its main functions were to manage flood flows, generate hydroelectric power, and serve as a water reservoir for Northern California's roughly 20 million residents. In the case of a flood, the dam regulates water flow into the Feather river basin. The water impounded by the dam formed Oroville lake, which is by far the second largest in the United States. The dam has two spillways, a primary spillway and an emergency spillway. Its main spillway has an estimated release capacity of 300,000 cubic feet (8,000 cubic meters) of water per second (Koskinas et al., 2019). The emergency spillway overflows whenever the water level of the reservoir gets above the maximum operating level, consequently preventing the crest of the dam from overtopping. Even with such an astronomical spillway design and capacity, the dam encountered a rather bizarre moment in early 2017. Extreme weather conditions compound poor or incomplete spillway designs, resulting in considerable damage in most situations (Bhattarai et al., 2016). That was the rare case of the Oroville dam.

On the stormy February 7, 2017, the reservoir began to rise above the optimal operating level. Excessive inflows were released by opening spillway gates. Unfortunately, the chutes connecting the spillway gates to the Feather river basin got compromised halfway. Operators of the dam were faced with a tough decision on whether or not to allow small flows of water over the damaged concrete chutes, taking into consideration a possible overflow of the dam. Small flows were released into the main spillways, which eventually moved entire parts of the concrete chute and walls downstream.

Lake Oroville's surface level continued to rise, eventually surpassing and overflowing the emergency spillway crest. Dam operators fully opened the main spillway gates to severely reduce the dam's surface elevation level, risking complete destruction of the spillway, in order to prevent the dam from completely overflowing. Fortunately, the dam's spillways were able to quickly release extra water, preventing the dam from overflowing and failing.

2.2.2 The Failure of the Edenville and Sanford Dams

The Edenville dam was a large earthen embankment dam built in 1924, in Michigan, United States. It impounded water from the Tittabawassee River and the Tobacco River to form Wixom Lake. Its primary purpose was to generate hydroelectric power and control flooding in Midland County. It is connected to the Sanford dam downstream, which impounds water to form Sanford lake. Following several dam failures and the inability of the dam to pass Probable Maximum Flood (PMF), the dam's operational license was revoked in 2018. PMF is the maximum amount of flooding that might be expected.

The days that ushered the dams' failure were somewhat rainy and stormy. Forecasts had predicted heavy rains for May 18^{th}, 2020. As predicted, the rains came in and with it a heavy storm. The deluge eventually increased the level of water in Wixom lake, which drew the attention of the dam's operator. All spillway gates were opened to let out excess floodwaters downstream. However, the level of the water in the lake continued to rise regardless.

Over the next few days, the situation worsened, with the dam's water level reaching its highest level ever recorded, causing significant erosion along the dam's east spillway. The eastern side of the dam suffered extensive erosion as the water level surged to barely a foot from the dam's crest. The dam was breached as a result of the erosion, and water began to flow through it. In a matter of hours, the entire Wixon Lake was drained through the opening.

The water downstream, forming Sanford lake, increased uncontrollably as the Wixon lake drained away. The water level in the Sanford dam eventually surged above the crest, overtopping the dam and causing it to fail. The failures of the two dams resulted in a deluge that wiped out the cities surrounding the dams. Over 10,000 people were inundated before the deluge destroyed over 2500 structures worth thousands of dollars.

2.3 Truncation of Probability Distributions

A truncated distribution is a probability distribution that is restricted to a certain range or interval. It is obtained by removing all the probability mass outside of this range and re-normalizing the remaining probability mass so that it adds up to 1. Truncated distributions are often used in statistical modelling and inference when the data is known to be confined to a certain range, or when certain values are not allowed. For example, a truncated exponential distribution is an exponential distribution with a limited range, such as between two certain values.

Rare incidents such as in Section (2.2), draw attention to the possibility of truncation of probability distributions, such as in Pareto tail modelling.

Aban et al. (2006) obtained the conditional estimator based on Maximum Likelihood Estimation to show that for ordered statistics from an independent and identically distributed sample, the M.L.E derived can be used for Pareto tail modelling. It expands the estimator proposed by Hill (1975), which was used for Pareto-type modelling. Aban et al. (2006)'s estimator covered the case of the Truncated Pareto distribution.

Three decades after Hill's proposed estimator, Nuyts (2010) derived a modification of Hill's estimator by trimming the estimator to make it more robust against outliers.

2.3.1 Commonly Used Truncated Probability Distributions

The exponential family of distributions is a widely used set of both discrete and continuous distributions. They possess very good statistical properties, making them ideal for model fitting.

2.3.1.1 Truncated Exponential Distribution

The truncated exponential distribution is a version of the exponential distribution that is limited to a specific range of values. It is used to simulate scenarios in which the relevant variable cannot have values below a predetermined threshold or above a predetermined maximum. The PDF of the truncated exponential distribution is:

$$f(x) = (1/\lambda) * e^{(-y/\lambda)} \text{ for y in } [\alpha, \beta], \tag{2.1}$$

Where y is the variable of interest, λ is the rate parameter, and α and β are the lower and upper bounds of the truncated range, respectively. The cumulative distribution function (CDF) can be found by integrating the PDF from α to y. The mean and variance of the truncated exponential distribution are different from the non-truncated distribution and can be found by using the CDF and PDF functions.

2.3.1.2 The Truncated Generalized Poisson distribution

The Truncated Generalized Poisson (TGP) distribution is a probability distribution that is a generalization of the Poisson distribution and is often used in modelling count data.It is frequently used for modelling count data that has a lower bound of zero, but an upper bound that is not known. The TGP distribution is defined by two parameters: the mean, which is the same as in the Poisson distribution, and the dispersion parameter, which controls the spread of the distribution. The TGP distribution is often used in applications such as modelling insurance claims, customer arrivals, and financial transactions.

2.3.1.3 Truncated Generalized Pareto Distribution (TGPD)

A truncated GPD is a statistical model that is used to describe the tail behaviour of a distribution. The TGPD is a variant of the GPD, which is frequently used to model extreme values in data sets. The difference between the two is that the GPD is defined over the complete real line, while the TGPD is defined only over a certain range of values, beyond a certain threshold.

The estimation of the parameters of the TGPD is usually done using the maximum likelihood method. Recent literature has focused on developing new methods for estimating the parameters of the TGPD, such as Bayesian methods, and also on comparing the performance of the TGPD with other models for extreme values, such as the GPD.In the Bayesian paradigm of estimation, Verster et al. (2012) fitted a TGPD to the tail of a diamond data, in order to predict high quantiles.

The mean and variance of the truncated GPD are different from the non-truncated distribution and can be found by using the CDF and PDF functions. We must remark that the truncated GPD is only defined for shape parameter $\gamma > 0$.

2.4 Applications of TGPD

The Truncated GPD is a statistical model that is commonly used to model extreme values and is widely used in fields such as finance, insurance, engineering and reliability analysis. New research focuses on the estimation of the parameters of the TGPD, developing new methods and comparing performance with other models. Conventional EVT models do not inculcate the possibility of truncation at very high or very low levels. Thus, becomes a limitation when modelling data with very high or low levels.

Ma et al. (2021) used the Peaks-Over-Threshold, Block Maxima, and Right-Truncated POT Distribution to predict the earthquake frequency and return earthquake periods in several Chinese zones. According to their findings, the right-truncated POT model was the best statistical model. In the field of telecommunication, Couturier and Victoria-Feser (2010) modelled the audience data of a Swiss radio station using the zero-inflated truncated Generalized Pareto distribution and showed that the proposed model accurately represents the non-zero observations. The model can also be applied to data in other fields such as hydrology. In application to hydrology, Minkah (2016) and Ocran et al. (2017) modelled the Akosombo dam data under conventional EVT models, without accounting for possible truncation. This work seeks to bridge that gap in the literature.

Chapter 3

Extreme Value Theory

In this chapter, we consider the various methods to be used in our analysis. In Extreme Value Theory literature, researchers fit either parametric, semi-parametric or non-parametric models for the Generalized Extreme Value Distribution or the Generalized Pareto Distribution.

3.1 Limiting Distribution of Maxima

The statistical theory of extreme values seeks to predict future extremes by analyzing current extremes (Gumbel, 1958). Extreme Value theory dates back to the 1920s with Fisher and Tippett (1928), who are known to have identified all the possible extreme value distributions. Their theorem is regarded as one of the cornerstones of EVT. It stated that for properly centred and scaled partial maxima, the Gumbel, Fréchet and Weibull distributions exhausted all possible limits.

Consider an unknown distribution function F of a random variable for a given sample $(X_1, X_2, X_3, ...X_n)$, we can examine the probability of events that are more extreme than any observed earlier. Let M_n be the partial maxima (i.e. either maximum or minimum), then

$$\mathrm{M}_n = max(X_1, X_2, X_3, \cdots, X_n) \tag{3.1}$$

We emphasize that although the focus of this study is on modelling the behaviour of sample maxima, outcomes for sample minima may be accessed through:

$$\mathrm{M}_n = min(X_1, X_2, X_3, \cdots, X_n) = -max(-X_1, -X_2, -X_3, \cdots, -X_n) \tag{3.2}$$

Intuitively, as $n \to \infty$ for some distributions, M_n would converge to the upper end-point of F, mathematically defined as $x^F = sup\{x \in \Re : F(x) < 1\}$.

The partial maxima, M_n, is bounded by the probability distribution:

$$P(M_n \leq x) = P(X_1 \leq x, \cdots, X_n \leq x) = P(X_1 \leq x) * \cdots * P(X_n \leq x) = \{F(x)\}^n \tag{3.3}$$

where;

$$\{F(x)\}^n = \begin{cases} 1 & x \geq x^F, \\ 0 & x < x^F. \end{cases} \tag{3.4}$$

Classical estimation leads to F. Heuristically, since $\{F(x)\}^n$ is a probability distribution depending on F, the error probability would surge as $n \to \infty$. This affects the robustness of M_n. In addition, it can be observed in Equation (3.4) that $\{F(x)\}^n$ is a degenerate function, which is not ideal for statistical analysis. Thus, there is a search for a family of distributions to characterise $\{F(x)\}^n$, rather than F.

A normalization idea similar to the Central Limit Theorem (CLT) is used to address this.

3.2 Approaches of Extreme Value Theory

There are two fundamental approaches in extreme value theory, based on two different theorems. Based on the works of Fisher and Tippett (1928) and Gnedenko (1943), the first theory of EVT was developed. The theory assumes data to be generated in full range. Here, the full data is divided into long blocks of equal sizes, and the maximum in each block is studied (Gumbel, 1958). Selected observations follow the Generalised Extreme Value Distribution.

Conversely, the second theory was developed by Balkema and De Haan (1974) and Pickands III (1975). Here, observations in data are only used when it crosses a specific deterministic threshold, bringing forth the Peaks-Over Threshold (POT) approach. Selected observations follow the Generalised Pareto Distribution.

3.3 Fisher-Tippett Theorem

Theorem 1: *Considering a sequence of i.i.d. random variables, $X_1, X_2, X_3, \cdots, X_n$ with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and $\sigma^2 \in +\Re$. Then,*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma}} \xrightarrow{d} Z \sim \mathcal{N}(0,\, 1) \tag{3.5}$$

Applying the concept in Equation (3.5) above, we can normalize M_n with normalizing constants, a_n and b_n, where $a_n > 0$ and $b_n \in \mathbb{R}$.

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{3.6}$$

Thus, the distribution F is approximated as:

$$\lim_{n\to\infty} (M_n^*) \tag{3.7}$$

Let $F(x) = Pr(X \leq x)$, Gnedenko (1943) formalised the theory above into the extreme value distribution, stated as:

$$\lim_{n\to\infty} P = \lim_{n\to\infty} F(a_n + b_n x)^n = \left(\frac{max(X_1, \cdots, X_n) - b_n}{a_n} \leq x\right) = G(x) \tag{3.8}$$

where $G(x)$ converges in distribution to either one of the following;

$$\Psi : G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, -\infty < x < \infty;$$

$$\Phi : G(x) = \begin{cases} 0 & x \leq b, \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, & x > b; \end{cases}$$

$$\Omega : G(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{\alpha}\right]\right\} & x < b, \\ 1, & x \geq b. \end{cases}$$

We must remark that a > 0, b and for families Φ and Ω, $\alpha >$ o. The three families of distribution above are jointly classified as the **extreme value distributions**, with type Ψ, Φ and Ω representing the Gumbel, Fréchet and Weibull family of distributions respectively. .

Fisher and Tippett (1928) demonstrated that a non-degenerate distribution function would converge in distribution to one of the extreme value distributions. However, they established no conditions for convergence. Von Mises (1923) and Gumbel (1958) are credited for establishing convergence conditions for the limiting distributions. Convergence was based on the Extreme Value Index (EVI), also known as the shape parameter.

Although the Extreme Value Distribution exhausted all possible limit laws, it was extremely difficult to work with. Complicated methods had to be applied to select a domain of attraction to fit the data. Jenkinson (1955) introduced the Generalised Extreme Value distribution, to serve as a single family for the three families of distributions. Here, we fit a model to the data and the data informs us of the most appropriate domain of attraction. Under some appropriate stationarity and regularity conditions, the G.E.V distribution governs the behaviour of block maxima. Here, blocks can represent months, weeks or years.

3.3.1 The Block Maxima Approach

Block maxima samples are fitted to the Generalized Extreme Value distribution using this method. The Block maxima (or minima) approach partitions a sample of size n into a specified number of non-overlapping blocks B. These blocks usually represent time periods, usually years. However, they may be used to represent other characteristics, such as speed, strength and weakness (Vicente, 2012). Let $B_1, B_2, \cdots, B_n$ be a set of blocks, where

$$\dot{\bigcup_{n\in\mathbb{R}}} B_n \text{ and } B_1 \perp\!\!\!\perp B_2, \cdots \perp\!\!\!\perp B_n$$

Attention is restricted to the maximum of each block, which is used to fit the GEV distribution. This distribution is based on the first theory of Extreme Value Theory. Although this method has excellent asymptotic features, it is well known to waste data because only the maximum in each block is used for model fitting. In order to balance the bias-variance connection, careful consideration must be given to the selection of block lengths. Large block lengths might have fewer sample maxima, which would increase variation and decrease bias. It also applies in reverse. It is a three-parameter model characterised by the following distribution.

$$G(x) = \begin{cases} \left(1+\gamma(\frac{x-\mu}{\sigma})\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ e^{-(x-\mu)/\sigma} & \text{if } \gamma = 0 \end{cases} \tag{3.9}$$

with γ, μ and σ representing the shape, location, and scale parameters of the distribution function respectively. It should be noted that $\mu > 0$, $1+\gamma(x-\mu)/\sigma > 0$, $\gamma \in \mathbf{R}$ and $\mu \in \mathbf{R}$. The parameter γ, referred to as the Extreme Value Index determines the nature of the tail distribution.

Domains of Attraction

The Frechet domain ($\gamma > 0$) has an infinite endpoint. The Gumbel domain ($\gamma = 0$) has a more steep infinite endpoint. The Weibull domain ($\gamma < 0$) is short-tailed, with an infinite right endpoint (Coles et al., 2001).

3.3.2 Statistical Choice for GEV models

In fields where the focus is on extremely rare events, model selection is crucial. The test on γ is of primordial importance in establishing the domain of attraction of a distribution (Minkah, 2016). Here, we consider the following hypotheses.
For $\gamma \neq 0$. Where $\gamma < 0$ or $\gamma > 0$

$$H_0 : \gamma = 0 \;\; \text{against} \;\; H_1 : \gamma \neq 0 \tag{3.10}$$

However, we can utilize the following hypotheses if our goal is to study the condition where $\gamma < 0$.

$$H_0 : \gamma = 0 \;\; \text{against} \;\; H_1 : \gamma < 0 \tag{3.11}$$

The Gumbel test and LRT would be utilized in selecting the most appropriate domain of attraction to fit the data.

3.3.2.1 The Gumbel Test

Consider a random variable $\Upsilon = \Upsilon_1, \Upsilon_2, \cdots, \Upsilon_n$. Let $\Gamma = \Gamma_1, \Gamma_2, \cdots, \Gamma_m$ represent the number of exceedances over a specified threshold ϑ, obtained from Υ. Then, the Gumbel Statistic is obtained as:

$$GSt = \frac{max(\Gamma)}{median(\Gamma)} = \frac{\Gamma_{m:m}}{\Gamma_{[m/2]+1:m}} \tag{3.12}$$

Consider GSt^* which is the Gumbel statistic under equation (3.10) as:

$$GSt^* = \frac{GSt - B_m}{a_m} \xrightarrow[m\to\infty]{d} \varphi$$

where

$$B_m = \frac{\ln(m) + \ln(\ln(2))}{\ln(\ln(m)) - \ln(\ln(2))}$$

For test (3.11), given an significance level α, (H_0) is rejected if $GSt^* \leq G_\alpha$. P-values for the test can be derived using the relation below:

$$p(GSt^*) = \varphi(GSt^*)$$

3.3.2.2 Likelihood Ratio Test (LRT)

The generalized extreme-value distribution has a relatively heavier upper tail when γ < 0. It is crucial to discern between these two conditions when characterizing natural occurrences since the cataclysmic events of a certain intensity may be projected to occur more frequently often if $\gamma < 0$ than if $\gamma = 0$. The LRT can be used to test the validity of the Gumbel model and other GEV models. The LRT in the GEV context is a ratio of the Gumbel distribution to the Generalized Extreme Value Distribution. Consider the likelihood function $\ell(\gamma,\, \mu,\, \sigma_\mu)$ of the GEV.

$$\ell(\theta) = \begin{cases} \ell(0, \mu, \sigma_\mu | \Gamma_1, \cdots, \Gamma_m) & \text{for Gumbel} \\ \ell(\gamma, \mu, \sigma_\mu | \Gamma_1, \cdots, \Gamma_m) & \text{for GEV} \end{cases}$$

The likelihood ratio test statistic is obtained as:

$$LRT = -2(\ell(\text{Gumbel}) - \ell(\text{GEV})).$$

Hence under H_0:

$$LRT \xrightarrow[m\to\infty]{d} \chi^2_{(1)}.$$

However, Hosking (1984) suggested using Bartlett's correction to the χ^2 approximation in order to increase the accuracy of estimation. The improved statistic is as follows:

$$L^* = \frac{L}{1 + \frac{2.8}{m}} \xrightarrow[m\to\infty]{d} \chi^2_{(1)} \tag{3.13}$$

At a significance level, α, H_0 in equation (3.11) is rejected if:

$$L^* \geq \chi^2_{1,1-\alpha} \tag{3.14}$$

P-value for the test is computed using: $p(L^*) = 1 - \chi^2_{(1)}(L^*)$

3.3.3 Estimation Under Generalized Extreme Value Distribution

We can proceed with estimations from the brief description of the GEV distribution in (3.9). Literature is cluttered with many methods of estimation using the GEV distribution. Prominent among them are the Maximum Likelihood (ML), Probability Weighted Moments (PWM), L-Moments and Bayesian methods. The EVI (γ), the scale (σ), and location (μ) parameter estimates from these methods can be used to obtain other parameters of interest. In terms of accuracy, the GEV distribution fits extreme data points very well.

3.3.3.1 Maximum Likelihood Estimation

Under this approach, the parameter estimators ($\hat{\gamma}$, $\hat{\mu}$, $\hat{\sigma}$) of the parameters (γ, μ, σ) are derived by maximizing the GEV distribution's likelihood function.

Consider a random variable $(X = X_1, X_2, \cdots, X_n)$, following the GEV distribution. We can obtain the parameters of interest by forming the likelihood function

$$L(\mu,\, \sigma,\, \gamma|x) = \prod_{i=1}^{n} G(x) \tag{3.15}$$

$$L(\mu,\, \sigma,\, \gamma|x) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\gamma}\text{exp}\big(- [1 + \gamma(\frac{x-\mu}{\sigma})]^{\frac{-1}{\gamma}}\big)\big[1 + \gamma(\frac{x-\mu}{\sigma})\big]^{\frac{-1}{\gamma}-1} & \text{if } \gamma \neq 0 \\ \prod_{i=1}^{n} \frac{1}{\gamma}\text{exp}\big(- \frac{x-\mu}{\sigma}\big)\text{exp}\big(- \text{exp}(-\frac{x-\mu}{\sigma})\big) & \text{if } \gamma = 0 \end{cases} \tag{3.16}$$

When $\gamma \neq 0$, we obtain the log-likelihood function as:

$$L(\theta/x) = -nlog\sigma - (\tfrac{1}{\gamma} + 1)\sum_{i=1}^{n} log\big(1 + \gamma(\tfrac{x-\mu}{\sigma})\big) - \sum_{i=1}^{n} \big(1 + \gamma(\tfrac{x-\mu}{\sigma})\big)^{\frac{-1}{\gamma}} \tag{3.17}$$

and when $\gamma = 0$,

$$L(\mu,\, \sigma,\, \gamma|x) = -nlog\sigma - \sum_{i=1}^{n} \text{exp}\big(- \tfrac{x-\mu}{\sigma}\big) - \sum_{i=1}^{n} \big(\tfrac{x-\mu}{\sigma}\big). \tag{3.18}$$

The estimators of the parameters$(\gamma,\, \mu,\, \sigma)$, are obtained by maximizing equation (3.18) with respect to the individual parameters of interest. For small samples, the parameter estimates suffer from biasedness. However, under certain regularity conditions such as high sample sizes, this approach has the best asymptotic properties such as consistency, efficiency and normality (Coles et al., 2001). Therefore as sample size increases, estimates converge faster to the true parameter estimates.

3.3.3.2 Estimation of Other Parameters

In addition to the parameters estimated above, we also extend our interest to other important parameters such as return levels, return periods and exceedance probability.

Quantiles are functions of a random variable obtained by inverting the distribution function of the original distribution. They are commonly known as "return levels" in hydrology. When assessing the likelihood that specific uncommon events will occur, extreme quantiles are particularly very useful. Estimates from extreme quantiles are utilized in flood control, finance, and insurance, to name a few.

Return levels enlighten us about the odds of extreme events happening, which in our case is water inflows. The return level can be calculated as the $(1-p)^{th}$ quantile, by inverting (3.18). Let Z_p represent the return level (i.e.$(G^{-1}(1-p;\mu,\sigma,\gamma)$.

$$Z_p = \begin{cases} \mu - \frac{\sigma}{\gamma}[\ln(1-p)]^{-\gamma} - 1 & \text{if } \gamma \neq 0 \\ \mu - \sigma \ln\big[-\ln(1-p)\big]\gamma & \text{if } \gamma = 0 \end{cases} \tag{3.19}$$

A return period describes the average time in between events, such as floods earthquakes, etc. It generally indicates the rarity of an extreme event.

The inverse of the return period yields the exceedance probability. Consider a random variable X_p with an extreme value X_n, the probability of exceeding X_n is termed the exceedance probability. Simply put, say we are interested in the exceedance probability for a three-year return period, we can formulate it as $\frac{1}{3}$= 0.333 or 33%. However, we may be interested in finding the exceedance probability over a period of time. We can estimate that using;

Exceedance probability $= 1-(1-p)^n$

where p - exceedance probability; n - time period; $(1-p)$ - non-exceedance probability.

