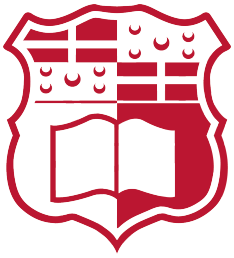


Univariate and Multivariate Extreme Value Analysis of Düsseldorf Hydrological Data

Owen Vella

Supervised by Dr Fiona Sammut

Co-supervised by Dr David Paul Suda

Prof. Pierre-Sandre Farrugia

Department of Statistics

Faculty of Science

University of Malta

March 2023

A dissertation submitted in partial fulfilment of the requirements for the Master of Science in Statistics.

Abstract

The rise in catastrophic climate events during the late 20^{th} century prompted an increase in the application of statistical methods based on extreme value theory (EVT) in the fields of hydrology, climate, and meteorology. Several statistical models have been developed over the years. This dissertation presents an in-depth review of the fundamental univariate and multivariate techniques that rely on asymptotic EVT results.

This dissertation focuses on univariate methods, including the Block Maxima (BM) method, the K largest Order Statistics (KLOS) method, and the Peak over Threshold (POT) method. Also, the Component-wise Block Maxima (CWBM) method and the General Copula-based (GCB) method are covered as multivariate methods.

These methods are applied to the monthly mean of river discharge observations and the collective impact of snow melt and precipitation excess observations that were collected from Düsseldorf stations. Through these models, the return period and return level metrics are used to assess whether flood risk mitigation measures are sufficient for Düsseldorf and if the univariate analysis is still important to be taken into consideration when implementing flood protection measures in light of the inter-relationship between extreme events.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my supervisors Dr Fiona Sammut, Dr David Paul Suda and Prof Pierre-Sandre Farrugia for their guidance and constant contributions at every step of the process. Also, I would like to extend this gratitude towards all individuals that supported and assisted with valuable input at some point or another.

Finally, I would also like to thank my family, especially my parents, who always gave me unconditional support throughout my academic years, from whom I get my drive and strength to push my boundaries to reach higher goals and to never give up. Their encouragement and tolerance was and still will be my source of motivation to do more.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

- Abbreviations

ACF	Autocorrelation Function
AD	Anderson-Darling
AIC	Akaike Information Criterion
AMBAV	Agrometeorological Calculation of Current Evaporation
BfG	Bundesanstalt für Gewässerkunde
BIC	Bayesian Information Criterion
BM	Block Maxima
CCF	Cross Correlation Function
CI	Collective Impact
CLT	Central Limit Theorem
CWBM	Component-wise Block Maxima
DWD	Deutscher Wetterdienst
EVC	Extreme Value Copula
EVI	Extreme Value Index
EVT	Extreme Value Theory
GCB	General Copula-based
GEV	Generalized Extreme Value

GP	Generalized Pareto
GRDC	Global Data Runoff Center
HS	Hazard Scenario
IID	Independent and identically distributed
IFM	Inference Functions for Margins
IPCC	Intergovernmental Panel on Climate Change
KLOS	K Largest Order Statistic
MDA	Maximum Domain of Attraction
MEVT	Multivariate Extreme Value Theory
MLE	Maximum Likelihood Estimation
MRD	Monthly River Discharge
MRL	Mean Residual Life
MRS	Monthly Rolling Sum
NSD	New Snow Depth
OS	Order Statistics
P	Precipitation
PCC	Pair Copula Constructions
PE	Potential Evapotranspiration
$\mathbf{PE}_{AMBAV}$	Potential Evapotranspiration estimated by the AMBAV Method
$\mathbf{PE}_{HS}$	Potential Evapotranspiration estimated by the HS Method
POT	Peaks-Over-Threshold
PP	Phillips-Perron
RD	River Discharge

RL	Return Level
RP	Return Period
RS	Rolling Sum
SD	Snow Depth
SMWE	Snow Melted Water Equivalent
T	Temperature
TSD	Total Snow Depth
UEVT	Univariate Extreme Value Theory
WV	Wind Velocity
WL	Water Level
WSV	German Federal Waterways and Shipping Administration

- Sets of numbers

$\mathbb{N}$	Naturals
$\mathbb{R}$	Reals

- Statistical operators

$\mathbb{E}(\cdot)$	Expectation
$\mathbb{P}(\cdot)$	Probability

- Miscellaneous

$\xrightarrow{\mathbf{p}}$	Convergence in probability
$\xrightarrow{\mathbf{a.s.}}$	Convergence Almost Surely
$\xrightarrow{\mathbf{d}}$	Convergence in Distribution
$\mathbb{I}_A$	Indicator Function
$\mathcal{N}(\mu, \sigma^2)$	Normal Distribution with mean μ and variance σ^2

Chapter 1

Introduction

In general, the main objective of any form of classical statistical theory is to develop a representation that quantifies the usual behaviour of the observed events. In fact, to reduce any bias in model estimates, observations that show extreme behaviour are often considered as outliers and hence disregarded to restrict their influence [101]. However, in disciplines such as engineering [6], finance [82], medicine [23], climatology [9], agriculture [127], actuarial science [128], the characterization of extreme observation's behaviour is instead desired.

This is due to the fact that extreme observations, which lie at the tails of the distribution, are expected to occur with low frequency but with high potential for devastating consequences. For instance, the drought that occurred in the US between 1987 and 1989 incurred around $39 billion in damages [112]. The heavy snowfalls that took place in Bavaria around January of 2006 generated several deaths and caused catastrophic damages to buildings and infrastructure [150]. According to Swiss Re Institute, in 2021 the estimated worldwide economic losses from natural catastrophes were around $280 billion, with the destructive floods in July causing losses of around $41 billion to Germany and its neighbouring countries [81]. As extreme observations are rare and located at the tails of the distribution, then any standard statistical theory that does not account for this type of data will produce unreliable inferences. Given the impact of extreme events and the increase in their frequency and magnitude, the Intergovernmental Panel

on Climate Change (IPCC) 2022 report [143] emphasized the importance of accurately modelling and estimating the frequency and impact of these events. This will enable effective adaptation efforts and continue to reduce the impact of the residual risks.

A probabilistic framework that is suitable to investigate and infer results about the properties of extreme observations is extreme value theory (EVT). In fact, this dissertation is primarily concerned with the investigation of hydrometeorological extreme events explored within this framework. From the comprehensive literature addressing extreme value theory, the general introduction of the present text refers to Coles [26], Leadbetter et al. [102], and Salvadori et al. [134], whereas Falk et al. [43], Resnick [126] and Laurens & Ferreria [70] are referred to when more detail on the probabilistic background was required. Additional literature on copula theory that is relevant to extreme analysis can be found in McNeil et al. [106] and Nelsen [111].

1.1 Historical Development

1.1.1 Univariate Extreme Value Theory

The mathematical discipline of EVT is relatively young compared to other branches of probability and statistics. Much of the initial development in the analysis of extreme events can be attributed to astronomers trying to identify methods for distinguishing and handling outliers in astronomical data [95, 97]. At that time, it was already noted that there was a lack of literature on statistical techniques that properly dealt with extreme observations. However, it was not until 1922 that the analysis of extremes achieved a systematic development in the context of a general framework, later to be known as the classical univariate EVT (UEVT), with the submission of a crucial paper written by Bortkiewicz [14].

In this paper, the concept that the largest values obtained from Gaussian distributed samples are observations represented by an entirely different distribution was investigated. Eventually, in 1927, Maurice Fréchet [49] was the first to establish the asymptotic

parametric distribution of the sample maxima from a sample of independent and identically distributed (IID) observations generated from a class of particular distributions. In the following year, independently from Fréchet's publication, Ronald Fisher and Leonard Tippett [47] have produced the same limiting distribution for the sample maxima for that particular class of distribution functions, and developed two other types of limiting distributions for other classes of distribution functions. It was later found that the three types of limiting distributions found in [47] covered all possibilities when considering all classes of distribution functions. Therefore, the content of this paper is regarded as the backbone of classical UEVT.

Subsequently, Gnedenko [58] formalized the statistical framework of UEVT by establishing the necessary and sufficient conditions for the weak convergence of the sample maxima to the three types of non-degenerate limiting distributions. Eventually, this work was refined by Mejzler [108] and Marcus & Pinsky [105]. Until then, the statistical literature about the analysis of extreme observations using UEVT was mainly concerned with the mathematical formulation of the theory, and the application side of the framework was scarce. Nevertheless, Gumbel was at the statistical forefront to apply the advantages of the UEVT to practical applications such as the extreme duration of human lifetimes [62], radioactive emissions [63] and flood analysis [64, 66]. Several of his contributions were collected in a single book [65] and for some time this text was considered as the main inferential text.

In 1955, Jenkinson [86] introduced the Generalized Extreme Value (GEV) distribution with the intention to be used instead of the three distinctive limiting distributions as it parametrically unifies them into one. The estimation of the parameters of the GEV distribution using maximum likelihood estimation was studied by various researchers, such as Jenkinson [85], Prescott & Walden [120] and Hosking [76], just to name a few. There was other published material that examined other methods to estimate the parameters of the GEV distribution [25, 75]. Up until then, the research community was still lacking contributions in expanding the univariate theory. However, things turned around when, in the early 1970's De Haan [68, 69] published an innovative theoretical

approach to model the excess of the extreme observations; the approach was based on the same assumptions defined in [47]. Despite that, it was Balkema & De Haan [8] and Pickands [79] who established that the asymptotic distribution of the excess of IID observations exceeding the limiting threshold is of the Generalized Pareto (GP) distribution form. The estimation of the GP distribution parameters using maximum likelihood estimation was thoroughly reviewed by various researchers, such as Davison & Smith [34, 35], Hosking et al. [77], Grimshaw [60] and Castillo et al. [19]. Similar to the GEV case, alternative methods to estimate the parameters of the GP distribution were also proposed [69, 141]. Accordingly, the classical UEVT is generally split between two parametric approaches. The first approach is the block maxima (BM) which is based on the Fisher and Tippett theorem and consequently the GEV distribution is used. The second approach is the peaks over threshold (POT) method which is based on the Pickands, Balkema and de Haan theorem and subsequently the GP distribution is used.

1.1.2 Multivariate Extreme Value Theory

Multivariate extreme analysis plays a crucial part in the analysis of extreme events, and it is of particular interest in several fields such as finance, actuarial science, hydrology, climatology, and engineering. Coles and Tawn pointed out that climate-related processes are often composed of multiple underlying univariate processes [28]. An extreme event involving the co-occurrence of multiple underlying events is called a compound event. Based on the IPCC risk framework, Zscheischler et al. [168] defined this term as "the combination of multiple drivers and/or hazards that contribute to societal or environmental risk" as shown in Figure 1.1.

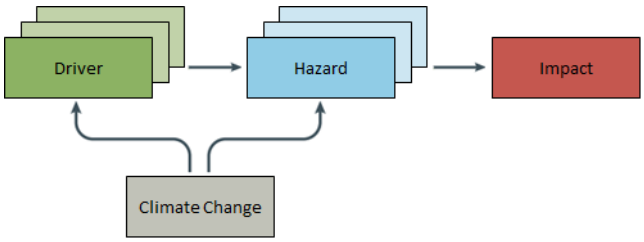


Figure 1.1: Components that constitute a compound event.

For instance, in a hydrological and climate context, the drivers of a compound flooding could be the combination of precipitation, river flow, and wind speed, in which the individual drivers might not be extreme but their combined occurrence creates an extreme hazard. In this case, the hazard would represent the flood event that will trigger the negative impact (flood damage) on human and natural systems. The joint behaviour underlying a compound event cannot be examined by the univariate models mentioned in Section 1.1.1, as the inter-relationships are not factored into the univariate models [22]. UEVT is not able to completely apprehend the complexity of the probability structure of extreme events.

If the relationships between multiple variables are ignored then this can result in an underestimation of the risks involved. For instance, an assessment of the storm event that occurred in June 2016 at Shoalhaven Estuary highlighted that, as a consequence of disregarding the additional impact of the river discharge, the flood damage was underestimated by 30% [100]. In [2], it was demonstrated that the simultaneous occurrence of drought and heat magnifies the repercussions in the United States. Furthermore, Zheng et al. [166] noted that even if the dependence structure between the drivers of an extreme event is weak, the compound event may produce some serious impact. During the early stages of the development of multivariate extreme value theory (MEVT), the shift from univariate models to multivariate models created non-trivial difficulties. These issues hindered the progress of the research on multivariate extreme analysis.