3.3.3.3 Bayesian Estimation

Bayesian methods provide a principle approach to model uncertainties in parameters based on Baye's Rule. Statistical inferences about unknown parameters are made in terms of probability statements using a probability distribution of the parameter under consideration after the data is observed. The Bayesian approach to modelling considers both parameters and a given random variable as random and fixed. The random quantities are modelled with a probability distribution. Uncertainty of the Bayesian methods is expressed through the prior distributions. The Bayesian framework capitalises on the drawback of requiring a restraint on γ of the classical framework. It circumvents constraints on γ and simply and intuitively makes predictions of future observations.

Consider a random variable Y. Suppose that Y is modelled with a probability distribution function $g(y;\theta)$, with θ representing an unknown parameter. The Bayesian approach to making inferences about the unknown parameter θ is as follows.

1. Both θ and Y are assumed to be random.

2. Before observing data, the uncertainty in θ is modelled with an appropriate probability distribution, say $h(\theta)$. This distribution is termed a prior distribution.

3. After the data is observed, $h(\theta)$ is updated into a new distribution say $p(\theta|y)$ using conditional probability as

$$P(\theta|y) = \frac{P(y|\theta) * h(\theta))}{P(y)} \tag{3.20}$$

Where $P(y|\theta)$ – data likelihood often denoted as $L(\theta)$.

$$P(y) = \int_{-\infty}^{\infty} p(y|\theta)h(\theta)d\theta \tag{3.21}$$

$P(y)$ is termed normalization constant/ marginal likelihood/ model evidence. $h(\theta)$ is termed prior distribution.

From the definition of the posterior distribution, the following statements can be made.

- The posterior distribution is proportional to the product of the likelihood of the sample data and the prior distribution.

$$p(\theta|y) \propto p(y|\theta) * h(\theta) \tag{3.22}$$

- $P(\theta|y) \propto \frac{1}{p(y)}$

3.3.3.4 Prior Probability

Priors define the probability of an event before empirical data is collected. It assigns a probability to every possible value of each parameter to be estimated. Priors can be classified into informative, non-informative, weak and improper. García-Pérez (2019) provided evidence that the application of informative priors frequently results in the falsification of data. They continued by explaining that the inclusion of fabricated data, whose statistical properties are dictated by the prior's parameters, can be thought of as prior knowledge. Due to this, this work makes use of the non-informative prior (specifically the Maximal Data Information (MDI) prior)

Non-Informative Priors

Non-informative priors, also termed improper priors, consider the 'entity' itself to be the primary source of information for developing our prior distributions for the scale and shape parameters. As a result, we create priors based on formal criteria and let the data speak for itself. The three most commonly used non-informative priors are the MDI prior, the uniform prior and the Jeffreys' prior. Castellanos and Cabras (2007) demonstrated that among the just mentioned priors, only the latter yields proper posterior distributions, irrespective of the sample size. There is also the tendency for the MDI prior to increase without limit as the shape parameter decreases infinitely. However, this study adopts the MDI prior to constructing the posterior distribution mathematically denoted as:

$$\pi_M(\phi) \propto \exp\left\{E[log f(Y|\phi)]\right\} \tag{3.23}$$

Zellner (1995) obtained the MDI prior by maximizing the difference between the likelihood function and the prior density. Considering the MDI prior density in Equation (3.23), it can be observed that $\pi_M(\phi) \not\propto e^{-\xi(1+\gamma)}$ for $\gamma > \mathbb{R}$, which is a necessary condition for proper posterior densities.

$$\pi_M(\sigma, \gamma) \propto \frac{1}{\sigma} e^{-\xi(1+\gamma)}, \quad \sigma > 0,\ \mu,\ \gamma \in \mathbb{R}, \text{where } \xi = \text{Euler's constant} \tag{3.24}$$

A number of proposals have been developed to nullify the problem of improper posterior distributions. To diffuse this effect, Smith and Goodman (2000) proposed the truncation of the MDI prior. Consider a sample of m ordered thresholds from a GP distribution, the construction of a truncated MDI prior density with constraint, $m \geq 2$ produces a proper posterior density.

The truncated MDI prior in (3.25) was adopted to ensure a proper posterior distribution (Northrop & Attalides, 2016)

$$\pi_M^T(\sigma, \gamma) = \frac{1}{\sigma} e^{-\xi(1+\gamma)},\ \gamma \geq -1 \tag{3.25}$$

3.3.3.5 Posterior Probability

The derivation of posterior probability, which particularly facilitates conditional or Bayesian conclusions about the relevant parameters, is described in this section. The selection of suitable priors, however, is a critical challenge in Bayesian inference. Empirical Bayes, which derive prior variances from the data, can be used to refine the problem. Empirical Bayes requires a hierarchical model. In the absence of a hierarchical system in the observation model, the priors' shape is unknown and is considered to be flat. Maximum Likelihood estimation results from Bayesian computations with flat priors.

The available literature on estimation has still not revealed a perfect and accurate estimator, and though some techniques are comparably superior, they are still associated with some discrepancies. A recent trend of estimation procedures that have yielded superior results over existing ones has been the Bayesian estimation coupled with its advantageous predictive ability (Amin et al., 2015)

Suppose that prior beliefs about our parameters $\theta = (\delta, \gamma)$ can be formulated and described via a probability density function $p(\theta)$ without using the data.

With reference to (3.20), the posterior distribution for the MDI prior is

$$P(\theta|y)_{MDI} \propto I(\theta)P(\theta) =$$

$$\frac{1}{\sigma^{n+1}}\pi(\gamma)\text{exp}\left\{-\sum_{i=1}^{n}\left(1+\frac{\gamma(y_i-\mu)}{\sigma}\right)^{\frac{-1}{\gamma}}\right\}\prod_{i=1}^{n}\left(1+\frac{\gamma(y_i-\mu)}{\sigma}\right)^{1+\frac{-1}{\gamma}}\text{, where } \sigma > 0 \tag{3.26}$$

3.3.3.6 Posterior Distribution Sampling

Where necessary, we will use The Ratio-of-Uniforms (RoU) and the Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution obtained above.

The Markov Chain Monte Carlo (MCMC) Method

When handling the posterior distribution analytically is difficult, Markov Chain Monte Carlo provides a class of algorithms for Bayesian model inference. Markov Chain Monte Carlo (MCMC) methods, such as Metropolis-Hastings and the Gibbs sampler compute posterior moments and probabilities by taking samples from the posterior distribution. These samples are generated by proposing new states in the distribution and accepting or rejecting them based on the probability of the state. We set up a Markov chain with the posterior as its long-run distribution using MCMC techniques. The Markov chain is extended until it reaches the limiting (long-run) distribution. Therefore, after that initial run-in time, any value obtained is close to a random sample taken from the posterior distribution.

The Ratio-of-Uniforms (RoU) Method

Kinderman and Monahan (1977) developed the Ratio-of-Uniforms technique for pseudo-random number sampling from a large variety of statistical distributions. It is based on the idea of generating samples from a uniform distribution and then comparing them to the likelihood of the model to obtain samples from the posterior distribution.

This procedure, a fundamental, but computationally demanding algorithm, was utilized to set up an algorithm to sample from the posterior distributions described in equation (3.43).

By computing $X = \frac{S}{T}$ for a pair (S, T) that is uniformly distributed in the set δ, we can generate a random variable X with density $f(x)$ from the desired density. This is the basis for the ratio of uniforms.

Let $\gamma = (s,\ t) : 0 < t \leq f(s/t)$

The Ratio-of-Uniforms approach for creating random variables is as follows:

1. Generate v and u independently from $U(\theta, b)$ and $U(c, d)$
2. Set $x = \dfrac{v}{u}$ if $u^2 \leq p\left(\frac{v}{u}\right)$ then return to i.

The procedure for choosing the bounds b, c and d is extensively discussed in Kinderman and Monahan (1977). The efficiency of the RoU is given by the reciprocal of the expected number of trials required for generating each variate known as the Acceptance Probability (Barabesi, 1993).

3.3.4 General Comments

We must emphasize that although the Block Maxima approach to EVT is most appropriate for data that are already in blocks (i.e. years, months), the approach has major drawbacks.

- The approach is widely known to waste data since only the maximum of each block is used for model building.
- Extremes may tend to converge in specific blocks.

3.4 Balkema and De Haan Theorem

Theorem: *Let $X_1, X_2, \ldots$ be a sequence of independent and identically-distributed random variables, and let F_t be their conditional excess distribution function. Then, for a large class of underlying distribution function F, and a significantly large threshold, t, the conditional excess distribution function, F_t, is well approximated by the generalized Pareto distribution.*

The classical approach (block maxima approach) makes use of only the maximum observation in each block, thus reducing its efficiency (Nkrumah, 2017). Conversely, the Peaks-over threshold (POT) approach utilizes all the observations exceeding a specified threshold. Balkema and De Haan (1974) proved that the Generalized Pareto Distribution can describe exceedances above a significant threshold. It is a two-parameter model characterised by the following distribution.

$$F_\gamma(z) = \begin{cases} 1-(1+\gamma z)^{-1/\gamma} & \text{for } \gamma \neq 0, \\ 1-e^{-z} & \text{for } \gamma = 0. \end{cases}$$

with support $z \geq 0$ for $\gamma \geq 0$ and $0 \leq z \leq -1/\gamma$ for $\gamma < 0$

The parameter γ classifies the GPD into either one of the following distributions.

- Pareto distribution ($\gamma > 0$)

- Exponential distribution ($\gamma = 0$)
- The short-tailed Pareto distribution ($\gamma < 0$)

3.4.1 Peaks Over Threshold (POT)

The tails of a distribution could take any shape. Let F_t represent the distribution function of the amount by which the random variable X exceeds a certain threshold, say t. Then F_t is termed the conditional excess distribution function of X.
Pickands III (1975) proved that this conditional excess distribution function F_t for larger values of y, follows the GPD distribution for a large class of underlying distribution functions.

The excesses is obtained as $y = x - t$, with the right tail function F_t having bounds $x_F \leq \infty$.

$$F_t = P(\{X - t \leq y | X > t\});\ 0 \leq t \leq x_F - t \tag{3.27}$$

Rewriting F_t in terms of F gives;
Let $x = y + t$, then

$$\mathrm{F}_t(x) = P(X > x) = P(Y > x | Y < t)$$

$$\mathrm{F}_t(x) = P(X > x) = \frac{P(Y > x \cap Y < t)}{P(Y < t)}$$

$$\mathrm{F}_t(x) = P(X > x) = \frac{P(x \leq Y \leq t)}{P(Y < t)} = \frac{P(Y \leq t) - P(Y \leq x)}{P(Y < t)}$$

$$\mathrm{F}_t(x) = \frac{F_Y(x) - F_Y(t)}{1 - F_Y(t)} \tag{3.28}$$

3.4.2 Statistical Choice of GPD Models

The statistical choice for GPD models follows the same procedure as observed in the GEV models in Section 3.3.2.

However, this test places more emphasis on the exponential distribution function which models excesses surpassing a specific threshold and has garnered some interest in literature. Many tests have been developed to fulfil this hypothesis test. Prominent among them are tests by Gomes and van Montfort (1986), Marohn (2000) and Kozubowski et al. (2008). The Gomes and Van Monfort test and the Marohn test are popular tests for comparing the exponential distribution to the Generalized Pareto Distribution. The first mentioned test is a ratio of the maximum and the median of recorded exceedances. The latter test is obtained explicitly using the coefficient of variation. We consider the following hypotheses,
For $\gamma \neq 0$. Where $\gamma < 0$ or $\gamma > 0$

$$H_0 : \gamma = 0 \;\; \text{against} \;\; H_1 : \gamma \neq 0 \tag{3.29}$$

However, we can utilize the following hypotheses if our goal is to study the condition where $\gamma < 0$.

$$H_0 : \gamma = 0 \;\; \text{against} \;\; H_1 : \gamma < 0 \tag{3.30}$$

The Gomes and Van Monfort Test, the Marohn GPD Test and LRT would be utilized in selecting the most appropriate domain of attraction to fit the data.

3.4.2.1 Gomes and Van Monfort Test

Consider a random variable $\Upsilon = \Upsilon_1, \Upsilon_2, \cdots, \Upsilon_n$. Let $\Gamma = \Gamma_1, \Gamma_2, \cdots, \Gamma_m$ represent the number of exceedances over a specified threshold ϑ, obtained from Υ. Using the Gumbel statistic expressed in Equation 3.3.2.1, we can consider GSt^* which is the Gumbel statistic under equation (3.10) as:

$$GSt^* = log(2)\, GSt - log(m) \xrightarrow[m\to\infty]{d} \varphi$$

For test (3.11), given an significance level α, (H_0) is rejected if $GSt^* \leq G_\alpha$. P-values for the test can be derived using the relation below:

$$p(GSt^*) = \varphi(GSt^*)$$

3.4.2.2 Marohn GPD Test

Consider a random sample $Y_1, ..., Y_m$ and a specified threshold u. Let $Z_1, ..., Z_{\nu_{(m)}}$ represent the exceedances over the selected threshold u. Marohn (2000) obtained the optimal test statistic for testing the excesses $Z_j - u$ of the GPd as:

$$T_m = \frac{\nu_{(m)}}{2}\left(\frac{\Theta^2}{\iota(m)(\bar{Z}-u)^2} - 1\right) \tag{3.31}$$

Where;

$$\Theta = \sum_{j=1}^{\nu_{(m)}} (Z_j - \bar{Z})^2$$

$$\bar{Z} = \sum_{j=1}^{\nu_{(m)}} Z_j$$

$\nu_{(m)}$ = rescaled location parameter

(H_0) for test (3.11) is rejected at a significance level α, if $T_m \leq z_\alpha$. The associated p-value for the test can be computed with:

$$p(T_m) = \phi(T_m)$$

3.4.2.3 Likelihood Ratio Test (LRT)

The LRT for the GPD context is a ratio of the exponential d.f. to the GPD. The test statistic is obtained using the same methodology outlined in Section 3.3.2.2. However, Hüsler and Li (2007) suggested using Bartlett's correction to the χ^2 approximation in order to increase the accuracy of estimation. The improved statistic is as follows:

$$L^* = \frac{L}{1+\frac{4}{m}} \xrightarrow[m\to\infty]{d} \chi^2_{(1)} \tag{3.32}$$

At a significance level, α, the null hypothesis in equation (3.11) is rejected following the same methodology as observed in the case of the GEV.

3.4.3 Threshold Selection

Literature is littered with profound methods of selecting optimal thresholds for fitting models for the Generalized Pareto Distribution. The practical choice of the threshold $'t'$ must be taken with caution. An optimal threshold is advised to balance the Mean Square Error (MSE). With every method so far, a core trait is finding a threshold that provides a trade-off between bias and variance. Few excesses (realizations) would yield model parameters with small variations, but consequently a large bias. On the other hand, the converse holds (Tancredi et al., 2006). Recently, Wu and Qiu (2018) have elaborated concise methods for an optimal threshold selection. Some functions and graphical methods have been developed to efficiently choose appropriate thresholds. The Mean Excess (M.E) function, referred to as Mean Residual Life (M.R.L) function in survival analysis was developed by Benktander and Segerdahl (1960). It is typically used to facilitate the selection of a suitable threshold and also to evaluate whether a G.P.D model is appropriate for the excesses. Mathematically, the Mean Excess (M.E) function is computed as:

$$E(X - t|X > t) = \frac{\sigma_\mu + \mu\gamma}{1 - \gamma}$$

Davison and Smith (1990) cited a different approach to estimating the shape parameter, γ, from a sequence of ordered potential thresholds $t_1, t_2, ..., t_k$. They suggested utilizing the Mean Excess Plot to estimate an ideal threshold. The G.P.D model is then employed to fit all excesses above this cutoff. The Mean Excess Plot graphs the mean of the excesses (i.e. $E[X - t|X > t]$) against the ordered statistics. Coles et al. (2001) elaborated that a fair balance for the MSE is ideally the point where a linear pattern starts on the M.E plot. The goal is to determine the lowest threshold at which the plot is almost linear, accounting for the 95% confidence intervals. They also indicated that the level where the parameter estimates are roughly steady as we approach larger thresholds and the GP model's asymptotic features hold is an optimal threshold. However, this threshold selection approach depends on the researcher's subjective view. The pattern of the M.E plot indicates the heaviness of the tail distribution function. An upper pattern suggests a heavy-tailed distribution. On the other hand, a descending pattern stipulates a light-tailed distribution. A horizontal form hints at an exponential distribution. An alternate method for graphically choosing the best threshold is the Parameter Stability Plot (P.S.P.). For a detailed examination of the issues and some intriguing threshold selection techniques, the reader is directed to Beirlant et al. (2004), Coles et al. (2001) and Scarrott and MacDonald (2012).

Estimation Methods Under Generalized Pareto Distribution

The realm of estimation constitutes two fundamental schools of thought. These schools govern the underlying principles for parameter estimation. The Classical (Frequentist) approach expresses uncertainty through asymptotic distributional properties. The Bayesian approach makes probability statements about parameters using conditional distributions derived using Baye's rule. Both philosophical approaches have identical behavioural assumptions. However, they have distinct estimation and interpretation schemes.

Commonly used classical estimation methods include Maximum Likelihood Estimation (MLE) and Probability Weighted Moments (PWM). The need for a "γ" constraint to attain asymptotic characteristics of estimates using this estimating procedure is a significant drawback. MLE requires $\gamma > -\frac{1}{2}$, while PWM requires $\gamma < \frac{1}{2}$. In this thesis, we focus on MLE due to its superior asymptotic properties.

3.4.4 Maximum Likelihood Estimation

For $\gamma \neq 0$, the GPD has a log-likelihood function as;

$$L(\gamma, \sigma | X_i) = -n \ln \sigma + \left(\frac{1}{\gamma} - 1\right) \sum_{i=1}^{n} \ln\left(1 - \frac{\gamma(X_i)}{\sigma}\right), \tag{3.33}$$

and for $\gamma = 0$, the GPD has a log-likelihood function as;

$$L(\mu,\, \sigma,\, 0 | X_i) = -n \ln \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} X_i \tag{3.34}$$

Let $G(x) = f(x_i; \mu,\, \sigma,\, \gamma | x)$

$$L(\theta / x) = \prod_{i=1}^{n} G(x) \tag{3.35}$$

$$\ln L(\mu,\, \sigma,\, \gamma | x) = \prod_{i=1}^{n} \ln G(x) \tag{3.36}$$

$$\frac{d \ln L(\mu,\, \sigma,\, \gamma | x)}{d\mu,\, d\sigma,\, d\gamma} = 0 \tag{3.37}$$

Substituting (3.36) into (3.37), we obtain

$$\sum_{i=1}^{n} \ln\left(1+\gamma\frac{y_i}{\sigma}\right) = \gamma n \tag{3.38}$$

$$(1+\gamma)\sum_{i=1}^{n} \gamma\frac{y_i}{\sigma+\gamma y_i} = n \tag{3.39}$$

Maximizing (3.39) with respect to γ and σ yields MLE estimates for $\hat{\gamma}$ and $\hat{\sigma}$ respectively. However, solving for these parameter estimates would yield a system of equations with no distinct solutions. Hence, we propose using computational approaches to solve them iteratively. For more details on the computational methods for solving such equations, we refer the reader to Prescott and Walden (1983); Hosking et al. (1985) and Smith (1985).

3.4.5 Bayesian Estimation

Bayesian estimation for the GPD distribution follows the same procedures as outlined in Section 3.3.3.3

3.4.5.1 Prior Probability

The MDI prior was constructed by Zellner (1995), by maximizing the difference between the likelihood function and the prior density. Thus, the MDI prior to constructing the posterior distribution is mathematically denoted as:

$$\pi_M(\phi) \propto \exp\left\{E[logf(Y|\phi)]\right\} \tag{3.40}$$

Considering the MDI prior density in (3.41), it can be observed that $\pi_M(\phi) \not\propto e^{-(1+\gamma)}$ for $\gamma > \mathbb{R}$, which is a necessary condition for proper posterior densities.

$$\pi_M(\sigma,\gamma) \propto \frac{1}{\sigma}e^{-(1+\gamma)}, \quad \sigma > 0, \ \gamma \in \mathbb{R} \tag{3.41}$$

A number of proposals have been developed to nullify the problem of improper posterior distributions. To diffuse this effect, Smith and Goodman (2000) proposed the truncation of the MDI prior. The construction of a truncated MDI prior density with constraint, $m \geq 1$ produces a proper posterior density.

The truncated MDI prior in (3.42) was adopted to ensure a proper posterior distribution (Northrop & Attalides, 2016)

$$\pi_M^T(\sigma, \gamma) = \frac{1}{\sigma} e^{-(1+\gamma)}, \; \gamma \geq -1 \tag{3.42}$$

3.4.5.2 Posterior Probability

With reference to (3.3.3.3), the posterior distribution for the MDI prior is similarly obtained as:

$$P(\theta|y)_{MDI} \propto I(\theta)P(\theta) =$$

$$\frac{1}{\sigma^{n+1}} e^{-(1+\gamma)} \prod_{i=1}^{n} \left(1 + \frac{\gamma(x_1 - u)}{\sigma}\right)^{-(\frac{1+\gamma}{\gamma})} \tag{3.43}$$

3.4.5.3 Posterior Distribution Sampling

Where necessary, we will use The Ratio-of-Uniforms (RoU) and the Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution obtained above. The reader is referred to Section 3.3.3.6 for details on these methods.

3.4.6 Estimation of Other Parameters

Estimating other parameters follow the same procedure shown in Section (3.3.3.2), but with the GPD distribution function instead. As used in (3.19), we obtain other parameter estimates of interest for the GPD by inverting (3.36).

3.4.6.1 Extreme Quantiles

Consider a quantile function Φ_p of a random variable Y defined as:

$$P(Y \leq \Phi_p) = p$$

Recall the df of the GPD from (3.4), and λ = some threshold. $F_\gamma(z)$ = p

Hence,

$$F_\gamma(z) = 1 - P(Y > \lambda)\left(1 + \gamma\frac{\Phi_p - \lambda}{\sigma}\right)^{\frac{-1}{\gamma}} = p$$

Solving equation (3.4.6.1) yields:

$$\left(1 + \gamma\frac{\Phi_p - \lambda}{\sigma}\right)^{\frac{-1}{\gamma}} = \frac{1-p}{P(Y > \lambda)}$$

The p^{th} quantile, Φ_p, can then be estimated as:

$$\lambda + \frac{\sigma}{\gamma}\left\{\left[\frac{1-p}{P(Y > \lambda)}\right]^{-\gamma} - 1\right\}$$

3.4.6.2 Return Levels

The return level is calculated as the $(1-p)^{th}$ quantile, by inverting (3.36). Let M_p represent the return level (i.e.$(M^{-1}(1-p;\mu,\sigma,\gamma))$).

$$M_p = \begin{cases} \frac{\sigma}{\gamma}p^{-\gamma} - 1^k & \text{if } \gamma \neq 0 \\ -\sigma \ln p & \text{if } \gamma = 0 \end{cases} \tag{3.44}$$

To estimate the return levels over a period of time, the parameters (γ, μ and σ) in (3.44) are replaced by their sample estimates ($\hat{\gamma}$, $\hat{\mu}$ and $\hat{\sigma}$)

3.4.6.3 Return Period & Exceedance Probability

These parameters are similarly estimated using the functions expressed in 3.19 but with the GPD function instead.