The initial proposed multivariate extreme models found in the literature are all based on a direct extension of some univariate distributions. For instance, the multivariate normal distribution [3], the bivariate gamma distribution [164] and the bivariate exponential distribution [45] were frequently used to model multivariate extreme observations. These multivariate distributions suffer from some drawbacks, such as the assumption that the marginals belong to the same probability family and that the observations must follow a certain dependence structure. In addition, some extensions can only be made to the bivariate case. In order to overcome some of the difficulties and restrictions, MEVT was developed as an extension and a generalization of the

univariate extreme probabilistic framework. However, as interest in the research community continued to grow, it was found that contrary to the univariate framework, the class of multivariate extreme distributions cannot be captured by a finite parametric family. The initial contributions to the extension of the GEV distribution were in the bivariate domain [41, 53, 140]. Some years had to pass until the characterization of the class of asymptotic multivariate extreme value distributions was developed by De Haan & Resnick [36] and Pickands [80]. Nevertheless, it was Tawn [152] who formulated several parametric models and inferential statistical techniques for the component-wise block maxima (CWBM) model, which is a direct extension of the univariate BM model.

An alternative technique that is still in development is based on the extension of the GP distribution into the multivariate domain to exploit the corresponding characteristics introduced by the univariate POT method. In more recent literature [44, 129, 130], it was indicated that the class of multivariate GP distributions is adequate for the multivariate POT framework that models the exceedances of an IID random variable over a high threshold. Nonetheless, the class of multivariate GP distributions is still under scrutiny due to the variations in the proposed definitions, and the methods that select the proper threshold for this distribution still possess crucial problems [96]. In [159], it was noted that even though the concept of compound events is more frequently considered, in excess of 95% of the publications are still using UEVT.

A more recently proposed alternative to model multivariate extremes, which is becoming increasingly popular, does exist. An effective and convenient way to construct multivariate extreme distributions is by means of the copula function. Sklar's Theorem [145] allows the disentanglement of the marginal distributions from the corresponding dependence structure [50, 144]. The utilization of copulas in extreme data analysis allowed for high flexibility in estimating and summarizing the information on the dependence structure between the variables being studied [125, 138]. There are still some shortcomings in using this methodology, such as the fact that the copula structure only allows for homogeneous dependence between the multiple drivers or hazards of the extreme event, and some copulas cannot model dependence in the tails well. Also, there

is a plethora of bivariate copula families, but the set containing copula families with dimensions greater than two is limited [87]. However, there are numerous fields that capitalize on the advantages of the copula structure [78, 137]. Some possible limitations that might be experienced by using copulas motivated the introduction of a new way of applying multivariate models called the vine copula or pair-copula. The pair-copula constructions (PCC), an innovative approach introduced by Joe [89], provided a lot of flexibility by decomposing the multivariate distribution function into the product of a cascade of independent bivariate copulas and corresponding marginal distributions. Daneshkhah et al. [33] tested the performance of PCC against five standard multivariate copulas and demonstrated that the results obtained when approximating the joint distribution are considerably better when considering the PCC structure.

1.1.3 Application to Weather and Climate Science

As already mentioned, in the first half of the 20^{th} century, the research community was predominantly concerned with the development of the fundamentals of the theoretical part of the univariate extreme analysis. However, in the second half of the 20^{th} century, as a result of Gumbel's work [65], the application side was given much more importance. Regardless, the earliest implementations were limited and mainly interested in fields such as finance, mortality, hydrology and engineering. Considering the rise in awareness on climate change [92] and as weather and climate catastrophic events have been on an increasing trend at the end of the 20^{th} century [15], the research community was compelled to apply EVT to weather and climate extreme data [61, 93, 113].

As a consequence of the properties observed from the data, it was demonstrated that the IID assumption considered in UEVT is unrealistic. Hence, the research community was inclined to consider results that extended the classical UEVT, for example, the extremal index method [149] or the practical approach taken by Coles [26] and Cooley [30] to modify the classical UEVT distribution parameters that vary with time. At the same time, the initial theoretical results of MEVT were being developed. Accordingly, the majority of published material that applied MEVT to weather and climate extreme

data started at the beginning of the 21^{th} century [26, 71]. Similar to the UEVT, the IID property assumed in the MEVT models is impractical considering that the probability of compound events is changing. This assumption is being relaxed, and new methodologies are being introduced to suit the characteristics of the data. For instance, although the copula approach has been substantially implemented in other areas of study, only recently has it been used for the analysis of weather and climate extreme events [132]. This also applies to the PCC structure, which is frequently used in recent research papers [13, 83]. The presented literature review includes all techniques used in EVT, but this dissertation covers the preliminary models together with the copula structure. The PCC structure will not be covered in this dissertation due to the extensive nature of the required foundational work.

1.2 Problem of Interest

One of the largest rivers in central-western Europe is the river Rhine. This river starts from the southern-eastern Swiss Alps, then passes through both Germany Rhineland and the Netherlands, and finally it ends up in the North Sea. According to [155], the drainage area is about $185\,260\,\mathrm{km}^2$ and the total length of the river Rhine is about 1250 km. Furthermore, in Germany, the river is considered to be an important major artery for services of transportation, production of water for drinking and irrigation, tourist attraction, industrial production, and power generation by the hydroelectric power plants [24]. Hence, the attributes of the river Rhine have attracted a lot of people to the urban areas that extend along the river, including Düsseldorf.

The increase in human and industrial activities and climate change have influenced the biodiversity, ecology and the structural integrity of the river. Over the last 200 years, the most notable impact of these changes has been an increased vulnerability to flooding in nearby cities [51]. Prior to the 19^{th} century, flooding disasters in the vicinity of the river Rhine only affected a small portion of the population with negligible impact [155]. However, the winter flood disasters of 1993 and 1995, which resulted from several heavy rainfall events, have amounted to astronomical damages [16]. Consid-

ering the fact that these compound extreme events will constantly keep on recurring [117] and that they cause significant damages to the nearby cities while disrupting the activities alongside the river, it is key to statistically investigate the associated risks and the adequacy of the flood control measures in a univariate and multivariate context.

Nonetheless, there are several factors that interplay for the exceedance of the capacity of the river channel. Hence, investigating and quantifying the flood risk of the river Rhine at a particular flood prone area can be quite difficult. For this study, the daily observations of precipitation, snow melted water equivalent (SMWE), potential evapotranspiration (PE), water level (WL) and river discharge (RD) are used as a proxy for floods (see Figure 1.2).

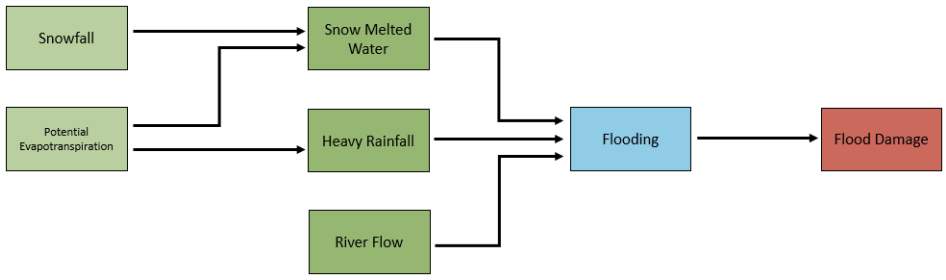


Figure 1.2: Overview of the considered compound event components.

Additional information and further analysis on the characteristics of the data used throughout this dissertation will be given in Chapter 2. There were several stations and river gauges scattered alongside the river that could serve as a case study in this dissertation. Considering the rarity of extreme events, the length of the data plays a central role in the reliability of any statistical analysis conducted on this data including model fitting and forecasting [94, 154]. It was thus decided to focus on the long-term-daily Düsseldorf dataset obtained from the station (ID: 1078) and river gauge (ID: 6335050) indicated in Figure 1.3. As the Rhine stream splits the city with a length of 42.1 km [20] and the geographical position of Düsseldorf is at the centre of the Lower Rhine basin between Cologne and Rees, in [98] it was also shown that Düsseldorf will be flooded under extreme conditions, providing further justification to choose Düsseldorf

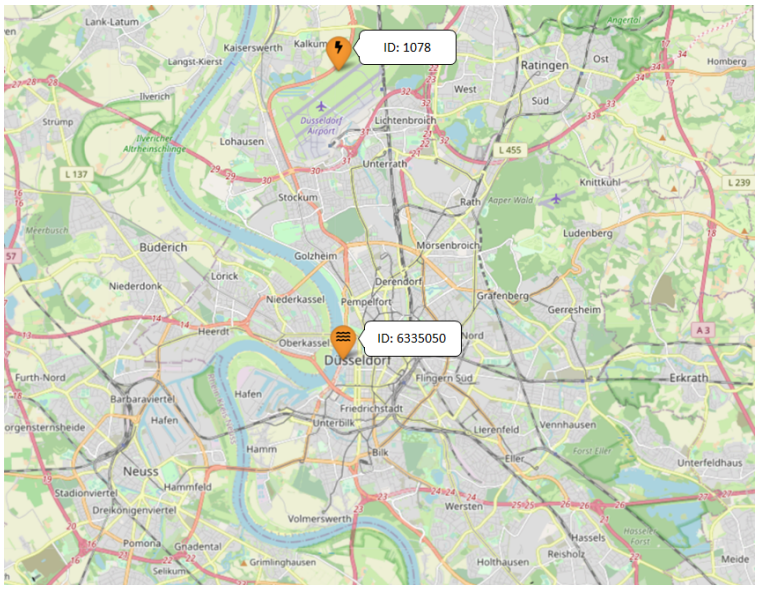


Figure 1.3: Düsseldorf meteorological station and river Rhine gauge.

for the case study for this dissertation. Subsequently, the aim of this dissertation is two fold. Firstly, the aim is to present in a rigorous manner the probabilistic framework behind the different methods used in the analysis of extreme events, in both univariate and multivariate dimensions. Secondly, the aim is to also implement the techniques covered in the theoretical sections, along with flood risk estimates, using the Düsseldorf observations collected from a nearby weather station and the Rhine river gauges as a case study.

1.3 Structure of the Dissertation

The remainder of this dissertation consists of four chapters, which are built upon the literature of extreme analysis. In Chapter 2, the description and the quality of the data being studied will be reviewed. The steps taken for pre-processing the Düsseldorf data in a format suitable for fitting statistical extreme models in all chapters will also be presented. In Chapter 3, a detailed introduction to the classical UEVT framework is given. This chapter covers the methodology behind the two main classical univariate methods, BM and POT. This chapter also introduces a modified version of the BM

method, called the K Largest Order Statistic (KLOS) method, which addresses some of the drawbacks of the BM method. In each section, the flood hazard is assessed in terms of the degree of rarity and the corresponding magnitude of the flood scenario through the return period (RP) and return level (RL) measures. After each section, the derived techniques are applied to the Düsseldorf data, and the results are then compared to each other. This chapter is crucial as it will set the foundations for the extension of the univariate theory into the multivariate domain.

In Chapter 4, the introduction to the classical MEVT framework will be presented by extending the arguments of the BM method derived in the previous chapter. Afterward, the corresponding copula framework is presented so that the extended version of the return period and return level concepts can be easily integrated. Dependence measures are introduced to assess the inter-relationships between observations. The two-step goodness-of-fit method will be used to assess the quality of the distributional fit. Subsequently, a General Copula-based (GCB) method will also be introduced to address some of the classical MEVT framework's drawbacks. Similar to the univariate case, after the derivation of a technique, it will be applied to the Düsseldorf data and the results from the different techniques are then compared.

In the final chapter, an overall summary of the work done, in particular highlighting the statistical results obtained by the fitted univariate and multivariate methods, will be laid out. In addition, the limitations and recommendations for further research directions that lead on from this work are put forward. Finally, Appendix A contains additional theory and results that support the progress of the main results produced in the previous chapters. While in Appendix B there is supplementary material that is referred to in Chapter 2 and in the case studies. All data extractions and results obtained for the Düsseldorf data are produced by routines in R that can be reproduced using the code uploaded in the GitHub repository [157].