Let x_ϕ be the return period with period ϕ and n the number of events during a specific time frame.

$$P(X > x_\phi) = \frac{1}{n\phi}\,.$$

Hence

$$P(X \leq x_\phi) = 1 - \frac{1}{n\phi}\,.$$

Applying the same concept from Equation (3.4.6.1) gives:

$$F_\gamma(z) = 1 - P(Y > \lambda)\left(1 + \gamma\frac{\Phi_p - \lambda}{\sigma}\right)^{\frac{-1}{\gamma}} = 1 - \frac{1}{n\phi}$$

Hence, the return period (x_ϕ) is established as:

$$x_\phi = \mu + \frac{\sigma}{\gamma}\left\{\left[\frac{1}{n\phi P(Y > \lambda)}\right]^{-\gamma} - 1\right\} \tag{3.45}$$

Exceedance probabilities are computed as (x_ϕ^{-1})

3.5 Truncated POT Distribution

At some levels, every probability distribution can get truncated. This will almost always result in new distributions, rather than one within the same family. If $F(x)$ is the distribution function of a random variable, then $F(y)$ is the distribution function of a new random variable Y, defined as the distribution of X trimmed to the semi-open interval (a, b). Even with the emergence of many methods of fitting tail distributions, like the POT approach, truncation effects are often not captured at high levels. Beirlant et al. (2017) demonstrated and proposed the use of a pseudo maximum likelihood approach to estimate the model parameters. Aban et al. (2006) also used the maximum likelihood approach to model parameters of a truncated Pareto distribution and further proved the existence and uniqueness of the estimators under certain conditions.

Consider a random variable $X = X_{(1)}, X_{(2)}, \cdots X_{(n)}$. Let $X_{(1)} \leq X_{(2)} \leq \cdots X_{(n)}$ be a sample of ordered statistics from X. Assuming a threshold as X_{n-k} and exceedances as X_{n-j+1}. The excesses $(E_{j,k})$ can be found as X_{n-j+1} - X_{n-k}.

Let

- Y - Original data before truncation
- T - Truncation point

- $F_Y(y) = P(Y \leq y)$

- Right Tail Function $= \overline{F}_Y(y) = 1 - F_Y(y)$

where Quantile function- $Q_Y(p) = inf\{y : F_Y(y) \geq p\}(0 < p < 1)$, and tail quantile function $U_Y(v) = Q_Y(1 - \frac{1}{v}(v > 1)$

Then;

$$\mathrm{X}_n \overset{d}{=} Y|Y < T \tag{3.46}$$

$$\begin{aligned}\overline{\mathrm{F}}_T(x) &= P(X > x) = P(Y > x|Y < T)\\ &= \frac{P(Y > x \cap Y < T)}{P(Y < T)}\\ &= \frac{P(x \leq Y \leq T)}{P(Y < T)} = \frac{P(Y \leq T) - P(Y \leq x)}{P(Y < T)}\end{aligned}$$

$$\overline{\mathrm{F}}_T(x) = \frac{\overline{F}_Y(x) - \overline{F}_Y(T)}{1 - \overline{F}_Y(T)} = (1 + D_T)\overline{F}_Y(x) - D_T \tag{3.47}$$

Thus, $D_T = \overline{F}_Y(T)/F_Y(T)$, and it is referred to as the odds of the truncated probability mass under the untruncated distribution Y.

$$P(\frac{X - t}{\sigma_t} > x|X > t) \longrightarrow \frac{(1 + \gamma x)^{1/\gamma} - (1 + \gamma k)^{1/\gamma}}{1 - (1 + \gamma k)^{1/\gamma}}, \tag{3.48}$$

Where k represents the top-order statistics.

Let $E_{1,k} = X_n - X_{n-k}$. Substituting $E_{1,k}/\sigma$ for k in (3.48)

$$logL_{k,n}(\gamma, \sigma) = log\left(\prod_{j=2}^{k} \frac{\sigma^{-1}(1 + \frac{\gamma}{\sigma}E_{j,k})^{-(1/\gamma)-1}}{1 - (1 + \frac{\gamma}{\sigma}E_{1,k})^{-(1/\gamma)}}\right) \tag{3.49}$$

Also, substituting $\tau = \gamma/\sigma$ in (3.49) and solving gives

$$\begin{aligned}logL_{k,n}(\gamma, \tau) = (k - 1)log\tau - (k - 1)log\gamma - \left(1 + \frac{1}{\gamma}\right)\sum_{j=2}^{k} log(1 + \tau E_{j,k})\\ -(k - 1)log(1 - (1 + \tau E_{1,k})^{-1/\gamma})\end{aligned} \tag{3.50}$$

3.5.1 Statistical Choice of Truncated GPD Models

The statistical choice for the truncated GPD models follows the same procedure as observed in the GPD models in Section 3.4.2.

3.5.1.1 Likelihood Ratio Test

Under the RT-POT, we will be performing the LRT for the two truncated GPD models (i.e. $\gamma \neq 0$ and $\gamma = 0$) using their respective likelihoods. For $\gamma \neq 0$. Where $\gamma < 0$ or $\gamma > 0$

$$H_0 : \gamma = 0 \ \text{ against } \ H_1 : \gamma \neq 0 \tag{3.51}$$

However, we can utilize the following hypotheses if our goal is to study the condition where $\gamma < 0$.

$$H_0 : \gamma = 0 \ \text{ against } \ H_1 : \gamma < 0 \tag{3.52}$$

3.5.1.2 Test between Truncated and Non-truncated Distributions

To aid in deciding the type of truncation present, Albrecher et al. (2017) cited a procedure to test between light truncation and rough truncation. The test is based on hypotheses:

$$\begin{aligned} &H_0 : \text{X is not truncated at the tails,} \\ &H_1 : \text{X is truncated at the tails.} \end{aligned} \tag{3.53}$$

For the Hill Estimator:

The test statistic is obtained as:

$$T_{k,n} = \frac{\sqrt{12k}(E_{k,n}(H_{k,n}) - \frac{1}{2})}{1 - (E_{k,n}(H_{k,n}))}, \tag{3.54}$$

where $E_{k,n} = \frac{1}{k}\sum_{j=1}^{k} \frac{X_{n-k,n}}{X_{n-j+1,n}}$ and $H_{k,n}$ = Hill estimator.

H_0 is rejected at an asymptotic level, α, if $T_{k,n} < -z_\alpha$. Associated p-values can be obtained as a product of the test statistic $T_{k,n}$ and the CDF of the standard normal distribution, Φ.

For the Maximum Likelihood Estimator:

The test statistic is obtained as:

$$T_{k,n} = k\left(1 + \hat{\tau}(X_{n,n} - X_{-k,n})\right)^{\frac{1}{\hat{\gamma}_k}}, \tag{3.55}$$

H_0 is rejected at an asymptotic level, α, if $T_{k,n} > \ln(\frac{1}{\alpha})$. Associated p-values can be obtained as the exponential of the negated test statistics $exp^{(-T_{k,n})}$.

3.5.2 Maximum Likelihood Estimation

Using Pseudo maximum likelihood estimation, we obtain $\hat{\gamma_k}, \hat{\tau_k}$ as:

$$\hat{\gamma_k} = \frac{1}{k-1}\sum_{j=2}^{k} log(1+\hat{\tau_k}E_{j,k}) + \frac{(1+\hat{\tau_k}E_{1,k})^{-1/\hat{\gamma}} log(1+\hat{\tau}E_{1,k})}{1-(1+\hat{\tau_k}E_{1,k})^{-1/\hat{\gamma}}} \tag{3.56}$$

$$\frac{1}{k-1}\sum_{j=2}^{k}\frac{1}{1+\hat{\tau_k}E_{j,k}} = \frac{1}{1+\hat{\gamma_k}}\frac{1-(1+\hat{\tau_k}E_{1,k})^{-1-1/\hat{\gamma}}}{1-(1+\hat{\tau_k}E_{1,k})^{-1/\hat{\gamma}}} \tag{3.57}$$

$\hat{\sigma_k}$ can be obtained from (3.57)

3.5.2.1 Truncation Odds

This parameter evaluates the likelihood that the truncated probability mass will remain within the scope of the original underlying distribution. Beirlant et al. (2016) discussed ways of estimating this parameter using the Truncated Hill estimator and the Truncated MLE. With F and T representing the Cumulative Density Function of the original distribution and the upper Truncation point respectively, D_T is mathematically defined as:

$$D_T = \frac{1-F(T)}{F(T)}$$

It is estimated using:

$$\hat{D}_J := max\left\{0, \frac{k}{n}\frac{(1+\hat{\tau_k}E_{1,k})^{-1/\hat{\gamma_k}} - \frac{1}{k}}{1-(1+\hat{\tau_k}E_{1,k})^{-1/\hat{\gamma_k}}}\right\} \tag{3.58}$$

3.5.2.2 Exceedance (Tail) probabilities

The likelihood that a specific value, w, will be surpassed in a given time frame in the future. It is computed using the estimated value of the shape parameter (γ), τ and D_J from (3.56), (3.57) and (3.58) respectively.

$$\hat{p}_J(w) = (1+\hat{D}_J)\frac{k}{n}\left(1+\hat{\tau_k}(c - X_{n-k}, n)\right)^{-1/\hat{\gamma_k}} - \hat{D}_J \tag{3.59}$$

3.5.2.3 Extreme Quantiles

Quantiles, $Q(1-p)$, describe the tails of marginal distributions. In this study, we will examine both quantiles of the truncated and original underlying distributions.

Estimating extreme quantiles of the original distribution Y requires reconstructing the parent distribution itself. Extreme quantiles for the unobserved quantities can then be estimated using estimates of γ, τ and D_J from (3.56), (3.57) and (3.58) respectively. These quantiles can be obtained using:

$$\hat{Q}_Y(1-p) = \hat{Q}_Y(1-[p-(1-p)\hat{D}_J]) \tag{3.60}$$

Considering observations truncated at a specific point, quantiles for the truncated distribution, $Q_J(1-p)$, can be obtained using

$$\hat{Q}_J(1-p) = X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left[\left\{\frac{\hat{D}_J + \frac{k}{n}}{\hat{D}_J + p}\right\}^{\hat{\gamma_k}} - 1\right] \tag{3.61}$$

The upper endpoint intuitively refers to the point where $p = 0$. The maximum point at which truncation occurs. It can be computed with:

$$\hat{T}_k = X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left[\left\{\frac{1-k^{-1}}{(1+\hat{\tau}_k E_{1,k})^{-1/\hat{\gamma_k}} - k^{-1}}\right\}^{\hat{\gamma_k}}\right] \tag{3.62}$$

3.5.3 Hill Estimation

When using Pareto-type modeling, high quantiles are computed using a projection from a fitted regression line through the point $(-\ln(((k+1)/(n+1)), \ln X_{n-k})$ on the Pareto QQ-plot. The slope, $H_{k,n}$, referred to as Hill's estimator is obtained from the fitted regression line on the Pareto QQ-plot. See Albrecher et al. (2017) for more computational details. The Hill estimator is shown in Albrecher et al. (2017) as:

$$H_{k,n} = \frac{1}{k}\sum_{j=1}^{k}(\ln(X_{n-j+1,n}) - \ln(X_{n-k,n})) \tag{3.63}$$

The Hill plot obtained from this estimator can be used to estimate the threshold and tail of a GPD. An adaptation of this estimator suited for the truncated GPD was proposed by Aban et al. (2006) and Beirlant et al. (2016). Nuyts (2010) proposed trimming Hill's estimator to make up for truncation. They suggested using the bounds $1 \leq r < k < n$ rather than $(r = 1)$. The modified Hill's estimator is obtained as:

$$H_{r,k,n} = \frac{1}{k-r+1} \sum_{j=r}^{k} (\ln(X_{n-j+1,n}) - \ln(X_{n-k,n})) \tag{3.64}$$

3.5.3.1 Estimation of other Parameters

Using estimates from the Hill estimator, we can obtain estimators for other parameters, just as derived using the MLE.

The odds of truncation can be computed as:

$$\hat{D}_J = \left\{ \frac{k+1}{n+1} \left(\frac{R_{r,k,n}^{\frac{1}{\gamma_k}} - \frac{1}{k+1}}{1 - R_{r,k,n}^{\frac{1}{\gamma_k}}} \right), 0 \right\}, \tag{3.65}$$

where $R_{r,k,n}^{\frac{1}{\gamma_k}} = \frac{X_{n-k,n}}{X_{n-r+1,n}}$. However in practice, the estimator $\hat{D}_J^{(0)} = \max\{D_J, 0\}$ is used to capture both truncated and the underlying distributions.

The upper endpoint can be obtained as:

$$\hat{T}_{k,n} = \max\left\{ X_{n-k,n} \left(\left(\frac{X_{n-k,n}}{X_{n,n}} \right)^{\frac{1}{\hat{\gamma}_k}} - \left(\frac{\frac{1}{k+1}}{1 - \frac{1}{k+1}} \right) \right)^{-\hat{\gamma}_k}, X_{n,n} \right\}. \tag{3.66}$$

The exceedance or tail probabilities are computed using:

$$\hat{p_j} = \frac{k+1}{n+1} \frac{\left(\left(\frac{q}{X_{n-k,n}} \right)^{-\frac{1}{\gamma_k}} - R_{r,k,n}^{\frac{1}{\gamma_k}} \right)}{1 - R_{r,k,n}^{\frac{1}{\gamma_k}}}. \tag{3.67}$$

Considering the random variable in equation (3.46), quantiles of the truncated distribution can be obtained for either rough or light truncation. Rough truncation relates to cases where the data will begin to show the divergence from the Pareto pattern caused by truncation at a high value after the threshold t has been exceeded. However, under light truncation, from the given threshold t onward, virtually no truncation is evident in the data.

For rough truncation, quantiles are computed using:

$$\hat{Q}(1-p) = X_{n-k,n}\left(\frac{\hat{D}_J + \frac{k+1}{n+1}}{\hat{D}_J + p}\right)^{\hat{\gamma}_k}, \tag{3.68}$$

Also, for light truncation, we obtain quantiles using:

$$\hat{Q}(1-p) = X_{n-k,n}\left(\frac{k+1}{(n+1)p}\right)^{\hat{\gamma}_k}. \tag{3.69}$$

Similarly, we can estimate quantiles for the parent distribution Y using:

$$Q_W(1-p) = X_{n-k,n}\left(\frac{\hat{D}_J + \frac{k+1}{n+1}}{p(1+\hat{D}_J)}\right)^{\hat{\gamma}_k}. \tag{3.70}$$

3.6 Confidence Interval for Parameters

We will create some confidence intervals for the estimates produced for the GPD parameters (γ, σ, μ) since all estimates need a margin of error to achieve asymptotic normality. Prior to the proposition by Beirlant et al. (2004) for drawing confidence intervals, the normal distribution was used as an approximation to the true sampling distribution for estimates obtained through Maximum Likelihood estimations, and credible intervals for Bayesian estimations. Let Υ be any parameter of interest, then the confidence interval was constructed traditionally using

$$\hat{\upsilon} = \pm 1.96 * \sqrt{\frac{\eta_{\hat{\upsilon}}}{m}}$$

Beirlant et al. (2004) suggested using the Profile Likelihood Based Confidence Interval because it is obtained directly from the likelihood function. In addition, it does not assume the normality of the estimator, as it is based on the asymptotic χ^2 distribution of the likelihood function, rather than the standard error as applied in Wald's C.I. The Profile Likelihood Based Confidence Interval is obtained using:

$$\ell_p(\Upsilon_i) = \max_{\Upsilon_i} \ell(\Upsilon_i, \Upsilon_{-i})$$

Remarking Υ as the parameter of interest, the $100(1-\alpha)\%$ confidence interval for Υ can be constructed using the profile-likelihood function as:

$$CI_{\Upsilon} = \left\{ \Upsilon : log\ell_p(\Upsilon) \geq log\ell_p(\hat{\Upsilon}) - \frac{\chi^2_{(1)}(1-\alpha)}{2} \right\} \tag{3.71}$$

3.7 Goodness of Fit Tests

These tests determine if sample data from a particular population fits a given distribution. The following hypotheses would be evaluated.

$$H_0 : \gamma = 0 \text{ vs } H_1 : \gamma \neq 0 \tag{3.72}$$

$$\begin{aligned} &H_0 : \text{The data comes from an Exponential distribution.} \\ &H_1 : \text{The data does not come from an Exponential distribution.} \end{aligned} \tag{3.73}$$

3.7.1 Kolmogorov-Smirnov (K-S) Test

The original Kolmogorov-Smirnov test was developed by Andrey Kolmogorov and Nikolai Smirnov. For two independent samples, it measures the disparity between the empirical distribution functions. However, the test can also determine if a sample is drawn from a hypothesized reference probability distribution.

In testing hypotheses (3.72), which is a classical exponential case, we compare the cumulative distribution of the exponential distribution to the cumulative distribution of the GPd. Hence under H_0, the K-S statistic is obtained using:

$$D_m = \max_{1 \leq i \leq m} \left(\left| 1 - \exp\left(-\frac{Y_{i:m}}{\hat{\sigma_\mu}} \right) - \frac{i}{m} \right|, \left| 1 - \exp\left(-\frac{Y_{i:m}}{\hat{\sigma_\mu}} \right) - \frac{i-1}{m} \right| \right) \tag{3.74}$$

Let σ_μ denote the estimated scale parameter σ_μ obtained from the exponential distribution and H_γ represent the GPD.

3.7.2 Cramer-Von Mises Test

Let $x_1, x_2, \cdots, x_n$ be the observed values, in increasing order. Then the test statistic is:

$$W_m^2 = \sum_{i=1}^{m} \left(H_{\hat{\gamma}} \left(Y_{i:m} | \hat{\sigma}_{\mu, H_\gamma} \right) - \frac{2i-1}{2m} \right)^2 + \frac{1}{12m} \tag{3.75}$$

3.7.3 Anderson-Darling Test

A variation of the Kolmogorov-Smirnov test is the Anderson-Darling test. It is more susceptible to tail variations in distributions. We proceed to test hypotheses (3.73). Let H_γ represent the GPD. The Anderson-Darling Test statistic is computed as:

$$A_m^2 = -m - \frac{1}{m} \sum_{i=1}^{m} \left\{ (2i-1) log \left(H_{\hat{\gamma}} \left(Y_{i:m} | \hat{\sigma}_{\mu, H_\gamma} \right) \right) + (2m+1-2i) log \left(1 - H_{\hat{\gamma}} \left(Y_{i:m} | \hat{\sigma}_{\mu, H_\gamma} \right) \right) \right\} \tag{3.76}$$

Chapter 4

Data Analysis

Results obtained from the Extreme Value Analysis of the data are presented in this chapter. This chapter is divided into three sections. Section 4.1 examines the distribution of the data using descriptive statistics and plots such as the mean excess plot, boxplot, histogram and exponential QQ plot. Section 4.2 provides results under the Block Maxima method. Section 4.4 provides results under the Block Maxima method. Section 4.5 provides a summary of this chapter.

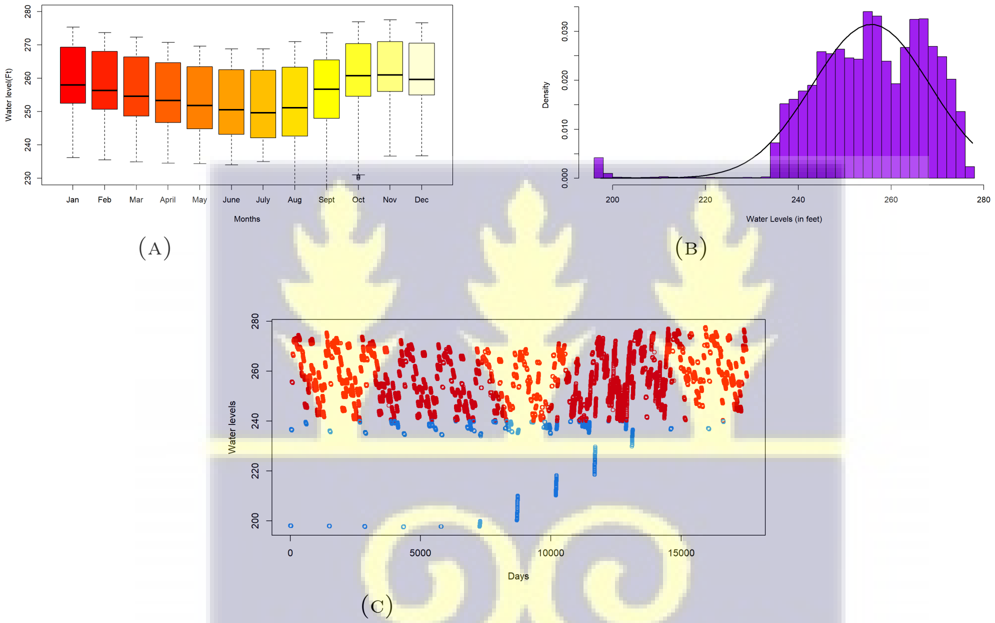
4.1 Descriptive Statistics

The descriptive disposition of the data under review is presented in this section, utilizing a variety of summaries and plots. Information on whether the data is appropriate for extreme value analysis is also provided here. Daily water levels (in feet) of the Akosombo dam between the periods of January 1st, 1965 and December 31st, 2013 would be used for analysis. The statistics are provided in Table (4.1) below.

Table 4.1: Summary statistics of the water level (daily) of the Akosombo dam

Statistics	Value (ft)
Minimum	197.40
1st quartile	247.20
Median	256.20
Mean	255.71
3rd quartile	265.70
Maximum	277.54
Variance	161.2486
Standard Deviation	12.70
Skewness	-1.12
Kurtosis	3.24

The data comprises of 17,533 daily water readings from the Akosombo dam, spanning the periods of January, 1965 and December 2013. The minimum and maximum water levels measured since the dam's construction are 197.40 feet and 277.54 feet respectively, with the maximum water level a few inches below the crest of the dam. Overall, the dam has been very efficient in retaining optimal water levels, averaging an elevation height of 256.2 feet for the period under investigation. Furthermore, the skewness (-1.12) and kurtosis (3.24) statistics indicate the presence of a radical shift from the mean of the distribution. This shift suggests that the data has a longer or fatter tail on the left side of the distribution as observed in Figure (4.1).



(A)

(B)

(C)

Figure 4.1: Summary plots of the data

The top panel shows the spread of the data. The bottom panel describes the appropriateness of the data for extreme value analysis. The top left (A) was used to identify extreme observations. The months of June, July and August recorded more variations in water levels. Besides the months of November and December, all other months had some extreme observations. The top right plot (B) indicates that most data points are above the critical operating level (240.00 ft.). The scatter plot on the bottom (C) depicts independence among the data points, as the points follow no apparent pattern.

To ensure that the data are consistent over periods of time, we check for stationarity using the "Augmented Dickey-Fuller (ADF)" test using the hypothesis test:

- H_0: Data is non-stationary

  H_1: Data is stationary

. Following the hypothesis test in 4.1, the ADF test indicated at a 95% confidence level that our data is stationary and fit for further analysis. See Table 4.2 for the table. Refer to Appendix (A.2.1) for R code.

TABLE 4.2: ADF Stationarity Test for water level

Variable	Test Statistic	P-value
Water level	-11.85	0.01

4.2 The Block Maxima Method

4.2.1 Preliminary Analysis

As a rudimentary step prior to Gumbel's method, it is essential to have an insight into the right tail of the underlying distribution. By completing a preliminary graphical analysis, we may determine which of the three kinds of extreme value distributions outlined in (3.3) is most present in our distribution function F. We must remark that Balkema and De Haan (1974) proved theoretically that, a quantile of the GEV distribution can be roughly expressed as a function of the GPD at a high threshold. Thus, the exponential distribution function can approximate the right tail of a Gumbel family of distributions.

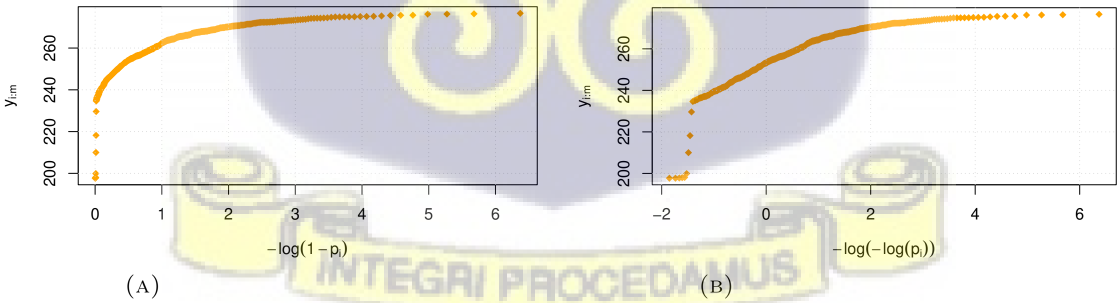


FIGURE 4.2: Exponential QQ plot (A) and Gumbel QQ plot (B) of Gumbel model

In Figure 4.2, the left graph (A) represents the plot of theoretical and empirical quantiles (QQ) of the Exponential model. The right graph (B) presents a plot of the theoretical and empirical quantiles of the Gumbel model. Both plots can be visually inspected to reveal a significant level of concavity, particularly for large values of the underlying distribution. It can clearly be observed in Figure 4.2 that the exponential QQ plot approximates the Gumbel model very well. Thus, the Gumbel model does not provide a good fit for the data. The majority of these conclusions are subjective and descriptive. To support these conclusions, we will require unbiased tests. We will put our data through more rigorous statistical tests in the section after this.

4.2.2 Statistical Choice For GEV models

Our search for an extreme value model has been reduced to two models as a result of the preliminary analysis in the previous section. Recall that in our preliminary report, the Exponential and Gumbel models both exhibited some concavity. Thus, we evaluate the Gumbel model's suitability objectively and also look for additional models that can accurately match the data. It is noteworthy that other models of interest as established in (3.3) are the GEV distributions with $\gamma < 0$ and $\gamma > 0$. However, motivated by our preliminary analysis, we will test,

$$H_0 : \gamma = 0 \text{ against } H_1 : \gamma < 0.$$

4.2.2.1 Gumbel Test

Following the procedure outline in Section 3.3.2.1, and using the R code described in Appendix A.5.1, we get the proceeding results:

```
Test Statistic= 1.48  Critical Value*= -2.28  p-value= 5.741256e-05
```

At $\alpha = 0.05$, the aforementioned findings result in the rejection of the null hypothesis, H_0. The Gumbel model is thus inappropriate for modelling this data. This result, however, validates our observations from Figure 4.2. We shall carry out another test to verify the Gumbel model's validity.

Summary

Therefore, based on our preliminary findings and the Gumbel test, we may conclude that the Gumbel model is unsuitable for modelling this data. Furthermore, we can agree that the Weibull model ($\gamma < 0$) is more suitable for fitting the data.

4.2.3 Parameter Estimations

The established distribution for modelling the water level of the Akosombo dam is the Weibull family of distributions, from which parameter estimations and other statistical deductions can be inferred. We will obtain parameter estimates using both the classical and Bayesian estimation techniques in this section and compare them. Using the R code in Appendices A.7.1 and A.7.2, we obtain the results shown in Table 4.3. In Table 4.3, we present the parameter estimates obtained using both MLE and Bayesian approaches, along with their respective Log-likelihoods, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The parameter estimates signal that there is little variation between the two estimation frameworks employed. They further establish that the data follow the Weibull domain of attraction (i.e. $\gamma < 0$). The difference in scale parameters, though small, suggests that the fit produced with Bayesian methodology has a more dispersed distribution. We determine that the Gumbel model, while good, is less suitable for fitting the data by comparing the log-likelihood, AIC and BIC estimates for the MLE estimates of the GEV model and Gumbel model. Refer to Appendix B.1 for Gumbel estimates. Here, besides the log-likelihood value, models with smaller estimates fit the data better. This validates the conclusions we reached in Section 4.2.2.

Table 4.3: Parameter estimation of the Generalised Extreme Value using MLE and Bayesian methods

	Maximum Likelihood Estimation			Bayesian Estimation		
Estimates	μ	σ	γ	μ	σ	γ
Parameters	254.13	13.65	-0.59	249.73	13.67	-0.59
Confidence Intervals	(252.95, 255.32)	(12.71, 14.58)	(-0.64, -0.55)	(252.79, 255.57)	(12.75, 14.67)	(-0.64, -0.54)
Standard errors	0.60	0.48	0.03			
Loglikelihood		-2247.28				
AIC		4500.56				
BIC		4513.67				

4.2.4 Goodness of Fit Tests

We employ the goodness of fit tests outlined in Section 3.7 to validate our findings in the tests mentioned above. Using the R code displayed in Appendix A.3.1 and the value of σ_μ obtained in Table 4.3, we obtain:

```
Kolmogorov-Smirnov statistic:  0.05  P-value: 0.074.
```

In accordance with the hypothesis test in 3.72, we fail to reject the null hypothesis that the data follow the Gumbel domain of attraction. The Kolmogorov-Smirnov test, however, is known to perform appallingly with sample sizes above 1000. As a result, we conduct the Likelihood Ratio Test.

4.2.4.1 The Likelihood Ratio Test

It is possible to compare the Gumbel model's validity to the GEV model using a different test that was proposed by Hosking (1984). The LR test adheres to the steps discussed in Section 3.3.2.2. With the R code presented in Appendix A.6.3, we obtain the following results:

```
LRT= 465.94  LRT*= 463.72  P-value= 7.45353e-103 .
```

The Likelihood Ratio test also rejects the Gumbel model at a 0.05 significance level.

4.2.5 Estimation of Other Parameters

With the parameter estimates acquired in the preceding subsection, we estimate other quantiles for both estimation methods. It can be shown from Table 4.4 that parameter estimates generated using both estimating techniques produced comparable outcomes. But both models with $\gamma = 0$ generated return levels and periods with high exceedance probability, as observed in Appendix B.2 and B.4, supporting the findings we made in Section 4.2.2. Results from the GEV models with $\gamma \neq 0$ were more credible. The Akosombo dam's water level rises to **275.54** and **275.66** feet at least once every century, with a 1% chance of it occurring. We see lower water levels with increased exceedance probabilities for shorter time periods. Using the MLE and Bayesian frameworks, we estimated the right upper endpoints (i.e., highest estimated level) as **277.01ft** and **277.25ft**, respectively, based on the highest water level that has been recorded in the Akosombo dam during the study period, which was **276.64ft**. The corresponding exceedance probability for the endpoints is extremely low (9.448319e-06 and 1.88541e-10, respectively). Therefore, there is very little chance that the water level in the dam will rise above its maximum operating water level (278.00ft).

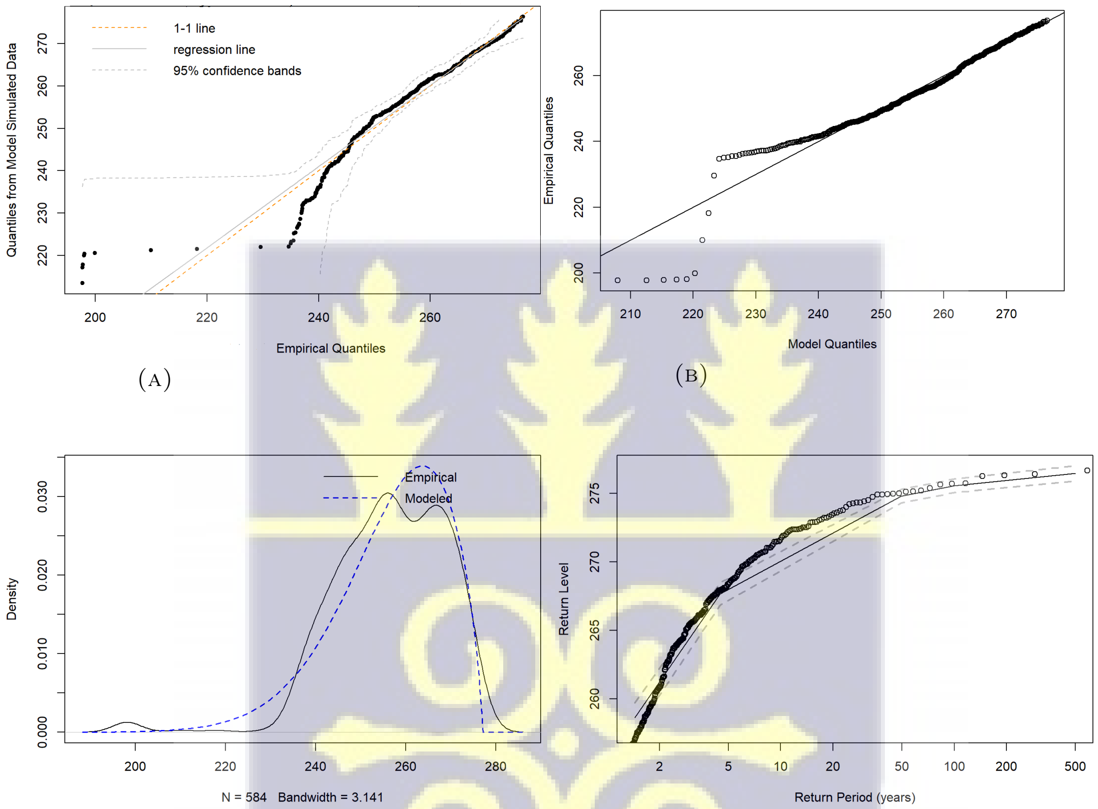
Table 4.4: GEV return level, period, upper endpoint and exceedance probabilities

	Maximum Likelihood		Bayesian	
Return Periods (Years)	Return Level	Exceedance Probability	Return Level	Exceedance Probability
2	258.67	0.50	258.35	0.49
5	267.66	0.20	267.52	0.19
10	271.03	0.10	270.98	0.099
20	273.12	0.05	273.17	0.049
50	274.78	0.02	274.86	0.019
100	275.54	0.01	275.66	0.009
	Value	Exceedance Probability	Value	Exceedance Probability
Maximum water level	277.54	0.00	277.54	0.00
Upper Endpoint (X^F)	277.01	9.448e-06	277.25	1.885e-10

4.2.6 Model Diagnostics

Figures 4.3 and 4.4 illustrate, respectively, how well the GEV model performs when using ML and Bayesian techniques. Both empirical quantities 4.3a and model quantiles 4.3b of the GEV model shown in Figure 4.3 illustrate a poor fit for the data. The density plot also portrays a significant variation between the empirical model (density) and the fitted model. The same is observed in the Figure 4.4 obtained under the Bayesian paradigm.

To test for convergence, Markov Chain Monte Carlo (MCMC) outputs obtained were plotted using trace plots. Figure 4.5 shows trace plots of the scale, shape and location parameter samples that were obtained from the posterior distribution. The plots indicate that sampling was done fairly well from the posterior distribution.



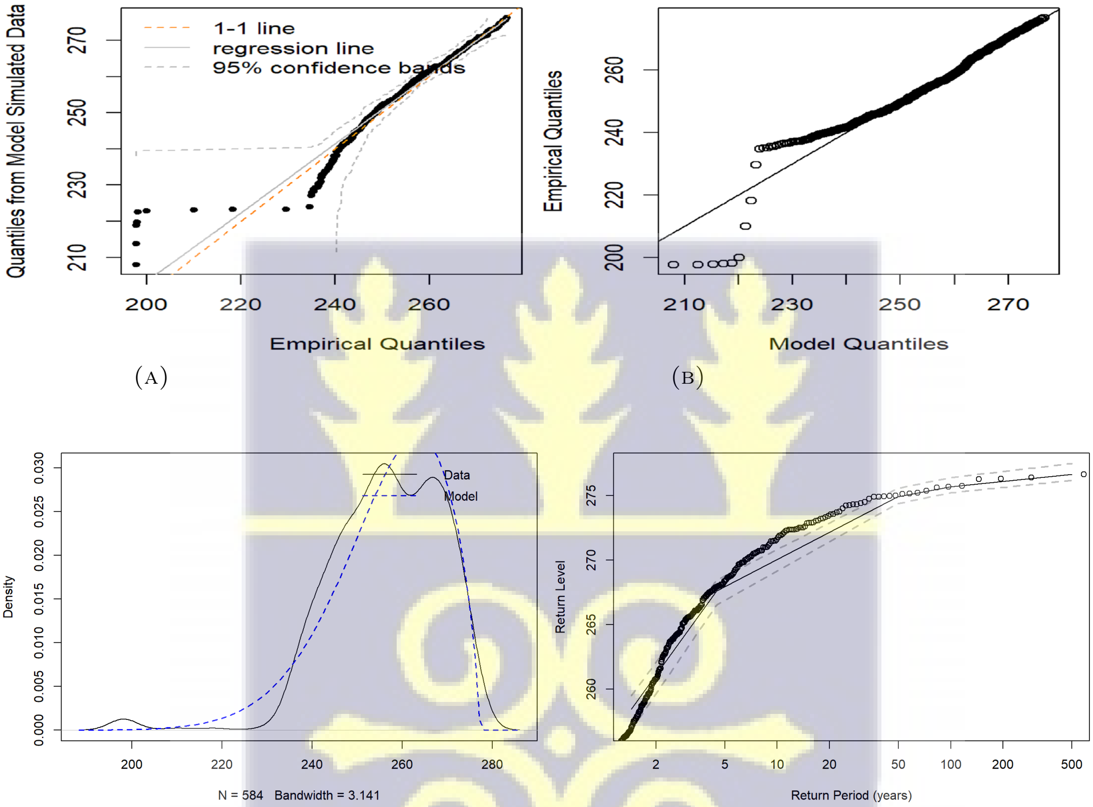
(A)

(B)

(C)

(D)

FIGURE 4.3: Diagnostic plots for GEV model obtained from Bayesian Estimation



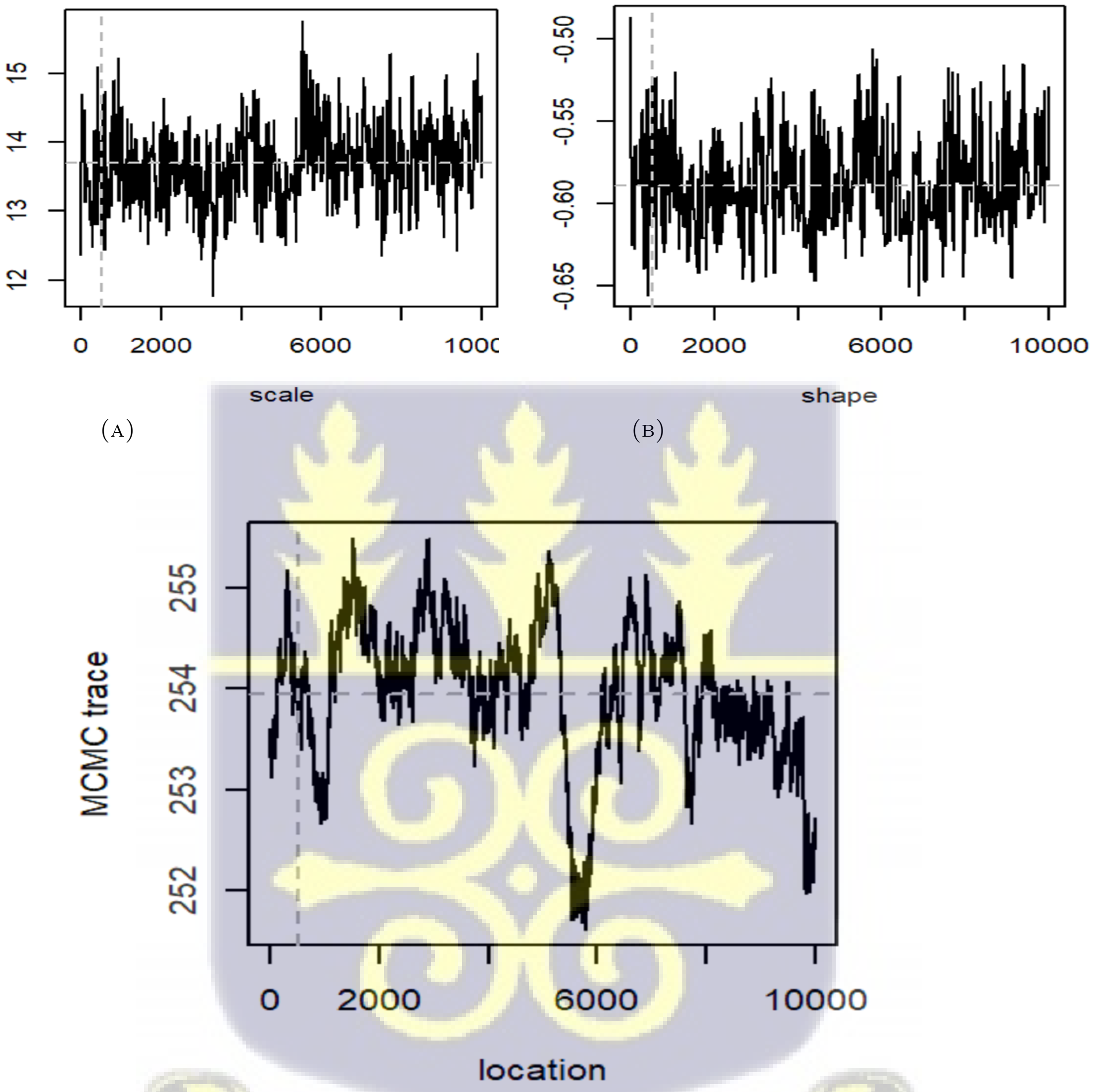
(A)

(B)

(C)

(D)

FIGURE 4.4: Diagnostic plots for GEV model obtained from Bayesian Estimation



(A)

(B)

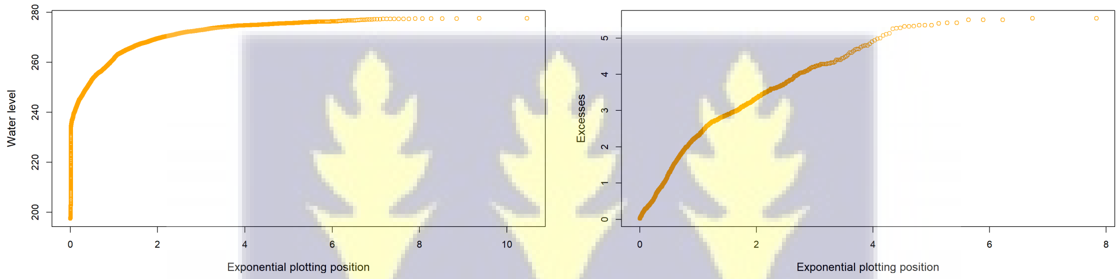
(C)

Figure 4.5: Trace plots for GEV model obtained from Bayesian Estimation

4.3 The POT Method

4.3.1 Preliminary Analysis

Obvious concavity in the exponential Q-Q plots of the data, as well as the excesses in (4.6), provide credible evidence that the data's underlying limit distribution has a lighter right tail (Beirlant et al., 2004). As a result, it suggests that the exponential model is unsuitable for characterizing the excesses of the data. Given that the GPD transforms into an exponential model when $\gamma = 0$, we anticipate that the data will follow a GPD with $\gamma < 0$. Noting that the Exponential QQ plots are illustrative but sensitive to interpretation, we subject the data to more rigorous statistical tests in the following section to justify our decision to choose a GPD model with $\gamma < 0$.



(A) (B)

FIGURE 4.6: Exponential QQ plots of water level and excesses over threshold

4.3.2 Statistical Choice of GPD model

Preliminary analysis for the possible GPD model suggested our data may not be appropriately fitted with the exponential model. Thus, we perform three statistical tests to determine the best domain of attraction for the GPD. As elaborated in the previous chapter, we perform the Marohn, Gomes and Van Monfort, and the likelihood ratio tests. The choice of an appropriate GPD model follows the hypothesis test:

$$H_0 : \gamma = 0 \text{ against } H_1 : \gamma \neq 0.$$

Recalling the random variable $\Upsilon = \Upsilon_1, \Upsilon_2, \cdots, \Upsilon_n$ and the exceedances $\Gamma = \Gamma_1, \Gamma_2, \cdots, \Gamma_m$, we obtain the excesses ϱ over the selected threshold ϑ as $(\Gamma - \vartheta)$. In the context of this work, the random variable Υ represents the water level of the water impounded by the Akosombo dam.

4.3.2.1 Gomes and Van Monfort Test

With suggestions in our preliminary analysis, we conduct the Gomes and Monfort test based on the one-sided hypothesis in equation (3.30). The test based on the Gumbel test statistic rejects the null hypothesis when $Gst* \leq G_\alpha$. See Section 3.4.2.1 for the concise procedure for the test. The results obtained from the test are presented in Table 4.5. Since our p-value (1.153889e-276) is less than 0.05, we reject H_0 and conclude that the Gumbel model is inappropriate for fitting the data.

TABLE 4.5: Gomes and Van Monfort test statistics

Statistics	Value
GSt	0.9968
GSt*	-6.4542
P-value	1.153889e-276

4.3.2.2 Marohn GPD Test

Using the R code presented in Appendix A.6.2 and the steps outlined in Section 3.4.2.2, we arrive at the findings that follow:

```
    Reject Ho

    t_m*= -9.235633 <=  p-value= 1.283816e-20
```

The null hypothesis, H_0, is rejected by the aforementioned findings at a 0.05 α level. Therefore, modelling this data using the Gumbel model is unsuitable.

4.3.2.3 Likelihood Ratio Test (LRT)

Hosking (1984) test can be used to compare the Exponential model's validity against that of the GPD model. The LR test follows the steps outlined in Section 3.4.2.3. The following outcomes are obtained using the R code shown in Appendix A.6.3:

```
[1] l= 237.67  l*= 236.92  p-value= 1.846293e-53
```

The Exponential model is likewise rejected by the Likelihood Ratio test at a 0.05 significance level.