Chapter 2

The Dataset

2.1 Introduction

Over recent years, the North Rhine-Westphalia state has experienced a number of river floods. Some of these floods occurred in the state capital, Düsseldorf. Consequently, it is quite natural to investigate on a microscale the dynamics behind these river floods in order to comprehend the conditions leading to the floods and the possible interplay between the fundamental flood drivers. If not stated otherwise, the datasets used in this dissertation will always consist of daily observations. The decimal places of the data will align with the precision specified by the instrument used to record the data.

In Germany, Deutscher Wetterdienst (DWD) [162] is responsible for monitoring and providing meteorological data. DWD was founded in 1952 and it manages around 182 primary weather stations scattered throughout Germany. The R package 'rdwd' includes a variety of functions that act as an interface to extract and read data from the Düsseldorf station (ID: 1078) through an FTP setup. Subsequently, the plots shown in Figures B.6 and B.7 were obtained to provide visual insight into the average weather conditions from the extracted Düsseldorf station data over the available period $1970 - 2021$. Upon first glimpse at the figures, it can be inferred that the weather in the winter season is typically cold but not freezing, windy, and mostly cloudy. The direction of the wind is the main determining factor: if winds coming from west pass

through, the temperature is mild, but if cold air masses from Russia passes through Düsseldorf, then the temperature drops drastically. In addition, snowfalls are quite common but not particularly substantial. However, the snow cover can accumulate and last for several days if a cold spell occurs.

In contrast, the spring season is unstable, as the southwesterly current can impose mild warm periods while the northwesterly current can impose mild cold periods. The summer season is typically partly cloudy with warm temperatures but with the highest amount of precipitation as opposed to other seasons. Although there is still plenty of rainfall, the autumn season starts pleasantly but progressively becomes colder, windier, and wetter. Throughout the year, the temperature typically varies between 3 °C to 20 °C with plentiful precipitation influenced by the North Sea climate. The region experiences constant calm winds, but they are notably higher from November to March. Also, there is very little sunshine during the winter months, especially from November to February. Hence, based on the Köppen-Geiger system, the climate of Düsseldorf is classified as an oceanic climate (Cfb) [99].

In general, when considering the topography of Düsseldorf, the floods occur when the river's discharge overflow the channel's capacity, as can be seen from Figure 2.1. In Germany, there are two main providers of this hydrological data. In 1988, the German Federal Institute of Hydrology (BfG) had established the Global Data Runoff Center (GRDC) [21] from which river discharge data is provided. The Federal Waterways and Shipping Administration (WSV) [17] provides the gauges water levels at the river Rhine. Traditionally, as the surface runoff arising from precipitation is considered to be the main mechanism that causes a significant rise in the streamflow, the assessors of river flood risk were only concerned about the heavy load that the extreme precipitation is able to exert on the river banks. In fact, if there is a prolonged rainfall, the ground will become saturated, causing the rainwater to enter the river stream more rapidly, which will increase the discharge that can result in a sharp rise in flood risk. Similarly, the rainwater from a sudden burst of heavy rainfall will not be immediately retained by the ground and instead it will enter the river streamflow, which may result in a flood.

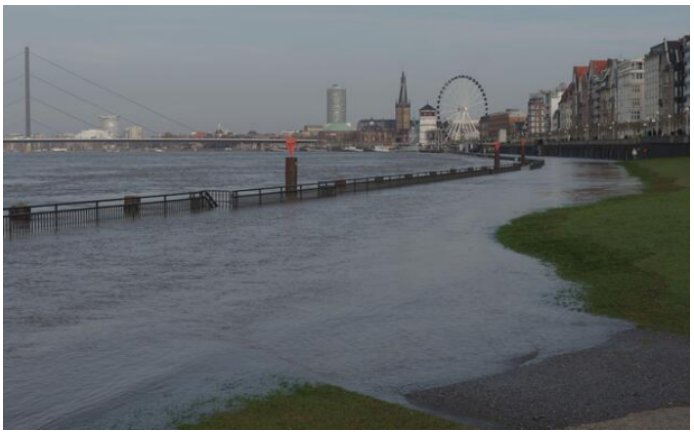


Figure 2.1: Rhine river in Düsseldorf under flooding conditions.

It should be emphasised that precipitation on existing snowpack will accelerate the melting of snow, which will augment to the precipitation entering the streamflow. This will contribute to a higher impact, especially in spring [99, 114]. The influence of snowmelt hazard is not usually accounted for, as several factors can influence the melting process, such as latitude, air and soil temperature, and wind speed, which makes it complicated to deal with. In addition, the monitoring and accurate assessment of evapotranspiration is essential to precisely determine the full consequences of the snowmelt hazard and precipitation. The conventional approach to determine the evapotranspiration observations is by the Gravimetric-based methods [158]. However, these methods use pin-point measurements. Variations in local terrain and vegetation properties cause these methods to be labour intensive and expensive to carry out. A widely used approach is to estimate the evapotranspiration by models that take into account the water dynamics and plants ecosystem by exploiting the hydrometeorological data.

Indeed, for this vital task, the DWD has created the agrometeorological calculation of current evaporation (AMBAV) model, which calculates the potential evapotranspiration. This model uses the classic Penman-Monteith equation (Definition $A.13$) to simulate the water balance in the soil system and the Richards equation (Definition $A.14$) to simulate the soil water dynamics [52]. Furthermore, the soil water and hydraulic conductivity properties are represented by the pedotransfer functions, where

the coefficients have been recalibrated for the field capacities and the wilting points of the German soil. This model is particular as it considers the phenological development of the plant and the impact of the drying soil on the water consumption of the plant.

The data calculated through this model is, however, limited as it starts at 1991. Nevertheless, daily potential evapotranspiration data recorded during the period $1951-2010$ can be obtained from the UFZ portal [118]. Even though this data is also based on observations extracted from the DWD Düsseldorf station, the potential evapotranspiration is estimated using the Hargreaves-Samani approach. The Hargreaves-Samani model is an empirical method that has been used in view of its simplicity and low data demand, as only temperature and solar radiation data are required. The procedure of how the data is obtained is detailed in [167]. As the data generated by the AMBAV model was considered to be more reliable, it is used as the main source, and where necessary, it is complemented with the data obtained from the UFZ portal.