Summary

As a result, we can infer from our preliminary results, the Gomes and Van Monfort Test, Marohn GPD Test, and the LRT that the Exponential model is unsuitable for modelling this data. We can also agree that the $\gamma < 0$ is a better model for fitting the data.

4.3.3 Threshold Selection

This approach assumes that we have a sufficiently high threshold for fitting the appropriate limiting distribution of the data. All excesses over this specified threshold are used to fit a limiting parametric distribution, from which parameters of interest specified in our methodology, high quantiles, return levels and period would be estimated.

As elaborated in Section (3.4.3), the choice of an acceptable threshold is facilitated visually by the parameter stability plots and mean residual excess plots. Figure (4.7) graphs a range of mean excess values against a range of ordered statistics. Figure (4.8) displays the Maximum Likelihood estimates of the shape parameter γ against a range of thresholds. In addition, Figure (4.9) displays the MLE estimates for the scale parameter against a range of thresholds. All three plots evaluated suggest a threshold value of 272 feet. This threshold yields 2138 exceedances, making up a proportion of 0.1219 (12.19%) of the data.

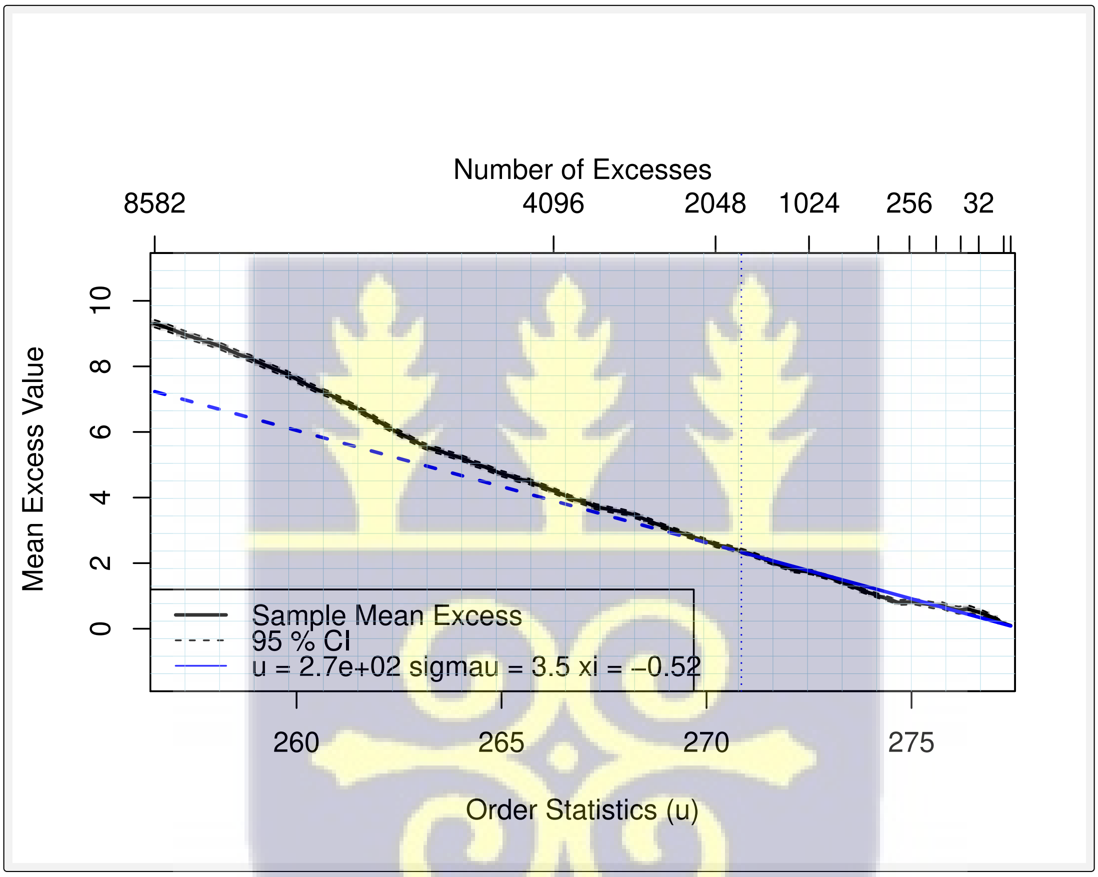


Figure 4.7: Mean Excess Plot of daily water level

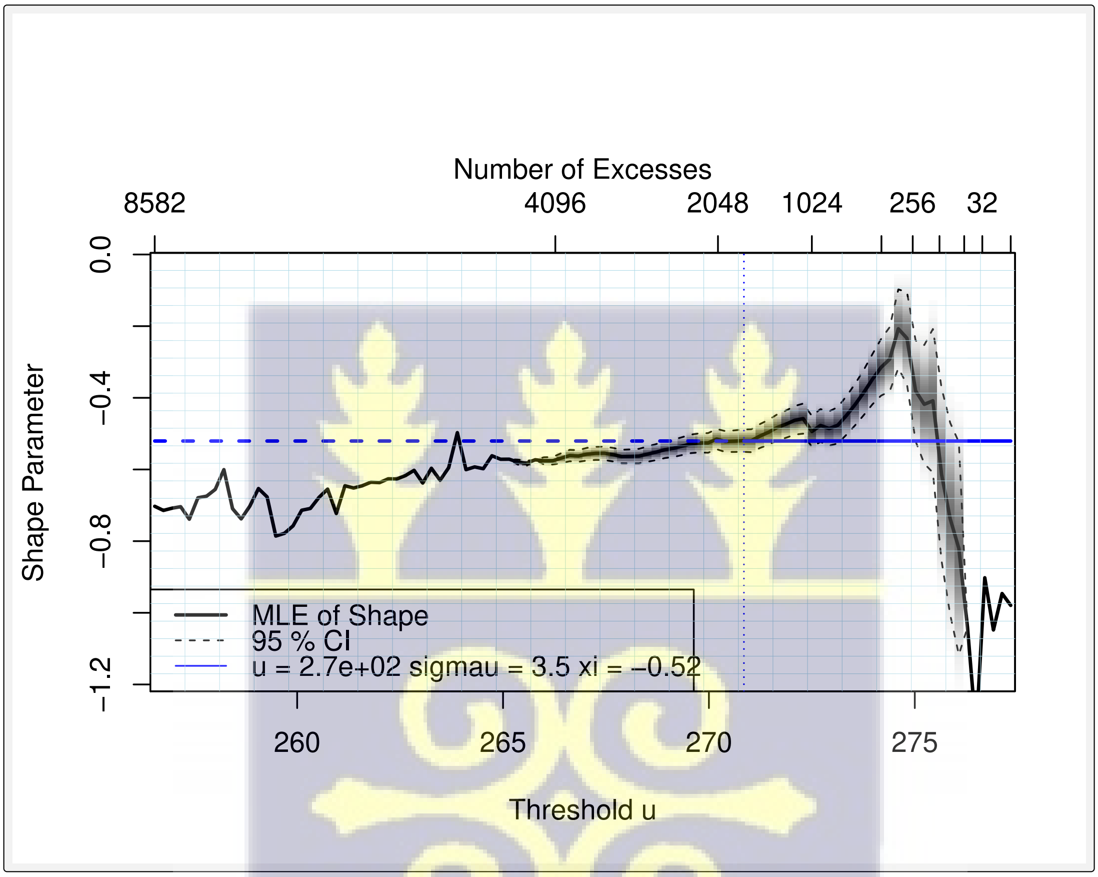


FIGURE 4.8: Shape Parameter Stability Plot: Shape parameter patterns based on threshold and excesses. The dashed lines surrounding the parameter estimations at each k indicate the 95% confidence interval.

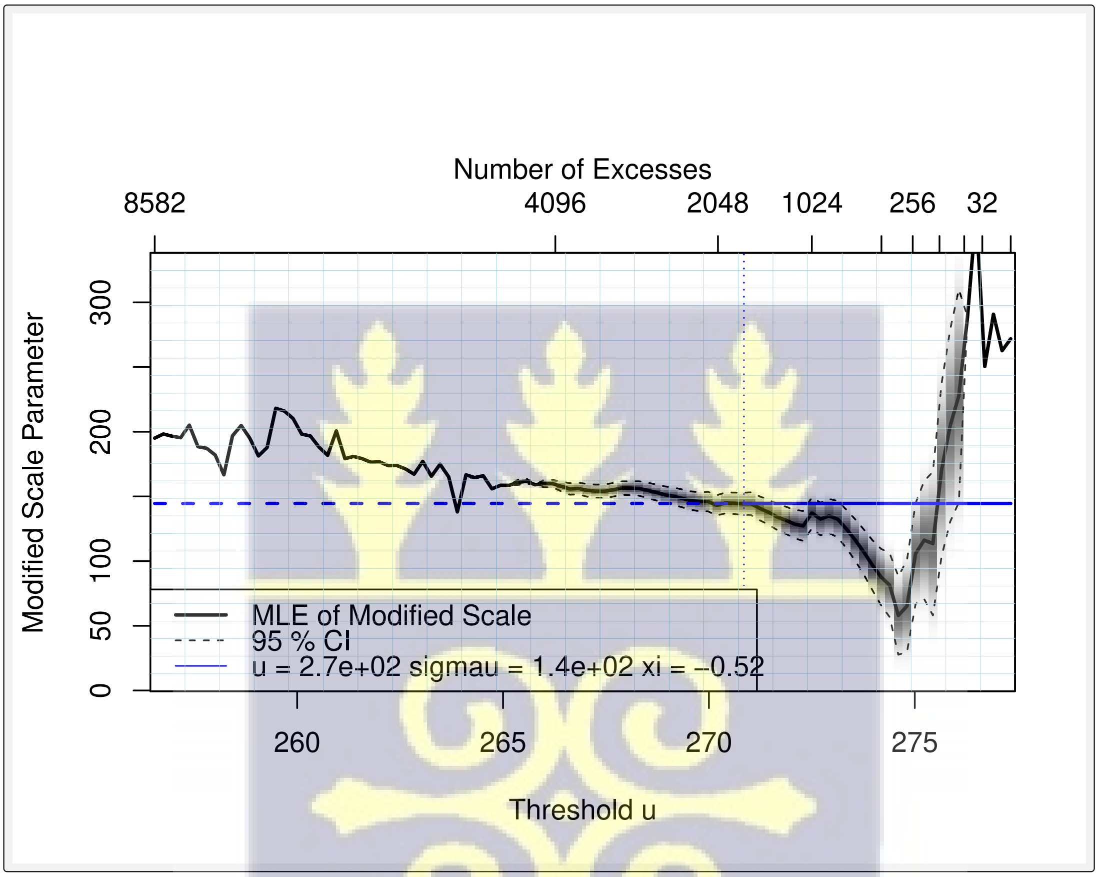


FIGURE 4.9: Scale Parameter Stability Plot: Scale parameter patterns based on threshold and excesses. The dashed lines surrounding the parameter estimations at each k indicate the 95% confidence interval.

4.3.4 Parameter Estimations

We have sufficient justification from the aforementioned section to proceed with Extreme Value analysis. Following the procedure outlined in our third chapter, we continue by fitting models for both the underlying truncated distribution and the non-truncated distribution.

Using the Maximum Likelihood Estimation and the Bayesian Estimation, we estimate the parameters for the appropriate limiting distribution (i.e.the shape and scale parameters). We further construct high quantiles and estimate for each their exceedance probability, return period and return level. Table 4.6 presents the MLE and Bayesian estimates for the GPD respectively. Both estimation methods produce similar results, suggesting that the data follow the Weibull domain of attraction. From Table 4.6, we observed that the GPD ($\gamma \neq 0$), produces better estimates, with smaller standard errors as compared to the GPD ($\gamma = 0$) as shown in Table B.3 . In addition, the Weibull class produces estimates with bigger log-likelihoods, with smaller AIC and BIC. Thus, the Weibull class fits the data better than the Exponential distribution, as observed in Table 4.6.

Table 4.6: Parameter estimation of the generalized Pareto model for the Akosombo dam with MLE and Bayesian methods

	Maximum Likelihood		Bayesian	
Estimates	σ	γ	σ	γ
Estimated parameters	2.72	-0.47	2.	-0.47
Standard errors	0.02	0.09	/	/
Confidence Intervals	(2.55, 2.87)	(-0.51, -0.43)	(2.55, 2.87)	(-0.50, -0.43)
Log-likelihood	-1937.55		/	
AIC	3879.10		/	
BIC	3889.39		/	

4.3.5 Goodness of Fit Tests

We employ the goodness of fit tests outlined in Section 3.7 to validate our findings in the tests mentioned above. Using the R code displayed in Appendix A.4.1 and the value of σ_μ obtained in Table 4.6, we obtain:

```
Kolmogorov-Smirnov statistic:  0.98
```

```
Cramer-Von Mises Statistic:    0.69  P-value: 1.594373e-05

Anderson-Darling statistic:    3.92   P-value:  2.863804e-05.
```

In order to reject the null hypothesis of an exponential distribution, the test statistic obtained must be compared to the critical values listed by Lilliefors (1969). If the test statistic exceeds the corresponding critical point, the null hypothesis is rejected. Table 4.7 presents some generated critical values. With 1286 recorded excesses (m), we obtain our critical values as:

$\frac{0.96}{\sqrt{1268}} = 0.0269, \quad \frac{1.06}{\sqrt{1268}} = 0.0298 \quad \frac{1.25}{\sqrt{1268}} = 0.0351$

For $\alpha = 0.1, \quad \alpha = 0.05$ and $\quad \alpha = 0.01$ respectively.

We reject the null hypothesis that the data follow the exponential distribution for all three alpha levels by comparing the calculated rejection values to the computed Kolmogorov-Smirnov statistic, **0.98**. However, it is universally acknowledged that the Kolmogorov-Smirnov (K-S) test has a low sensitivity to tail-related deviations from the hypothesized distribution. Thus, we support our results with the Cramer-Von Mises and Anderson-Darling tests. At a 0.05 significance level, both the Cramer-Von Mises and Anderson-Darling tests reject the null hypothesis with p-values of **1.594373e-05** and **2.863804e-05** respectively.

Table 4.7: Lilliefors (1969) Simulated critical values of the Kolmogorov-Smirnov statistic adapted to the Exponential distribution with unknown parameters.

		Significance level for D_m		
Statistic	m	0.1	0.05	0.01
D_m	5	0.406	0.442	0.504
	10	0.295	0.325	0.38
	15	0.244	0.269	0.315
	20	0.212	0.234	0.278
	30	0.174	0.192	0.226
	> 30	$0.96/\sqrt(m)$	$1.06/\sqrt(m)$	$1.25/\sqrt(m)$

4.3.6 Estimation of Other Parameters

We estimate other quantiles for both estimation methods using the parameter estimates obtained in the preceding subsection. Table 4.8 illustrates that parameter estimates generated using both estimating techniques produced comparable results. All models produced exceedance probabilities that increased as return levels and periods increased. However, the models $\gamma \neq 0$ produce more insightful return levels,

Table 4.8: GPD return level, period, upper endpoint and exceedance probabilities

	Maximum Likelihood		Bayesian	
Return Periods (Years)	Return Level	Exceedance Probability	Return Level	Exceedance Probability
2	275.50	0.138	275.51	0.137
5	276.29	0.055	276.31	0.055
10	276.70	0.028	276.72	0.027
20	276.99	0.014	277.02	0.014
50	277.26	0.006	277.2883	0.005
100	277.39	0.003	277.43	0.003
	Value	Exceedance Probability	Value	Exceedance Probability
Maximum water level	277.54	0.001	277.54	0.001
Upper Endpoint (X^F)	277.79	0	277.79	0

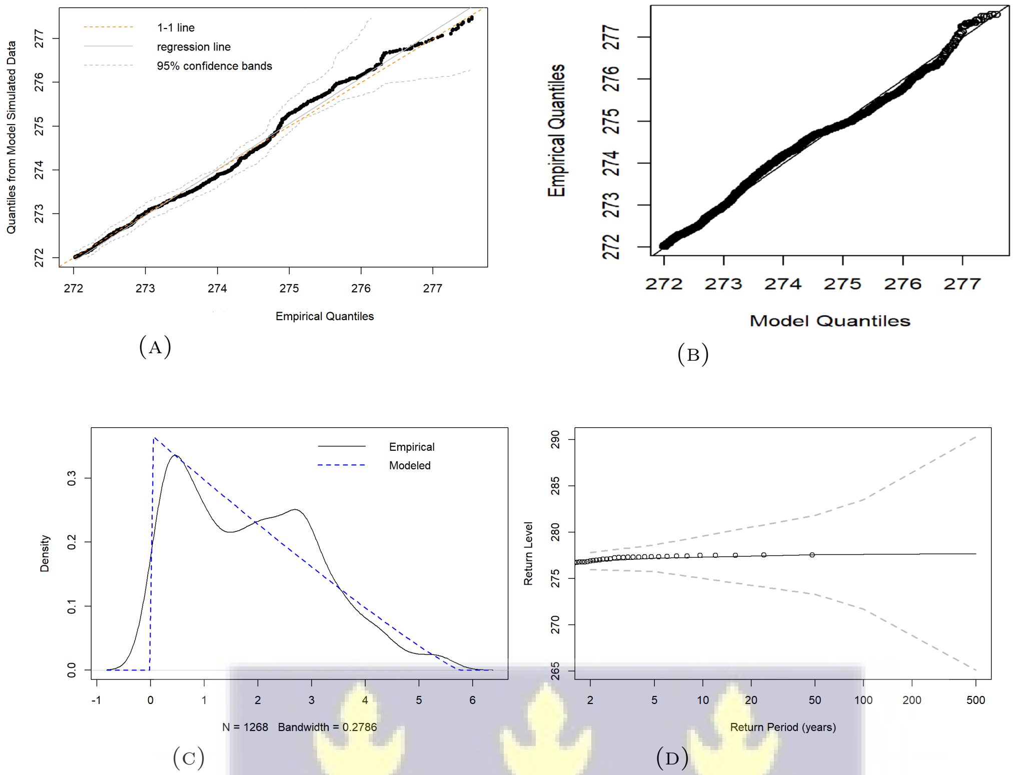
corroborating our findings in Section 4.3.2.3. This increases the credibility of results from GPD models with $\gamma \neq 0$. The water level at the Akosombo dam rises to **277.39** feet and **277.43** feet at least once every century, with a 0.2% chance of occurring. We see lower water levels with increased exceedance probabilities for shorter periods.

We estimated the right upper endpoints (i.e., highest estimated level) using the MLE and Bayesian frameworks to be **277.79ft** and **277.79ft**, respectively. During the study period, the highest water level recorded in the Akosombo dam (i.e.**277.54ft**) resulted in an exceedance probability of **0.01%**. The endpoints have a corresponding exceedance probability of 0.00%. As a result, there is no risk that the dam's water level will rise above its maximum operating water level (278.00ft).

4.3.7 Model Diagnostics

Figures 4.10 and 4.11 illustrate, respectively, how well the GPD model performs when using ML and Bayesian techniques. Both empirical quantities 4.10a and model quantiles 4.10b of the GPD model shown in Figure 4.10 illustrate a good fit for the data. The fitted model and the empirical model (density) exhibit some striking similarities, as shown by the density plot. Similar results are seen in the Figure 4.11 obtained using the Bayesian framework.

Trace plots were used to plot the MCMC outputs in order to check for convergence. Trace plots of the scale and shape parameter samples that were obtained from the



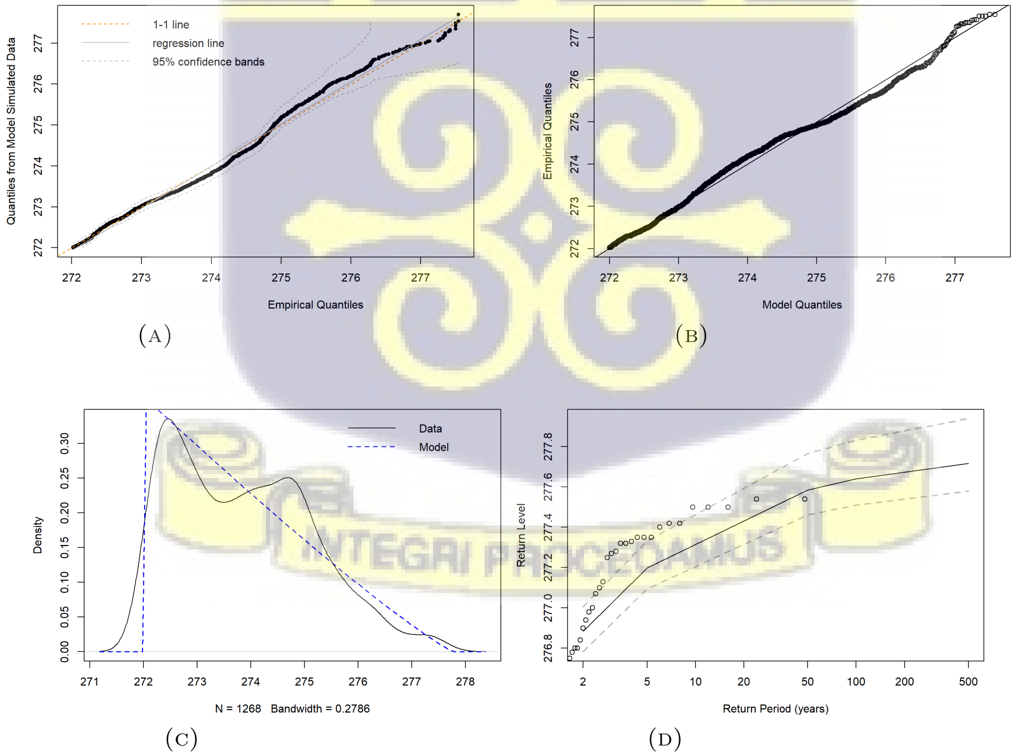
(A)

(B)

(C)

(D)

FIGURE 4.10: Diagnostic plots for GPD Model obtained under the MLE framework



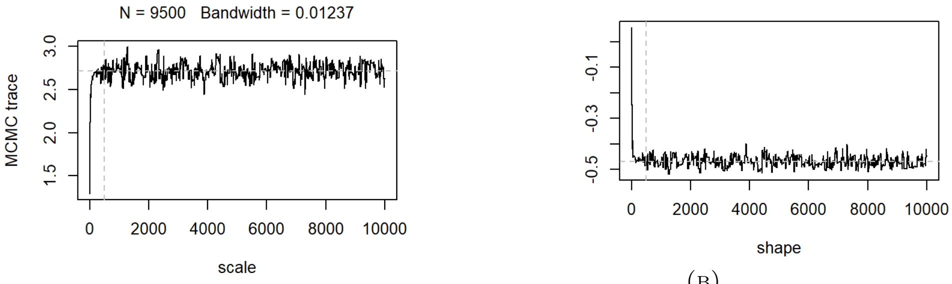
(A)

(B)

(C)

(D)

FIGURE 4.11: Diagnostic plots for GPD Model obtained under the Bayesian framework



(A)

(B)

Figure 4.12: Trace plots for GPD model obtained from Bayesian Estimation

posterior distribution are shown in Figure 4.12. The plots show that sampling from the posterior distribution was done particularly well.

4.4 The Right-Truncated POT Model

4.4.1 Preliminary Analysis

The data belongs to the Weibull domain of attraction according to preliminary analysis and the fitted GEV and GPD models. In light of this, Hill's estimator suggested in Section 3.5.3 is inappropriate for this data. Figure 4.15 illustrates the subpar performance. Although inconclusive, we can observe from Figure 4.13 that the truncated distribution fits the data better than the original non-truncated distribution. As a result, we fit the RT-POT with $\gamma = 0$ and $\gamma \neq 0$ parameters.

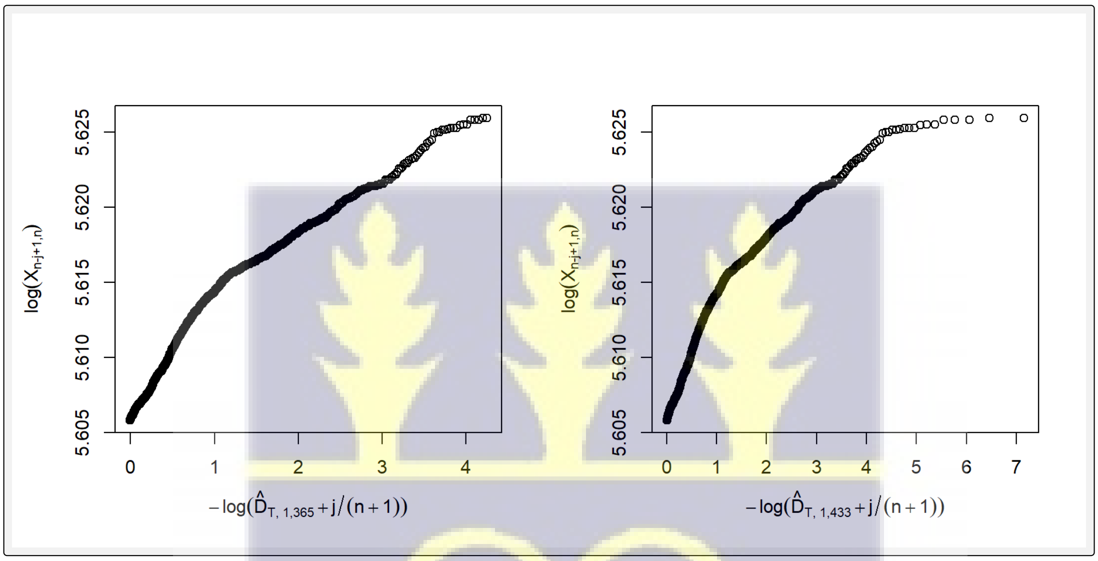


FIGURE 4.13: Pareto plots from Truncated and Original Distributions)

4.4.2 Statistical Choice For Truncated GPD Models

4.4.2.1 Likelihood Ratio Test

As explained in Section 3.5.1.1, we test the validity of the RT-POT model with $\gamma = 0$ against that of the RT-POT model with $\gamma \neq 0$. The LR test follows the steps outlined in Section 3.4.2.3. Using the R code displayed in Appendix A.6.3, the following results were attained:

```
[1] l= 219.3  l*= 218.61  p-value= 1.817952e-49
```

The RT-POT model with $\gamma = 0$ is rejected by the Likelihood Ratio test at a 0.05 significance level.

4.4.2.2 Test between Truncated and Non-truncated Distributions

We reject the null hypothesis that the Akosombo dam data is not truncated at the tails by following the process described in Section 3.5.1.2 and the code provided in Appendix A.10. Figure 4.14 illustrates the decision, and further buttressed with the mean p-values of the test as 0.02749057

Furthermore, Figure 4.14 clearly shows that a majority of k falls below the baseline p-value (0.05). As a result, we can definitively reject H_0 and assert that truncation exists in the data.

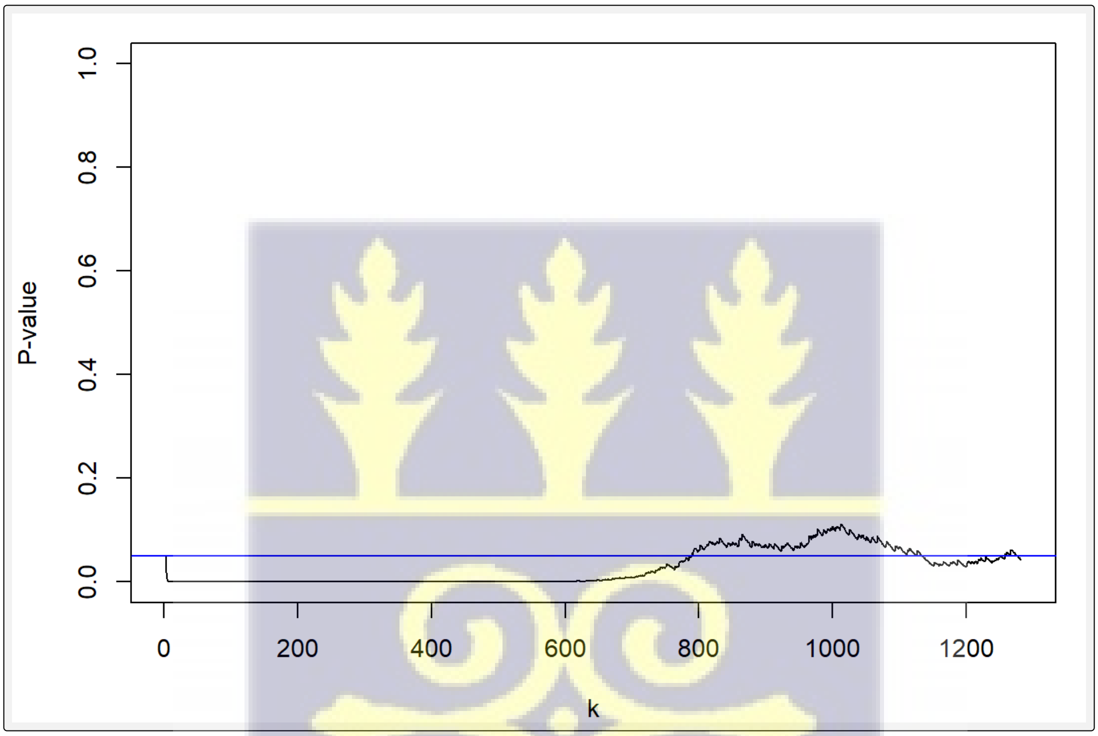


Figure 4.14: Test of Truncation effect

4.4.3 Parameter Estimation

To compare the efficiency of the truncated GPD described in Section 3.5 to the regular GPD, we also model the Akosombo dam data using the truncated GPD. In order to provide a relative position with the non-truncated GPD, we chose the threshold, u, at 272 feet. The parameter estimates for the truncated POT are given in Table 4.9. The shape parameter estimate ($\gamma = -0.47$) of the Right POT indicates that the Weibull domain of attraction is most significant for the Akosombo dam data. The upper endpoints are estimated to be **277.60** and **285.34** for the RT-POT($\gamma \neq 0$) and

RT-POT($\gamma = 0$) respectively. The estimated odds of truncation, D_T, were the same for both models**(0.002)**, suggesting that the likelihood that the truncated probability mass will remain within the scope of the original underlying distribution is very low. Comparing the log-likelihood, AIC and BIC estimates, it can be clearly observed that the Right Truncated POT with $\gamma \neq 0$ fits the data better than the model with $\gamma = 0$. After computing the tail index γ using the ML and Hill estimators, we can effortlessly use them to determine higher quantiles.

TABLE 4.9: Maximum Likelihood parameter estimation of the Right-Truncated generalized Pareto and Exponential models for the Akosombo dam

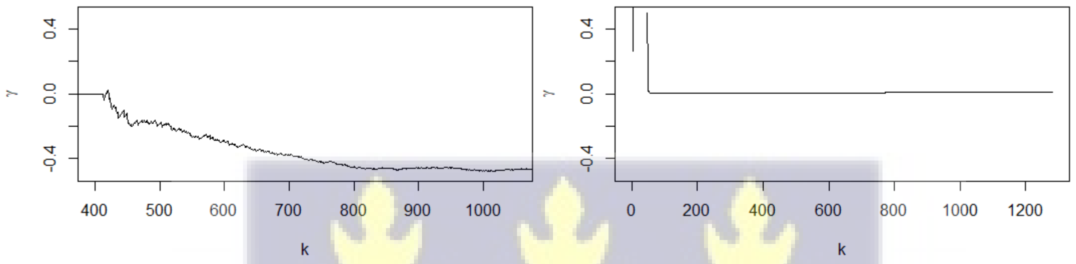
	Truncated Generalised Pareto Model ($\gamma \neq 0$)			Truncated Exponential Model ($\gamma = 0$)	
Estimates	σ	γ	τ	σ	γ
Estimated parameters	2.65	-0.47	-0.168	1.86	/
Standard errors	/	0.03	0.006	0.05	/
Confidence Intervals	/	(-0.49, -0.46)	(-0.15, -0.16)	(1.76, 1.97)	/
Log-likelihood	-1945.94			-2055.59	
AIC	3895.88			4113.18	
BIC	3906.19			4125.47	
T_k	277.60			285.34	
D_T	0.002			0.002	

4.4.4 Estimation of Other Parameters

We estimate additional quantiles for both estimation techniques using the parameter estimates obtained in the preceding Subsection. Table 4.10 demonstrates that the model produced by the RT-POT with $\gamma \neq 0$ had better parameter estimates. This model's log-likelihood was higher **(-1945.94)**. The model with $\gamma = 0$ is a better fit, as shown by the AIC and BIC, which quantitatively measure how much information is lost in a model. Consequently, we will continue with the RT-POT with $\gamma \neq 0$. Table 4.4 shows that there is a 0.01% chance of the water level at the Akosombo dam rising to **277.13** feet at least once every century. Lower water levels are observed along with shorter exceedance probabilities. Based on the highest water level ever recorded in the Akosombo dam during the study period, which was **277.54ft**, we estimated the right upper endpoints (i.e., highest estimated level) as **277.60ft** using the code in Appendix A.9.2. The upper endpoint's corresponding exceedance probability is very unlikely **(i.e. 0.0008)**. In light of this, there is a very low likelihood that the dam's water level will rise above its maximum operating water level (278ft).

TABLE 4.10: Return period, Return level, exceedance probability

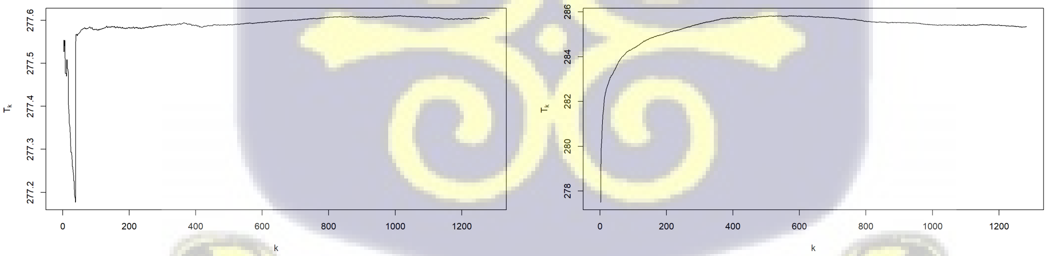
Return Period (Years)	Truncated POT ($\gamma \neq 0$)	
	Return level	Exceedance Probability
2	273.57	0.50
5	275.04	0.199
10	275.80	0.099
20	276.36	0.049
50	276.87	0.019
100	277.13	0.009
	Value	Exceedance Probability
Maximum water level	277.54	0.0008
Upper Endpoint (X^F)	277.60	4.465e-05



(A) (B)

FIGURE 4.15: Shape parameter estimates γ for the Truncated distribution using MLE and the Hill estimator

Model fitting the Akosombo dam data using the RT-POT with $\gamma \neq 0$, we observe that the shape parameter, γ, remains unsteady until it stabilizes below -0.4 for increasing values of k.



(A) (B)

FIGURE 4.16: Estimates of T_k evaluated for each k. Figure (A) represents $\gamma \neq 0$. Figure (B) represents $\gamma = 0$.

From Figure 4.16, we see that both models estimate the endpoints well. However, Figure 4.16a produces an upper endpoint closer to the maximum water level **(i.e. 277.54 ft)** recorded in the Akosombo dam till date. The model with $\gamma = 0$ yields an upper endpoint which greatly exceeds the maximum water level recorded in the

Akosombo dam.

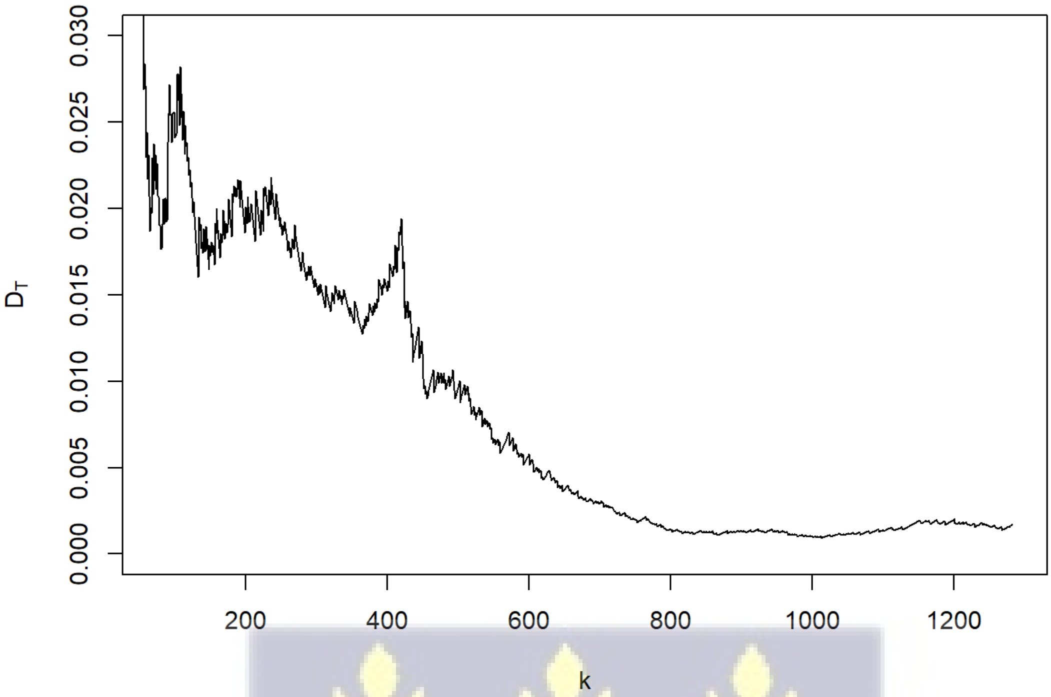


Figure 4.17: Odds of Truncation for $\gamma \neq 0$

The odds of truncation,D_T, for increasing values of k follows an unstable pattern for lower values of k, as observed in Figure 4.17. D_T stabilizes at odds very closes to zero, indicating that the the odds of the truncated probability mass under the untruncated distribution Y is almost negligible. Hence, the probability that the truncated probability mass will remain within the scope of the original underlying distribution is very small.

Using the maximum water level as a high quantile, we observe from Figure 4.18 that for higher values of k, the exceedance probabilities remains stable with negligible probabilities. Simply, this signals that there is a very low probability that the maximum water level would be surpassed. Figure 4.19 shows estimates of large quantiles for small exceedance probabilities such as 0.01. For higher values of k, the truncated distribution remains stable between the heights of 277.0 and 277.2 feets. As observed in Figure 4.20, the reconstructed parent distribution follows a similar pattern as the truncated distribution, but heavy on the first 450 values of k. The pattern stabilizes between the water levels of 277.0 and 277.5 feets.

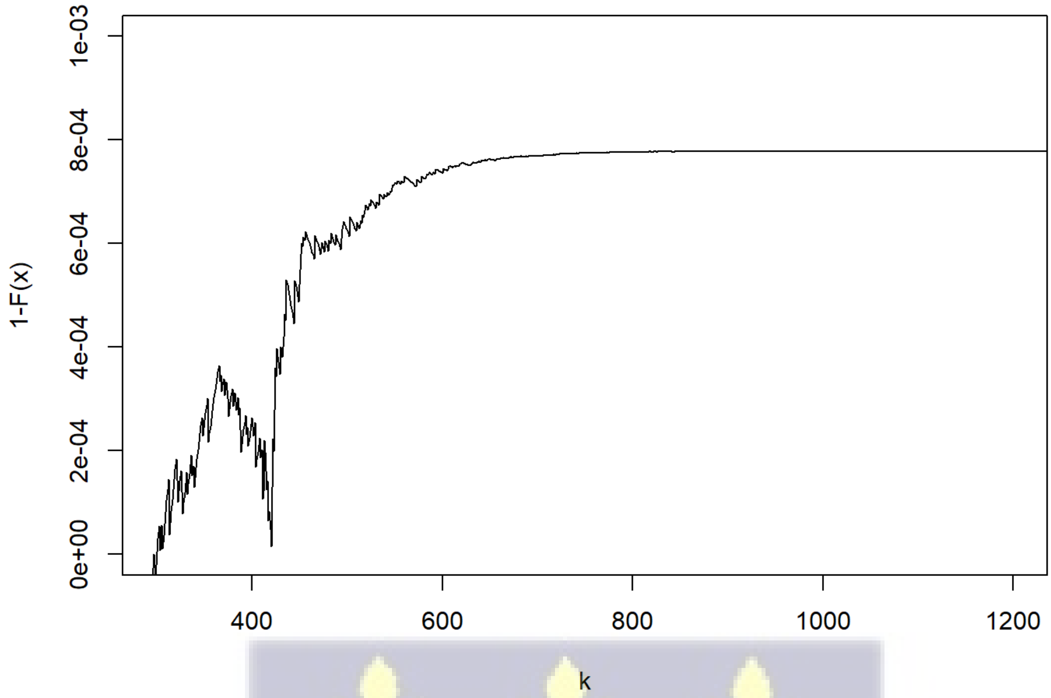


FIGURE 4.18: Estimates of small exceedance probabilities of high quantiles (277.54 ft), for $\gamma \neq 0$

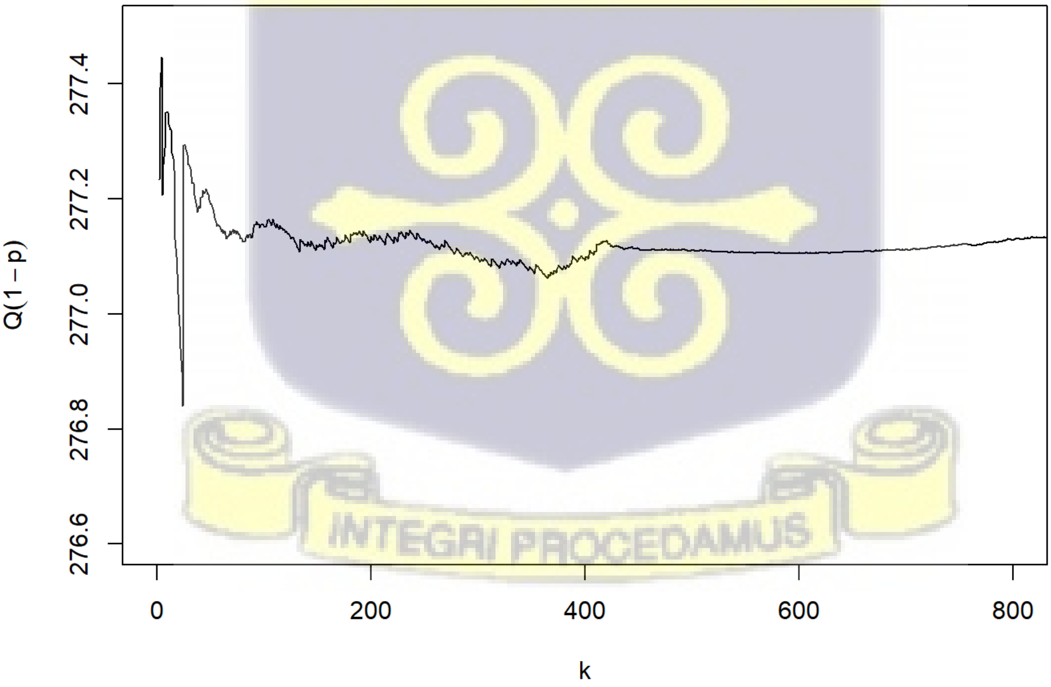


FIGURE 4.19: Estimates of large quantiles of the truncated distribution associated with small exceedance probabilities (0.01%), for $\gamma \neq 0$

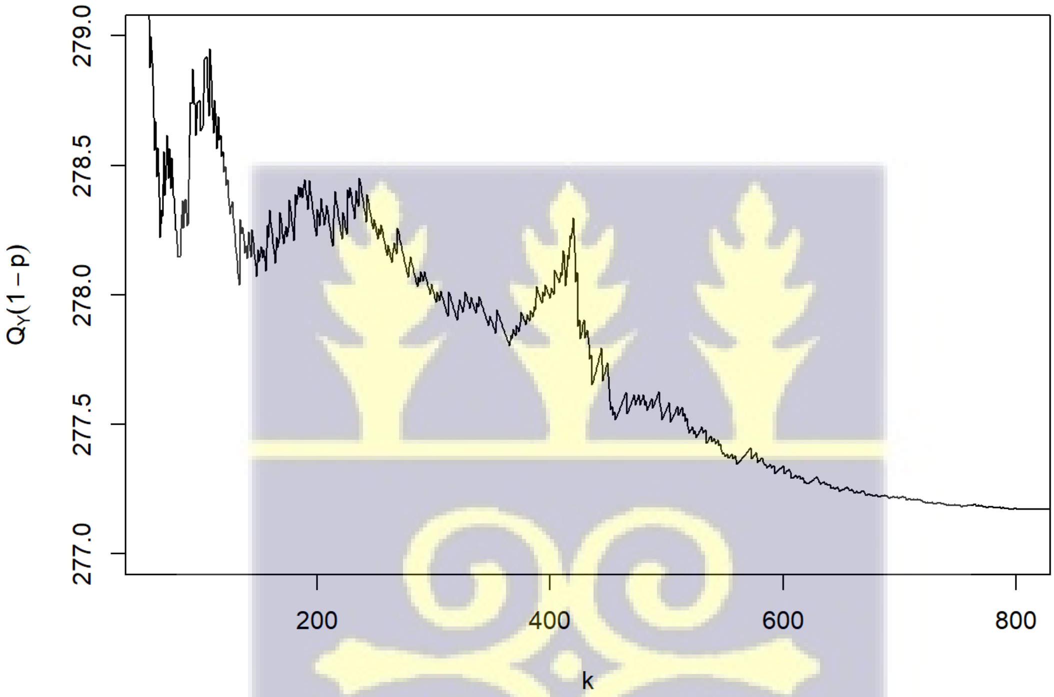


FIGURE 4.20: Estimates of large quantiles of the parent distribution associated with small exceedance probabilities (0.01%), for $\gamma \neq 0$

4.5 Summary of Chapter

This chapter obtained parameter estimates of the shape, scale and location parameters of the GEV and GPD under both the Bayesian and classical frameworks. Parameter estimates of the shape, scale and location parameters of RT-POT were also obtained using MLE. Extreme quantiles under each distribution were estimated using the two estimation approaches. We also access the accuracy of the estimates using standard errors and confidence intervals. Trace plots were also used to access the convergence of the MCMC samples. The study found that all the studied distributions yielded negative estimates of the shape parameter. This observation led to the conclusion that the Hill estimation method was unsuitable for further application in this study as it is valid for positive shape parameters.