2.2 Quantitative Analysis on Collected Data

To recapitulate, the focus of the case study of this dissertation is to apply the probabilistic models derived to analyse flood risk by taking into account the discharge of the river Rhine at Düsseldorf station and the corresponding combination of the precipitation, snow melted water equivalent, and potential evapotranspiration, as displayed in Figure 1.2. Accordingly, the following table encompasses all the information with regards to the data at our disposal, that will be used throughout this dissertation. Variables such as wind velocity and temperature will be used in the calculation of the snow water equivalent. More information will be given in Section 2.3.1.

From Table 2.1, it can be seen that the common time period of the available time series is from 1969 to 2019. However, as the relationship between the river discharge and the water level can be represented by a hydrological rating curve, the river discharge data from 2020 to 2021 can be estimated using the following relationship:

$$\mathbb{E}(RD \mid WL = wl) = C_r(wl + \alpha)^{\beta}, \tag{2.1}$$

Variable	Acronym	Unit	Source	Period	Measuring Device/Method
Precipitation	P	mm	Düsseldorf station	01/07/1969 - 31/12/2021	Hellmann and Precipitation sensor according to Joss-Tognini
Total Snow Depth	TSD	cm	Düsseldorf station	01/07/1969 - 31/12/2021	Snow ruler
New Snow Depth	NSD	cm	Düsseldorf station	01/07/1969 - 31/12/2021	Snow ruler
Temperature	T	°C	Düsseldorf station	01/07/1969 - 31/12/2021	PT 100 (air)
Wind Velocity	WV	m/s	Düsseldorf station	01/01/1952 - 31/12/2021	Windsensor Classic 4.3303
Potential Evapotranspiration	PE_{AMBAV}	mm	Düsseldorf station	01/01/1991 - 31/12/2021	AMBAV method
Potential Evapotranspiration	PE_{HS}	mm	UFZ portal	01/01/1951 - 31/12/2010	Hargreaves-Samani method
River Discharge	RD	m^3/s	Rhine River gauge	01/11/1900 - 31/12/2019	Rating curve conversion
Water Level	WL	cm	Rhine River gauge	01/11/1872 - 31/12/2021	Staff gage and water level sensor

Table 2.1: Summary of the available data.

where C_r and β denote the rating curve constants and α is a datum correction, that needs to be estimated [142]. The C_r constant is influenced by the characteristics of the river channel, α signifies the water level when there is no flow, and β is an exponent that outlines the slope of the hydrological rating curve. The relationship holds only if the data is in the correct standard units. So, in what follows, the WL observations will be converted to meters. Table 2.1 shows that the time period for the water level measurements started at 1872. However, a considerable amount of measurements were missing before 1930. To avoid any possibility of bias from using interpolated data, we decided to consider data from the time period 1930 to 2021.

Figure 2.2 displays the rating curve, where each color represents a group of 20 years, as indicated by the legend shown in the plot. From Figure 2.2, it can be deduced that the WL-RD relationship changes with time, a phenomenon known as shifting control. This movement can be caused by river erosion, water management, or any type of sediment deposition. Hence, to properly estimate the river discharge data from 2020 to 2021, the observations from the latest 10 years were considered. Non-linear least squares regression was used and the corresponding parameter estimates together with the estimated standard errors are shown in Table 2.2. The t-test for each parameter reveals a significant deviation from zero. The fit of the resulting rating curve is demonstrated in Figure 2.3. It is evident that the distance between the curve and the data points appears to be negligible, leading to the conclusion that the fit is suitable.

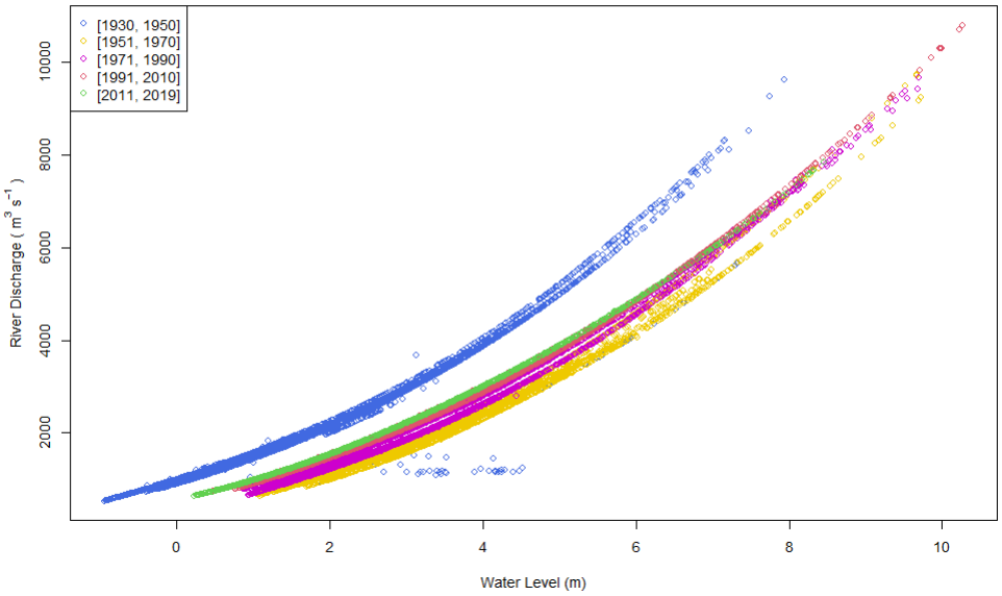


Figure 2.2: Düsseldorf hydrological rating curve.

Parameters	C_r	α	β
Estimates	0.02	262.10	1.85
Standard Error	< 0.01	0.32	< 0.01

Table 2.2: Rating curve parameter and standard errors estimates.