Chapter 5

Conclusion and Recommendations

This chapter focuses on the summary of our results and conclusions based on the study objectives. It ends with appropriate recommendations based on the findings and conclusions. In this paper, we attempt to fit three distinct distributions to model the extreme elevations of the Akosombo dam, based on available data. We employ Extreme Value Theory to fit the GEV, GPD and RT-POT distributions to model the Akosombo dam data. Data used for the study comprised of water levels of the Akosombo dam, between the periods January 1965 and December 2013. The Maximum Likelihood, Bayesian and Hill estimators were used.

5.1 Summary of Findings

This study constituted three objectives. The primary objective was to provide a model that would describe the extreme elevation of the Akosombo dam, taking into account truncation. Before embarking on this objective, we tested the data for the possibility of truncation at high elevations. The test showed that the data was truncated at high levels, giving us enough evidence to proceed with this objective. Under this objective, two conditions were evaluated, based on the shape parameter. Based on preliminary findings, we concluded that the data followed the RT-POT distribution with $\gamma \neq 0$, specifically $\gamma < 0$. Thus the Hill estimator was excluded as it is appropriate for modelling data with a non-negative shape index. To contrast the models in terms of performance, we fit the available data to the RT-POT model with $\gamma = 0$ and $\gamma < 0$. A threshold of **272 feet** was selected based on mean excess and parameter stability plots. The model with $\gamma < 0$ was selected as it had a bigger Log-likelihood and smaller AIC and BIC estimates. The shape parameter was estimated as **-0.47** with standard error **0.03** and confidence intervals **(-0.49, -0.46)**. The odds of truncation, which evaluates the likelihood that the truncated probability mass will remain within the scope of the original underlying distribution was estimated as **0.002**.

Our second goal compares the fitted Right-Truncated Peaks Over Threshold to the current GEV and GPD distributions. We performed various statistical tests using the Gumbel test on GEV models before fitting the GEV distribution to the data. The Gumbel test demonstrated that the Gumbel model $\gamma = 0$ is unsuitable for modelling this data at a 5% significance level. According to Table 4.3, the GEV model with the Weibull domain of attraction is the best suitable for fitting the data based on the log-likelihoods, AIC, and BIC of the two GEV models fitted. Both the Bayesian and ML estimators produced estimates that were approximately similar.

When evaluating the chosen model's goodness of fit, the Kolmogorov-Smirnov test stated in Section 3.7 revealed that, at a 95% level of confidence, the Weibull model is the most appropriate model for fitting the data. The Weibull model was found to be appropriate with a 95% level of confidence by the Likelihood Ratio test.

We carried out some statistical tests for GPD models using the Gumbel test before fitting the GPD distribution to the data. The Gumbel test demonstrated that the Gumbel model ($\gamma = 0$) is unsuitable for modelling this data at a 5% significance level. Thus, the GPD model with the Weibull domain of attraction is the best suitable for fitting the data, as shown in Table 4.6 and Table B.3, according to the log-likelihoods, AIC, and BIC of the two GPD models that were fitted. Estimates produced by the Bayesian and ML estimators were remarkably similar.

Diagnostic plots for all estimators indicated that the data were well-fitted. The GEV model obtained was examined using the Kolmogorov-Smirnov test. A GEV model with a negative EVI ($\gamma < 0$) was chosen as a result of the test. A GPD model with a negative EVI ($\gamma < 0$) was likewise chosen by the Marohn GPD, LRT, and the Gomes and Van Monfort diagnostic tests. The Weibull model is the most effective model for fitting the data when assessing the goodness of fit of the chosen model, according to the Kolmogorov-Smirnov test provided in Section 4.3.5.

To choose between the $\gamma = 0$ and $\gamma \neq 0$ models for the RT-POT, we similarly performed the Likelihood Ratio Test. Our results indicate that the $\gamma \neq 0$ model provided a better fit to the data. The Akosombo dam data is truncated at the tails, according to the test comparing truncated and non-truncated distributions, which also demonstrated this at a 95% significance level. In comparison to the $\gamma = 0$ model, the $\gamma \neq 0$ model provides a better fit to the data, according to the parameter estimates shown in Table 4.9.

The RT-POT model was demonstrated to fit the data better since it had better estimates when the AIC, BIC, and log-likelihoods of the GEV, GPD, and RT-POT models were compared.

Estimating high quantiles (truncated and underlying non-truncated), exceedance probability, and return periods are our third objective. The following is a presentation of estimates obtained using the ML and Bayesian approaches. With the GEVd, the Akosombo dam's water level rises to 275.54 and 275.66 feet at least once every century, with a 1% chance of happening. The highest estimated level for the right upper endpoints (i.e., highest measured level) was determined to be 277.01 feet and 277.25 feet, respectively, based on the highest water level ever recorded in the Akosombo dam during the study period, which was 277.54 feet. The probability of exceeding the endpoints is very low $(9.45 * 10^{-6} and 1.89 * 10^{-10})$. There is therefore very little chance that the dam's water level will rise over its highest operational level (278 ft) using either of the estimation techniques.

For the GPD, the Akosombo dam's water level rises to 275.54 and 275.77 feet at least once every century, with a 1% chance of happening, according to the ML and Bayesian estimation methods. The right upper endpoints were determined to be 277.01 feet and 277.25 feet, respectively, based on the highest water level ever recorded in the Akosombo dam during the study period, which was 277.54 feet. The resulting exceedance probability for the endpoints is very low $(9.45 * 10^{-6} and 1.89 * 10^{-10})$ respectively. As a result, there is very little chance that the dam's water level will rise over its highest operational level (278ft).

The Likelihood Ratio Test revealed that the RT-POT with $\gamma \neq 0$ provided a better fit to the data. We also reject the null hypothesis that the Akosombo dam data is not truncated at the tails, as explained in Section 4.4.2. We chose a threshold of 272 feet and estimated the upper endpoint as **277.60 feet**, which is a few inches lower than the measured height of the crest of the Akosombo dam **(278.00 feet)**. Given that the right endpoint estimate at the chosen threshold value is lower than the maximum operating level of the dam, it is possible to calculate exceedance probabilities and return periods near the maximum operating level. A negligible exceedance probability of ($4.47 * 10^{-5}$) is associated with the upper endpoint. The water level at the Akosombo dam has a 0.01 percent chance of rising to 277.13 feet at least once every century, according to the exceedance probability. Ocran et al. (2017) and Minkah (2016) obtained the right endpoints greater than the maximum operating water level (**279.07 feet** and **280.180 feet** respectively). We also obtained an upper endpoint of **277.60 feet**, a few inches lower than the maximum operating water level. This difference is attributed to the marginal difference in the shape parameters obtained. Ocran et al. (2017) and Minkah (2016) obtained ($\gamma = -0.47$ and $\gamma = 0.30$) respectively. The shape parameter, γ, obtained in this study is **-0.47**. However, our results are consistent in terms of exceedance probabilities.

Minkah (2016) showed that an increase in the maximum water level of the dam will reduce the return period of floods, specifically the 100-year return period. Similarly, in this current study, it is shown that even though the probability of exceeding the maximum operating water level is negligible, extending the dam's height may decrease the return period of floods.

5.2 Conclusion

We are able to reiterate a few key aspects of the findings in the previous chapter. Analytically and experimentally, we have demonstrated that when applied to samples of appropriate sizes, the findings produced by the Bayesian and classical frameworks are comparable. The numerical approaches used by the methodologies are the only thing that stands out as different. We should point out that, despite its convenience, applying prior distributions enhances parameter estimates and quantifies hypotheses, as evidenced by findings compared to the previous chapter.

For all three distributions under study, the shape and scale parameters were estimated using the two methods, and the results were remarkably similar. Smith (1985) demonstrated that ML estimators are regular and satisfy common asymptotic criteria for $\gamma > -\frac{1}{2}$. The EVI estimate obtained for all three distributions was marginally greater than -0.5. Our estimators are therefore regular and exhibit common asymptotic features.

The upper endpoints derived from all of the studied distributions were considerably below the 278-foot maximum operational level of the dam. Therefore, under the prevailing conditions, it is highly improbable that the dam's water level will overflow.

We can reiterate some conclusions using the highest water level that was recorded in the dam at the time of the study. Exceedance probabilities for some elevation levels were calculated using the estimated upper endpoint. These probabilities provided information on the likelihood that the water level will rise above the chosen elevation levels for dam management. According to the analysis, there is very little chance that the water level will rise over the maximum value that has been recorded (277.54).

We must note that while our findings indicated that the Akosombo dam data is truncated at high levels, Minkah (2016) and Ocran et al. (2017), who fitted the Akosombo dam data to the GPD and GEV distributions respectively, also achieved similar findings. Due to the catastrophic nature of extreme events, estimates of parameters with smaller confidence intervals are highly preferred. Thus, based on the smaller confidence intervals of the Truncated-POT, we can conclusively say that the Right-POT model is better for predicting the Akosombo dam data.

5.3 Recommendations & Limitations

- There is a small chance that the water would exceed the dam's crucial maximum water level. Engineers of the dam may consider increasing the height of the dam to avoid any severe unanticipated events.

- The study is limited by the use of incomplete historical data records, only considering water levels between 1965 and 2013. To make the results more reliable, data with a larger period and a more complete water level catalogue are needed.

5.4 Areas for Future Studies

- The study focused on truncation at very high-water levels that could result in flooding and a breach of the dam. Therefore, we focussed on the right tail of the distribution of water levels. In the future, we can look at the left tail which could address the closure of the dam due to inadequate water levels.

- A goal of this work was to apply Bayesian estimation methods for the RT-POT distribution. However, due to time constraints, attempts did not fully materialize.

- The present study assumes stationarity concerning the impact of climatic conditions on dam water levels. Thus, incorporating covariates like rainfall, temperature, inflow volume, and discharge volume could enhance estimation and statistical inference. However, further research is required to assess the advantages of including these covariates in the findings of the current study.

Appendix A

R Software codes

Project Data

GEV data

```
library(readxl)

Elevations_Max<- read_excel("Akosombo Daily Upstream Elevations Max.xlsx",

sheet = "GEV")

GEVData<-data.frame(unlist(Elevations_Max,use.names = FALSE))

GEVData<-sort(GEVData[!is.na(GEVData)])
```

GPD data

```
Projdata<-read.csv(file="CombinedData.csv", header = TRUE)

ProjectData<-data.frame(unlist(Projdata,use.names = FALSE))

ProjectData<-ProjectData[!is.na(ProjectData)]

waterlvl<-sort(ProjectData)
```

A.1 Plots

A.1.1 Summary Plots

```
### Boxplot of data

par(mfrow=c(2,2))

Box<-function(x){

  par(mar=c(4,4,1,1))
  boxplot(x$JAN,x$FEB,x$MAR,x$APRIL,x$MAY,x$JUNE,x$JULY,x$AUG,x$SEPT,x$OCT,
          x$NOV,x$DEC, xlab = "Months", ylab = "Water level(Ft)",
          ylim = c(230,280), col=heat.colors(12),
          names = c("Jan","Feb","Mar","April","May","June","July",
                    "Aug","Sept","Oct","Nov","Dec"))}
Box(Projdata)


### Histogram of data
histdat<-function(x){
  par(mar=c(4,4,1,1))
  hist(x,breaks=50,main="", xlab="Water Levels (in feet)",col="purple",freq=FALSE)
  xfit<-seq(min(x),max(x),length=50);  dfunc<-dnorm(xfit,mean=mean(x),sd=sd(x))
  lines(xfit, dfunc, col="black", lwd=2)}
histdat(waterlvl)


### Scatter plot
scatdat<-function(x){
  par(mar=c(4,4,1,1))
  plot(x,xlab = "Days", ylab="Water levels", col=ifelse(x>240,"red","dodgerblue"))}
scatdat(ProjectData)


### Exponential Q-Q plot
par(mar=c(4,4,1,1))
qqplot(qexp(ppoints(length(waterlvl))), line=TRUE,
       col="dodgerblue", xlab= "", ylab = "", main="",waterlvl)
title(xlab= "Exponential plotting position", line = 3,
      cex.lab=1.2, family="sans")
title(ylab= "Water level", line = 3,
      cex.lab=1.2, family="sans")
```

A.1.2 Gumbel QQ Plot and Exponential QQ Plot for GEVd

```
QQ<-function(data, method = c("Gumbel", "Exponential"))
```

```
{  data <- data
  mi<-length(data)
  ii<-c(1:mi)
  l<-ii/(mi+1)
  if (method == "Exponential"){
    QGEV<-sapply(l, FUN = function(x) -log(1-x))
    plot(QGEV,data,pch=18, col= "orange", xlab=expression(-log(1-p[i])),
    ylab=expression(y[i:m]))
    grid() }
  else {
    QGEV<-sapply(l, FUN = function(x) -log(-log(x)))
    plot(QGEV,data,pch=18, col= "orange", xlab=expression(-log(-log(p[i]))),
    ylab=expression(y[i:m]))
    grid()}}
par(mar=c(4,4,1,1))
par(mfrow=c(1,2))
QQ(GEVData, method = "Exponential")
QQ(GEVData, method = "Gumbel")
```

A.1.3 Sample Mean Excess Plot

```
library(evir)
par(mar=c(4,4,1,1))
Meplot<-meplot(ProjectData, main = "", type= "l")
grid(25,25,lwd=0.0005,col = "lightblue")
```

A.1.4 Shape Stability Plot

```
library(evmix)
par(mar=c(4,4,1,1))
shapeplot<-tshapeplot(ProjectData, tlim = NULL,legend.loc = "bottomleft",
           main = "", xlab = "Threshold u", ylab = "Shape Parameter")
grid(25,25,lwd=0.0005,col = "lightblue")
```

A.1.5 Scape Stability Plot

```
par(mar=c(4,4,1,1))
tscaleplot<-tscaleplot(ProjectData, tlim = NULL,
           legend.loc = "bottomleft", main = "", xlab = "Threshold u",
           ylab = "Modified Scale Parameter",alpha = 0.05)
grid(25,25,lwd=0.0005,col = "lightblue")
```

A.2 Augmented Dickey-Fuller (ADF) Test

A.2.1 ADF Test for GEV data

```
library(tseries)
adf.test(sort(GEVData))
```

A.2.2 ADF Test for GP data

```
library(tseries)
adf.test(waterlvl)

#####################################################
#################  GOODNESS OF FIT TESTS ############
#####################################################
```

A.3 FOR GEV

A.3.1 Kolmogorov-Smirnov Test

```
    library(EnvStats)
    gofTest(GEVData, distribution = "gev",test = "ks")
```

A.4 FOR GPD

A.4.1 Kolmogorov-Smirnov Test

```
K.S<-function(data, u){
  if(!require(stats4)){
    install.packages("stats4")
    library(stats4)}
  u<-u
  uplus<-data[which(data>u)]
  excess<-uplus-u
  n<-length(excess)
  ii<-c(1:n)
  Exp<-function(lambda){
    ll <- n * log(lambda) - lambda*sum(excess)
    return(-ll) }
  suppressWarnings(
    exp.est <- mle(Exp, start=list(lambda=mean(excess))))
  lambda<-as.numeric(exp.est@coef)
  rate<-1/lambda
  Dm<- max(abs(1-exp(-(excess/rate)) - 1/ n),
           abs(1-exp(-(excess/rate)) - (ii-1)/ n))
  return(cat("Kolmogorov-Smirnov statistic: ",Dm,"\n"))}
KolSmir<-K.S(data = ProjectData, u = 272)
```

A.4.2 Cramer Von Mises Test

```
CVMTest<-function(data, u){
  if(!require(eva)){
    install.packages("eva")
    library(eva)}
  u<-u
  uplus<-data[which(data>u)]
  excess<-uplus-u
  me<-length(excess)
  CVMT<-  gpdCvm(excess)
  CVMstat<-CVMT$statistic
  CVMpval<-CVMT$p.value
  return(cat("Cramer-Von Mises Statistic: ",CVMstat,"\n",
             "P-value: ",CVMpval,"\n"))}
CVMTest(data = ProjectData, u = 272)
```

A.4.3 Anderson- Darling Test

```
ADTest<-function(data, u){
  if(!require(eva)){
    install.packages("eva")
    library(eva)}
  u<-u
  uplus<-data[which(data>u)]
  excess<-uplus-u
  me<-length(excess)
  ADT<-gpdAd(excess)
  ADstat<-ADT$statistic
  ADpval<-ADT$p.value
  return(cat("Anderson-Darling statistic: ",
             ADstat,"\n","P-value: ",ADpval,"\n"))}
ADTest(data = ProjectData, u = 272)

#####################################################
#########   STATISTICAL CHOICE OF MODELS   ##########
#####################################################
```

A.5 FOR GEV

A.5.1 Gumbel Test

```
GumTest<-function(data){
  if(!require(goft, evd)){
    install.packages("goft", "evd")
    library(goft, evd)}
  data<- data
  mi<-length(data)
  bm<-(log(mi)+log(log(2)))/(log(log(mi))-log(log(2)))
  am<-1/(log(log(mi)))
  GTS<-ev_test(data, dist = "gumbel", method = "ratio",
              N= 10000)
  GTstar<- (GTS$statistic - bm)/am
  pvalue<-pgumbel(GTstar)
  cat("[1] Test Statistic=",GTS$statistic," Critical Value*=",GTstar,
  " p-value=",pvalue,"\n")}
GumTest(data = GEVData)
```

A.6 FOR GPD

A.6.1 Gomes and Van Monfort(1986) test

```
GMT<-function(data,u){
  u<-u
  uplus<-data[which(data>u)]
  excess<-uplus-u
  n<-length(excess)
  Gvm<-uplus[n]/uplus[floor(n/2)+1]
  Gvmstar<-log(2)*Gvm-log(n)
  Pval<-evd::pgumbel(Gvmstar)
  Decision<-if (Gvmstar<=Gvm){cat("Reject Ho","\n",
                                " Gm*=",Gvmstar, "<=",
                                " Gm=",Gvm,"\n",
                                "P-value=",Pval,"\n")
  } else    {
    cat("Fail to reject Ho","\n",
        " Gm*=",Gvmstar, ">=",
        " Gm=",Gvm,"\n",
        "P-value=",Pval,"\n")}
  return(Decision)
}
```

```
GMT(data = ProjectData, u= 272)
```

A.6.2 Marohn (2000) GPD Test

```
Marohn<-function(data,u){
  u<-u
  uplus<-data[which(data>u)]
  excess<-uplus-u
  n<-length(excess)
  Mar<-0.5*((var(uplus)*(n-1)/n)/(mean(uplus)-u)^2-1)
  Marstar<-sqrt(n)*Mar
  pvaluniMarstar<-pnorm(Marstar)
  Decision<-if (Marstar<=pvaluniMarstar){cat("Reject Ho","\n",
                                         " t_m*=",Marstar, "<=",
                                         " p-value=",pvaluniMarstar,"\n")
  } else    {
    cat("Fail to reject Ho","\n",
        " t_m*=",Marstar, ">=",
        " p-value=",pvaluniMarstar)}
  return(Decision)

  }
Marohn(data = ProjectData, u = 272)
```

A.6.3 Likelihood Ratio Test For GEV, GPD and tGPD

```
LRT<-function(data, type = c("GEV", "GPD", "tGPD"), u = NULL, tGPDLL = NULL, tGPDOLL = NULL
              {
  if(!require(extRemes,ismev)){
    install.packages("extRemes","ismev")
    library(extRemes,ismev)}
  data<- data
  uplus<-data[which(data>u)]
  excess<-uplus-u
```

```
n<-length(excess)

if (type =="GEV")

{

  me<-length(data)

  Gumb<-fevd(data, type = "Gumbel", method = "MLE")

  GEVm<-ismev::gev.fit(data)

  loglikgum<--Gumb$results$value

  loglikgev<--GEVm$nllh

  LRgev<-round(-2*(loglikgum-loglikgev),2)

  LRgevstar<-round(LRgev/(1+2.8/me),2)

  pvalueLRgev<-pchisq(LRgevstar,1,lower.tail=F)


  print(cat("LRT=",LRgev," LRT*=",LRgevstar," P-value=",pvalueLRgev,"\n"))


} else

if (type == "GPD")

{

  if(!is.null(u)) u <- u


if(!require(stats4)){

  install.packages("stats4")

  library(stats4)}


ii<-c(1:n)

Exp<-function(lambda){

  ll <- n * log(lambda) - lambda*sum(excess)

  return(-ll)

}

suppressWarnings(

exp.est <- mle(Exp,

               start=list(lambda=mean(excess))))
  Expll<--exp.est@min

  loglikgpd<--GP$Loglikelihood

  n<-length(excess)

  LRT<-round(-2*(Expll-loglikgpd),2)

  LRTstar<-round(LRT/(1+4/n),2)

  pvalueLRT<-pchisq(LRTstar,1,lower.tail=F)


  print(cat("[1] l=",LRT," l*=",LRTstar," p-value=",pvalueLRT,"\n"))

  }

  if (type == "tGPD")

  {

    if(!is.null(tGPD0LL)) tGPD0LL <- tGPD0LL

    if(!is.null(tGPDLL)) tGPDLL <- tGPDLL

    n<-length(excess)

    LRT<-round(-2*(tGPD0LL-tGPDLL),2)

    LRTstar<-round(LRT/(1+4/n),2)

    pvalueLRT<-pchisq(LRTstar,1,lower.tail=F)

    print(cat("[1] l=",LRT," l*=",LRTstar," p-value=",pvalueLRT,"\n"))


  }

  }


LRT(data = GEVData, type = "GEV")

LRT(data = ProjectData, type = "GPD", u = 272)

LRT(data = ProjectData, type = "tGPD", u = 272, tGPD0LL = Est$Loglikelihood, tGPDLL = likk$
```

A.7 GEV Parameter Estimation

A.7.1 Maximum Likelihood Estimation

For $(\gamma \neq 0)$

```
library(extRemes)

fitGEVMLE<-fevd(GEVData, type = "GEV", method = "MLE")

fitGum<-fevd(GEVData, type = "Gumbel", method = "MLE")

ci(fitGEVMLE, type = "parameter")

ci(fitGum, type = "parameter")
```

A.7.2 Bayesian Estimation

For $(\gamma = 0)$

```
init <- c(1,1)

logpost0<-function(data, params)

{  n<-length(data)

  sigmar<-params[1]

  mu<-params[2]

  ll<--n*(log(sigmar)) - sum(exp(-((data-mu)/sigmar)))-sum(((data-mu)/sigmar))

  const<-0.5772156649

  logprior<-1/(sigmar*exp(-const))

  logpost<-ll+ logprior

    return(logpost)}

Est <- ru(logf = logpost0, d = 2, n = n, init = init,

        lower = 0, rotate = FALSE, trans = "BC", data = GEVData,

        a_method = "Nelder-Mead", var_names = c("Sigma", "Mu"))

summary(Est)
```

For $(\gamma \neq 0)$

```
library(revdbayes)

prior_dist <- set_prior("mdi", "gev", min_xi = -1)

post_dist <- rpost(50000, "gev", GEVData, prior_dist)

summary(post_dist)
```

A.8 GPD Parameter Estimation

A.8.1 Maximum Likelihood Estimation

For both ($\gamma \neq 0$) and ($\gamma = 0$)

```
    MLEGPD<-function(data, u, type =c("Exponential", "GP"))
{
  if(!require(data.table)){
    install.packages("data.table")
    library(data.table)
  }
  u<-u
  uplus<-data[which(data>u)]
  excess<-uplus-u
  n<-length(excess)
  if (type == "Exponential")
  {
    MLE.Exp<-function(data, params)
    {
      sigmar<-params[1]
      logl<--n*log(sigmar)-(1/sigmar)*sum(excess)
      return(-logl)
    }
      OptExp<-optim(f=MLE.Exp,
                            par=0.1,
                            lower = 0,
                            upper = 10,
                            hessian=TRUE,
                            method = "Brent",
                            # Custom Inputs
                            data = data)
      MLEGPD_par <- OptExp$par
      invMat<-  solve(OptExp$hessian)
      MLEGPD_SE <- sqrt((diag(invMat)))
      MLE <- data.table(param = c("sigma"),
                        estimates = MLEGPD_par,
                        se = MLEGPD_SE)
      print(MLE)
    }

  else if (type == "GP")
  {
    MLE.GPD<-function(data, params)
    {
      xi<-params[1]
      sigmar<-params[2]
      #if (xi == 0 | sigmar <= 1)
       # return(1e7)
      logl<-(-n*log(sigmar)-((1/xi)+1)*sum(log(1+(xi*excess)/sigmar)))
      return(-logl)
    }
      MLEGPD_opt <- optim(fn=MLE.GPD,
                          par=c(1,3),
                          lower = c(-Inf, -Inf),
                          upper = c(Inf, Inf),
                          hessian=TRUE,
                          method = "Nelder-Mead",
                          # Custom Inputs
                          data = data)
      MLEGPD_par <- MLEGPD_opt$par
      invMat<- solve(MLEGPD_opt$hessian)
      MLE_SE <- sqrt((diag(invMat)))
      MLE <- data.table(param = c("xi", "sigma"),
                        estimates = MLEGPD_par,
                        se = MLE_SE)
      print(MLE)
    }
  }


####### Estimations
GPExp<-MLEGPD(data = ProjectData, u = 272, type = "Exponential")
GP<-MLEGPD(data = ProjectData, u = 272, type = "GP")


####### Exp Parameters
scaleGPExp<-GPExp$estimates


####### GP parameters
shapeGPD<-GP$estimates[1]
scaleGPD<-GP$estimates[2]



######      Confidence Intervals    ########
############################################
conf.l <- 0.95
# For GP
critval <- qnorm((1 + conf.l)/2)
invfishGP<-GP[[2]]

shapeGPD + c(-1, 1) * critval * sqrt(invfishGP[1,1])
scaleGPD + c(-1, 1) * critval * sqrt(invfishGP[2,2])

###   For Exponential
critval <- qnorm((1 + conf.l)/2)
invfishExp<- GPExp[[2]]
scaleGPExp + c(-1, 1) * critval * sqrt(invfishExp[1,1])
```

A.8.2 Bayesian Estimation

For $\gamma \neq 0$

```
library(revdbayes)

u<-272

GPDprior_dist <- set_prior(prior = "mdi", model = "gp", min_xi = -1)

GPDpost_dist <- rpost(n = 10000, model = "gp", prior = GPDprior_dist,

data = ProjectData, thresh = u)

summary(GPDpost_dist)
```

A.9 Truncated RT-POT Parameter Estimation

For $\gamma \neq 0$

A.9.1 Parameter Estimation

```
myTruncorig<-function (data, start, eps = 10^(-10),...)

{

  X <- sort(data)

  n <- length(X)

  gamma <- numeric(n)

  tau <- numeric(n)

  K <- 1:(n - 1)

  conv <- numeric(n)

  start_orig <- start

  for (k in (n - 1):2) {

    if (k != (n - 1)) {

      start <- c(gamma[k + 1], tau[k + 1])

    }

    E <- X[n - (1:k) + 1] - X[n - k]

    E[1] <- X[n] - X[n - k]

    lik <- function(x) {

      gammapar <- x[1]

      taupar <- x[2]
      a <- 1 + taupar * E[-1]
      beta <- 1 - (1 + taupar * E[1])^(-1/gammapar)
      if (gammapar/taupar < eps | min(a) < eps | 1 + taupar *
          E[1] < eps) {
        L <- -10^6
      }
      else {
        L <- (k - 1) * log(taupar/gammapar) - (1 + 1/gammapar) *
          sum(log(a)) - (k - 1) * log(beta)
      }
      return(-L)
    }
    if (!is.numeric(lik(start)) | !is.finite(lik(start)) |
        is.nan(lik(start)) | lik(start) == 10^6) {
      start <- start_orig
    }
    sol <- optim(start, fn = lik, method = "Nelder-Mead", hessian = T)
    conv[k] <- sol$conv
    gamma[k] <- sol$par[1]
    tau[k] <- sol$par[2]
    invMat<- solve(sol$hessian)
    MLE_SE <- sqrt((diag(invMat)))
    AIC<-(2*2) - 2*-sol$value
    BIC<- - 2*-sol$value +2*log(n)
    critval <- qnorm((1 + 0.95)/2)
    c.int<- sol$par + c(-1, 1) * critval * sqrt(invMat)
    names(sol$par)<- c("Gamma","Tau")
    #plot(x = k, y = sol$par[1], type = "l", xlab = "k", ylab = "gamma")
    return(list(Estimates = sol$par, Confidence_Interval = c.int,
    LL = -sol$value, SE= MLE_SE, AIC = AIC, BIC = BIC))
    }  }
```

```
    myfun<-myTruncorig(prodat, start = c(1, 1), eps = 10^(-10),
       plot = TRUE, add = FALSE, main = "",
       ylim=c(0,4), xlim=c(120,450))
    gam<- as.numeric(myfun$Estimates[1])
    taup<-as.numeric(myfun$Estimates[2])
    sig<-gam/taup
```

MLE Estimator For γ as a Function of k

```
    par(mar=c(4,4,1,1))
TruncMLE<-function(data, u){
  prodat<-data[data>=u]
  prodat<-sort(prodat)
  TMLE<-trMLE(prodat, start = c(1, 1), eps = 10^(-10),
                   plot = TRUE, add = FALSE, main = "",
                   ylim=c(-0.5,0.5), xlim=c(400,1050)) #Estimates for EVI
  return(TMLE)
}
TrunMLE<-TruncMLE(data = ProjectData, u = 272)
```

A.9.2 Estimation of Other Parameters

For $\gamma \neq 0$ For Upper Endpoint (T_k)

```
    par(mar=c(4,4,1,1))
TruncEnd<-function(data, u){
  prodat<-data[data>=u]
  prodat<-sort(prodat)
  EndTrunc<-trEndpointMLE(prodat, gamma=TrunMLE$gamma,
                        tau=TrunMLE$tau, plot=TRUE,
              ylim=c(277.1,277.7), main="")
return(EndTrunc)
}
Endpointr<-TruncEnd(data = ProjectData, u = 272)
EndpointEst<-tail(Endpointr$Tk,10)[1]
```

For Small Exceedance Probability

```
    par(mar=c(4,4,1,1))
SEP<-function(data, u){
  prodat<-data[data>=u]
  prodat<-sort(prodat)
  Sep<-trProbMLE(prodat, gamma=TrunMLE$gamma,
          tau=TrunMLE$tau,
          DT=DTTrunc$DT,
          plot=TRUE, q=273.573, main = "", ylim = c(0,0.025))
}
crush<-SEP(data = ProjectData, u = 272)
```

For Truncation Odds

```
    par(mar=c(4,4,1,1))
TruncDT<-function(data, u, gamma, tau){
  prodat<-data[data>=u]
  prodat<-sort(prodat)
  TruDT <- trDTMLE(prodat, gamma=gamma,
                 tau=tau, plot=TRUE, ylim=c(0,0.03), xlim=c(75, length(prodat)), main = "")
  return(TruDT)
}
DTTrunc<-TruncDT(data = ProjectData, u = 272, gamma=TrunMLE$gamma, tau=TrunMLE$tau)
```

For Large Quantiles of Parent Distribution

```
    par(mar=c(4,4,1,1))
par(mfrow= c(1,2))
LQ<-function(data, u){
  prodat<-data[data>=u]
  prodat<-sort(prodat)
  trQuantMLE(prodat, gamma=TrunMLE$gamma,
           tau=TrunMLE$tau, DT=DTTrunc$DT, plot=TRUE,
           p=c(0.01), xlim= c(0,800), ylim=c(276.6,277.5), main = "")
}

gof<-LQ(data = ProjectData, u = 272)
```

For Large Quantiles of Truncated Distribution

```
    par(mar=c(4,4,1,1))

par(mfrow= c(1,2))

LQOri<-function(data, u){

  prodat<-data[data>=u]

  prodat<-sort(prodat)

  trQuantMLE(prodat, gamma=TrunMLE$gamma,

             tau=TrunMLE$tau, DT=DTTrunc$DT, plot=TRUE,

             p=c(0.01), xlim= c(65,800), ylim=c(277,279),

             Y=TRUE, main = "")

}

LQOri(prodat, u = 272)
```

For $(\gamma 0)$

A.9.2.1 Parameter Estimation

```
Estimator<-function(data, thresh)

{

  if(!require(data.table)){

    install.packages("data.table")

    library(data.table)

  }

  exceeds <- data[which(data>thresh)]

  k <- exceeds - thresh

  N <- length(k)

  nll0 <- function(param){

  sigmar <- param[1]

  a <- log(1 - exp(k[1]/sigmar))

  ll<- -(N - 1)*log(sigmar)- (sum(k[-1]/sigmar))

  - (N - 1)*a
  return(-ll)
  }
  truncest <- optim(par = 1, fn = nll0, method = "Brent",
  upper = 4, lower = 0, hessian = T)
  MLEtGPD_par <- truncest$par
  invMat<-  solve(truncest$hessian)
  MLEtGPD_SE <- sqrt((diag(invMat)))
  MLE <- data.table(param = c("sigma"),
                    estimates = MLEtGPD_par,
                    se = MLEtGPD_SE)
  AIC<-(2*1) - 2*-truncest$value
  BIC<- - 2*-truncest$value +2*log(N)
  critval <- qnorm((1 + 0.95)/2)
  c.int<- MLEtGPD_par + c(-1, 1) * critval * sqrt(invMat)
  print(list(MLE_estimate = MLE, Confidence_Interval = c.int,
  Loglikelihood= -truncest$value, AIC= AIC, BIC= BIC))
}
Est<-Estimator(data = ProjectData, thresh = 272)
```

A.9.2.2 Other Parameter Estimation

For Upper Endpoint

```
    myTkfun<-function (data, sigma, plot = FALSE, add = FALSE,
    main = "Endpoint Estimates", ...)
{
  X <- sort(data)
  n <- length(X)
  K <- 1:(n - 1)
  E <-  X[n] - X[n - K]
  a <- exp(E[1]/sigma)
  b <- K - a
  Tn <- X[n - K] + sigma * log(1 + K*(a - 1/ b))
  plot(K, Tn[K], type = "l", xlab = "k", ylab = expression(T[k]))
  return(Tn)
}
Tkk<-myTkfun(prodat, sigma = Est$MLE_estimate$estimates)
tail(Tkk,n=1)
```

A.10 Truncation Test

```
testTr<-trTestMLE(prodat, gamma=TrunMLE$gamma, tau=TrunMLE$tau, main = "")
mean(testTr$Pval)
```

Appendix B

Tables

B.1 Gumbel Model Parameter Estimates

TABLE B.1: Maximum Likelihood and Bayesian Estimates for the location and scale parameters

	Maximum Likelihood Estimation		Bayesian Estimation	
Estimates	μ	σ	μ	σ
Parameters	249.73	17.37	249.6	17.67
Confidence Intervals	(248.23, 251.22)	(16.52, 18.21)	(249.2, 250.2)	(17.12, 17.69)
Standard errors	0.765	0.429		
Loglikelihood	-2480.25			
AIC	4964.5			
BIC	4973.24			

TABLE B.2: Return level, period, upper endpoint and exceedance probabilities for Gumbel model

	Maximum Likelihood		Bayesian	
Return Periods (Years)	Return Level	Exceedance Probability	Return Level	Exceedance Probability
2	256.0909	0.50	256.0506	0.50
5	275.7733	0.20	275.9989	0.20
10	288.8048	0.10	289.2065	0.10
20	301.3049	0.05	301.8754	0.05
50	317.4851	0.02	318.2741	0.02
100	329.6098	0.01	330.5626	0.01
	Value	Exceedance Probability	Value	Exceedance Probability
Maximum water level	276.64	0.019	276.64	0.1946478
Upper Endpoint (X^F)	-Inf	1	-Inf	1

B.2 Exponential Model Parameter Estimates

Table B.3: Parameter estimation of the generalized Pareto model for the Akosombo dam with MLE and Bayesian methods (Exponential case)

	Maximum Likelihood		Bayesian	
Estimates	σ	γ	σ	γ
Estimated parameters	1.862	/	1.835	/
Standard errors	0.05	/	/	/
Confidence Intervals	(1.759, 1.965)	(-0.510, -0.432)	/	/
Log-likelihood	-2056.385		/	
AIC	4114.769		/	
BIC	4127.06		/	

Table B.4: Return level, period, upper endpoint and exceedance probabilities for Exponential model

	Maximum Likelihood		Bayesian	
Return Periods (Years)	Return Level	Exceedance Probability	Return Level	Exceedance Probability
2	275.7031	0.50	275.6490	0.50
5	277.4094	0.20	277.3304	0.20
10	278.7002	0.10	278.6023	0.10
20	279.9909	0.05	279.8742	0.05
50	281.6972	0.02	281.5556	0.02
100	282.9880	0.01	282.8276	0.01
	Value	Exceedance Probability	Value	Exceedance Probability
Maximum water level	277.54	0.05104802	277.54	0.04884645
Upper Endpoint (X^F)	-Inf	1	-Inf	1

References

Aban, I. B., Meerschaert, M. M., & Panorska, A. K. (2006). Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, *101*(473), 270–277.

Albrecher, H., Beirlant, J., & Teugels, J. L. (2017). *Reinsurance: Actuarial and statistical aspects*. John Wiley & Sons.

Amin, N. A. M., Adam, M. B., & Aris, A. Z. (2015). Bayesian extreme for modeling high pm10 concentration in johor. *Procedia Environmental Sciences*, *30*, 309–314.

Balkema, A. A., & De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, *2*(5), 792–804.

Barabesi, L. (1993). *Random variate generation by using the ratio-of-uniforms method*. Nuova immagine.

Beirlant, J., Alves, I. F., & Gomes, I. (2016). Tail fitting for truncated and non-truncated pareto-type distributions. *Extremes*, *19*(3), 429–462.

Beirlant, J., Alves, I. F., & Reynkens, T. (2017). Fitting tails affected by truncation. *Electronic Journal of Statistics*, *11*(1), 2026–2065.

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2004). *Statistics of extremes: Theory and applications* (Vol. 558). John Wiley & Sons.

Benktander, G., & Segerdahl, C.-O. (1960). On the analytical representation of claim distributions with special reference to excess of loss reinsurance. *Transactions of the h˜ ternational Congress of Actuaries*.

Bhattarai, S., Zhou, Y., Zhao, C., & Yadav, R. (2016). An overview on types, construction method, failure and key technical issues during construction of high dams. *Electronic Journal of Geotechnical Engineering*, *21*(26), 10415–10432.

Blanchet, J., He, F., & Murthy, K. (2020). On distributionally robust extreme value analysis. *Extremes*, *23*(2), 317–347.

Castellanos, M. E., & Cabras, S. (2007). A default bayesian procedure for the generalized pareto distribution. *Journal of Statistical Planning and Inference*, *137*(2), 473–483.

Castillo, E. (2012). *Extreme value theory in engineering*. Elsevier.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). Springer.

Couturier, D.-L., & Victoria-Feser, M.-P. (2010). Zero-inflated truncated generalized pareto distribution for the analysis of radio audience data.

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

Demirer, R., Gkillas, K., & Suleman, T. (2019). Economic policy uncertainty and extreme events in the us stock market. *Unpublished working paper.*

Dodd, E. L. (1923). The greatest and the least variate under general laws of error. *Transactions of the American Mathematical Society*, *25*(4), 525–539.

Dosio, A., Mentaschi, L., Fischer, E. M., & Wyser, K. (2018). Extreme heat waves under 1.5 c and 2 c global warming. *Environmental Research Letters*, *13*(5), 054006.

Energy Commission, G. (2016). Energy supply and demand outlook for ghana.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical proceedings of the Cambridge philosophical society*, *24*(2), 180–190.

Frank, P. (1954). The work of richard von mises: 1883-1953. *Science*, *119*(3102), 823–824.

García-Pérez, M. Á. (2019). Bayesian estimation with informative priors is indistinguishable from data falsification. *The Spanish journal of psychology*, *22*.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, 423–453.

Gomes, M. I., & van Montfort, M. A. (1986). Exponentiality versus generalized pareto-quick tests. *Proc. III Internat. Conf. Statistical Climatology*, 185–195.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia university press.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163–1174.

Hosking, J. (1984). Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika*, *71*(2), 367–374.

Hosking, J., Wallis, J., & Wood, E. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, *27*(3), 251–261.

Hüsler, J., & Li, D. (2007). *Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields*. Birkhäuser.

IDMC. (2017). Dams and internal displacement [[Online; accessed 23-January-2023]]. https://www.internal-displacement.org/publications/case-study-series-dam-displacement

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

Kinderman, A. J., & Monahan, J. F. (1977). Computer generation of random variables using the ratio of uniform deviates. *ACM Transactions on Mathematical Software (TOMS)*, *3*(3), 257–260.

Koskinas, A., Tegos, A., Tsira, P., Dimitriadis, P., Iliopoulou, T., Papanicolaou, P., Koutsoyiannis, D., & Williamson, T. (2019). Insights into the oroville dam 2017 spillway incident. *Geosciences*, *9*(1), 37.

Kotz, S., & Nadarajah, S. (2000). *Extreme value distributions: Theory and applications*. World Scientific.

Kozubowski, T. J., Panorska, A. K., Qeadan, F., Gershunov, A., & Rominger, D. (2008). Testing exponentiality versus pareto distribution via likelihood ratio. *Communications in Statistics-Simulation and Computation*, *38*(1), 118–139.

Leščešen, I., & Dolinaj, D. (2019). Regional flood frequency analysis of the pannonian basin. *Water*, *11*(2), 193.

Lewin, J., Ballard, G., & Bowles, D. S. (2003). Spillway gate reliability in the context of overall dam failure risk. *USSD Annual Lecture*, 1–17.

Li, Y., & Jones, B. (2019). The use of extreme value theory for forecasting long-term substation maximum electricity demand. *IEEE Transactions on Power Systems*, *35*(1), 128–139.

Lilliefors, H. W. (1969). On the kolmogorov-smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, *64*(325), 387–389.

Ma, N., Bai, Y., & Meng, S. (2021). Return period evaluation of the largest possible earthquake magnitudes in mainland china based on extreme value theory. *Sensors*, *21*(10), 3519.

MacAfee, A. W., & Wong, S. W. (2007). Extreme value analysis of tropical cyclone trapped-fetch waves. *Journal of applied meteorology and climatology*, *46*(10), 1501–1522.

Marohn, F. (2000). Testing extreme value models. *Extremes*, *3*(4), 363–384.

Mettle, M. (2011). *Forced resettlement in ghana: The dam and the affected people: The bui hydroelectric power project in ghana* (Master's thesis). Norges teknisk-naturvitenskapelige universitet, Fakultet for . . .

Minkah, R. (2016). An application of extreme value theory to the management of a hydroelectric dam. *SpringerPlus*, *5*(1), 1–12.

Muela, S. B., Martın, C. L., & Sanz, R. A. (2017). An application of extreme value theory in estimating liquidity risk. *European Research on Management and Business Economics*, *23*(3), 157–164.

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

Nolde, N., & Zhou, C. (2021). Extreme value analysis for financial risk management. *Annual Review of Statistics and Its Application*, *8*, 217–240.

Northrop, P. J., & Attalides, N. (2016). Posterior propriety in bayesian extreme value analyses using reference priors. *Statistica Sinica*, 721–743.

Nuyts, J. (2010). Inference about the tail of a distribution: Improvementon the hill estimator. *International Journal of mathematics and mathematical sciences*, *2010*.

Ocran, E., Doku-Amponsah, K., & Nortey, E. (2017). Estimating exceedance probability of extreme water levels of the akosombo dam.

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

Prescott, P., & Walden, A. (1983). Maximum likeiihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. *Journal of Statistical Computation and Simulation*, *16*(3-4), 241–250.

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

Skakun, S., Kussul, N., Kussul, O., & Shelestov, A. (2014). Quantitative estimation of drought risk in ukraine using satellite data. *2014 IEEE Geoscience and Remote Sensing Symposium*, 5091–5094.

Smith, R. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, *72*(1), 67–90.

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

Smith, R., & Goodman, D. (2000). Bayesian risk analysis. chapter 17 of extremes and integrated risk management, edited by p. embrechts.

Spearing, H., Tawn, J., Irons, D., Paulden, T., & Bennett, G. (2019). Ranking, and other properties, of elite swimmers using extreme value theory. *arXiv preprint arXiv:1910.10070*.

Tancredi, A., Anderson, C., & O'Hagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, *9*(2), 87–106.

Verster, A., de Waal, D., Schall, R., & Prins, C. (2012). A truncated pareto model to estimate the under recovery of large diamonds. *Mathematical Geosciences*, *44*(1), 91–100.

Vicente, S. L. G. (2012). *Extreme value theory: An application to sports* (Doctoral dissertation).

Vitanov, N. K., Angelova, P., & Dimitrova, Z. I. (n.d.). Analysis of extreme water levels at the mississippi river in the usa.

Von Bortkiewicz, L. (1922). Variationsbreite and mittlerer fehler. *Sitzungsber. Berli. Math. Ges*, *21*, 3–11.

Von Mises, R. (1923). Über die variationsbreite einer beobachtungsreihe. *Sitzungsberichte der Berliner Mathematischen Gesellschaft*, *22*(3).

Wu, G., & Qiu, W. (2018). Threshold selection for pot framework in the extreme vehicle loads analysis based on multiple criteria. *Shock and Vibration*, *2018*.

Zellner, A. (1995). *Past and recent results on maximal data information priors.* University of Chicago, Graduate School of Business, Department of Economics.