The resulting rating curve (2.1) was thus then used to predict the RD data from 2020 to 2021 and these estimated observations are shown by the red dots in Figure 2.4. Consequently, the time period considered for this dissertation is from 1969 to 2021. Following the discussion on the time period to be taken into consideration, the next important step in any statistical analysis is to become more familiar with the dataset, as it will assist in the model analysis. This can be achieved by exploring characteristics that can be observed from descriptive statistics and plots. It is also important to get an idea about the quality of the data used. At this stage, we assume that the time series exhibit stationarity. Nonetheless, we will examine this assumption in detail in Section 2.3.2. Recall that DWD is the authority responsible for monitoring and providing meteorological data. Uncertainties in the data mainly arise from changes in meteorological measuring equipment, different observation reading time, and data

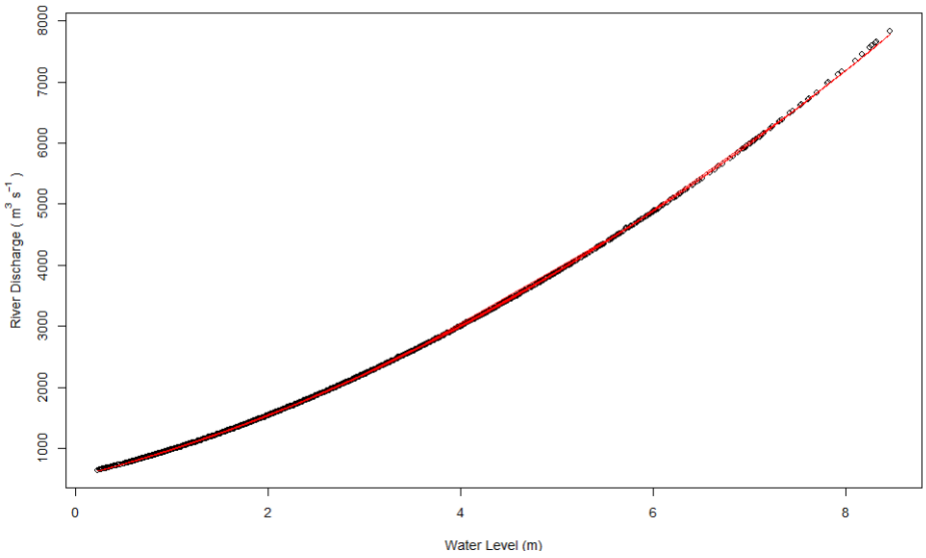


Figure 2.3: Non-linear rating curve fit on the data for the period $2011 - 2019$.

transfer errors. DWD addresses these uncertainties by following guidelines set by the World Meteorological Organization and the International Civil Aviation Organization. In fact, according to their portal, the data collected by any German station is verified by an interactive computer-aided monitoring system that ensures successful quality assurance. The quality assurance process consists of four sequential steps, each assigning a quality flag. According to the DWD portal, the variables mentioned in Table 2.3 had the best quality check level. For more detail, refer to [91]. Time series plots for the five variables considered are shown in Figure B.1 to Figure B.5 in Appendix B, and the following are the corresponding summary statistics.

Statistics / Variable	Minimum	Mean	Maximum	Standard Deviation	Median
Precipitation (mm)	0.00	2.07	57.40	4.17	0.10
Total Snow Depth (cm)	0.00	0.15	27.00	1.16	0.00
New Snow Depth (cm)	0.00	0.04	15.00	0.44	0.00
Temperature (°C)	−14.60	10.81	31.20	6.79	10.90
Wind Velocity (m/s)	0.20	3.93	14.40	1.81	3.70

Table 2.3: Summary statistics of the DWD data.

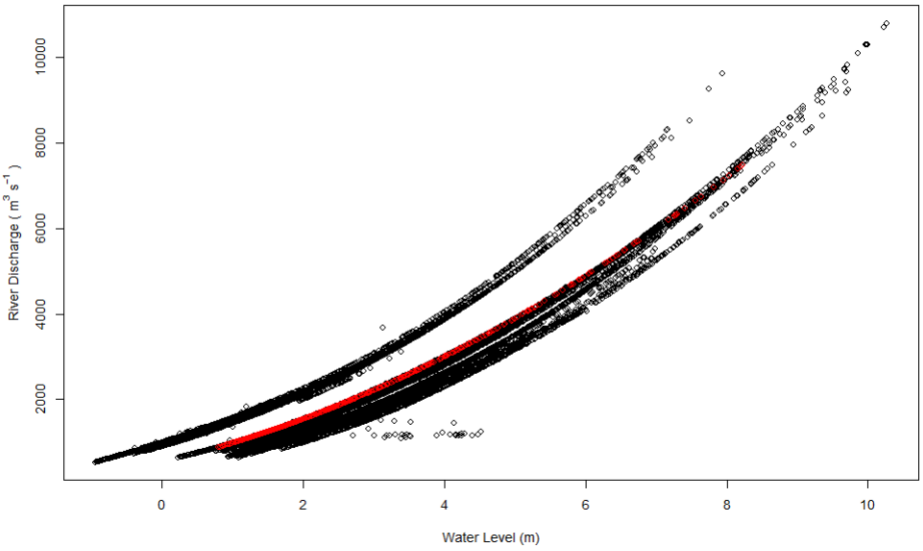


Figure 2.4: The estimated river discharge data from $2020 - 2021$ in red circles.

As mentioned earlier, the potential evapotranspiration estimates (PE_{AMBAV} and PE_{HS}) used in this dissertation are extracted from two separate sources, which are calculated by different approaches: the Hargreaves-Samani approach and the AMBAV approach. The accuracy of PE_{AMBAV} and PE_{HS} estimates relies on the attributes of the respective model but, more importantly, on the quality of the data that is utilized by the models. Both models were implemented on DWD data that holds the best quality check level. The complete time series for evapotranspiration is shown in Figure B.8 and the corresponding summary statistics are provided in the table below.

Variable \ Statistics	Minimum	Mean	Maximum	Standard Deviation	Median
$PE = PE_{AMBAV} + PE_{HS}$ (mm)	0.00	1.98	9.30	1.60	1.55
PE_{AMBAV} (mm)	0.00	1.93	9.30	1.57	1.50
PE_{HS} (mm)	0.07	2.05	8.16	1.63	1.64

Table 2.4: Summary statistics of the potential evapotranspiration data.

The values in Table 2.4 and the comparison of the datasets over an overlapping time period that is shown in Figure B.9, suggest that the two approaches share similar characteristics. In fact, the histograms depicted in Figure 2.5 exhibit similar shape and

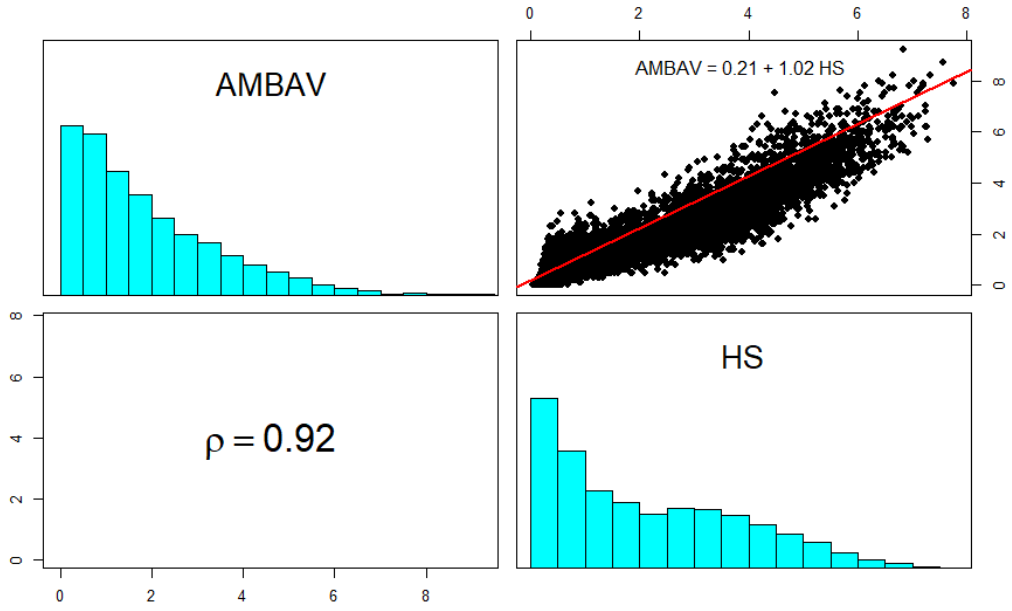


Figure 2.5: AMBAV and HS scatter and histogram plots for the two datasets on the overlapping time period.

tail decay. It can be noted that there is a linear relationship between the two datasets, as demonstrated by the red line representing the best fit with an R^2 of 0.85. The gradient of the regression line is almost one, and the Pearson's correlation coefficient is 0.92 (p-value < 0.01). There is indeed some scattering around the line of best fit, but such behaviour is expected since the two models are based on different principles. Additionally, the evapotranspiration values quoted by each model are averages over significantly different terrains; as such, the original values already have significant uncertainties. Taking this into account, along with the need to have a time series that is as long as possible for more reliable extreme value analysis, using the data sourced from [162] as an extension to the data sourced from [118] for the evapotranspiration, is considered reasonable. After comparing the results of the models presented in Chapters 3 and 4, which were applied to the data with and without the inclusion of the evapotranspiration observations, it was determined that the impact was relatively small for that would not make a difference in a flood scenario. Hence, any difference in the values given by the two models can be expected to have a negligible effect on the final

results. It is important to emphasize that, notwithstanding the fact that the effect of evapotranspiration can be small, it is still important to consider it in the model for the sake of completeness and to be as comprehensive as possible.

Finally, let us consider the last two variables in Table 2.1. As previously mentioned, the water level and river discharge observations are provided by the BfG and WSV authorities. According to the PEGELONLINE portal and the Global Data Runoff Center, the authorities apply plausibility checks to identify and eliminate errors. Therefore, the data obtained is assumed to be checked, and the corresponding time series plots are shown in Figures B.10 and B.11. The corresponding summary statistics are provided in Table 2.5.

Variable \ Statistics	Minimum	Mean	Maximum	Standard Deviation	Median
River Discharge (m^3/s)	641	2153	10800	1123.74	1860
Water Level (m)	0.23	2.94	10.27	1.37	2.68

Table 2.5: Summary statistics of the collected river data.

2.3 Pre-processing the Data

In this section, four further pre-processing steps will be outlined, with the aim to create the final dataset that will be used in the case study investigations. Data summary and assumption checks that are required by the theoretical models that will be covered in Chapters 3 and 4 will also be made.

2.3.1 Reconstruction

As stated in Sections 1.2 and 2.2, the two main variables selected to describe the flood regime in Düsseldorf are the discharge of the Rhine River and the collective impact (CI) of the snow melted water equivalent (SMWE) and precipitation excess. The river discharge values were already available, but the collective impact time series, denoted by T_{CI}, had to be constructed for the purpose of this dissertation. In [11], precipitation

excess is defined to be the arithmetic difference between precipitation and potential evapotranspiration, which is assumed to have a lower bound of zero. Meanwhile, the SMWE component refers to the amount of water content that is released from the melted snow. The daily SMWE observations had to be calculated. Since, the mass of a sample of snow remains constant after the snow is converted to water, the SMWE observations can be expressed through the definition of density. Let ρ and V denote the density and volume of the sample in the physical state indicated by the subscript, then it follows that

$$\rho_{\text{water}} V_{\text{water}} = \rho_{\text{snow}} V_{\text{snow}}. \tag{2.2}$$

Given that $\rho_{\text{water}} = 1000\,\text{kg/m}^3$, then equation (2.2) can be rearranged as follows

$$V_{\text{water}} = V_{\text{snow}} \cdot \frac{\rho_{\text{snow}}}{1000}. \tag{2.3}$$

Furthermore, the volume of the sample in each respective state can be expressed as the base area times the depth of the sample in the corresponding state. However, as the base area remains the same for each state, the expression (2.3) can be further simplified as follows

$$SMWE = SD \cdot \frac{\rho_{\text{snow}}}{1000}, \tag{2.4}$$

where SD (snow depth) represents the depth of the sample in each state. The snow depth in equation (2.4) can be obtained by observing the interplay between the total snow depth (TSD) and the new snow depth (NSD) as follows

$$SD_i = TSD_{i-1} + NSD_i - TSD_i, \tag{2.5}$$

where i represents an arbitrary day and $i-1$ represents the previous day. The density of the snow in equation (2.4) can be estimated using the following formula

$$\rho_{\text{snow},i} = \left(90 + 130\sqrt{TSD_i}\right) \cdot \left(1.5 + 0.17 \times \sqrt[3]{T_i}\right) \cdot \left(1 + 0.1\sqrt{WV_i}\right), \tag{2.6}$$

where T represents the temperature and WV denotes the wind velocity [109]. Consequently, considering the above substitutions and the calculation of the precipitation excess, the target time series T_{CI}, for the collective impact was obtained. Time series for the river discharge, T_{RD}, and collective impact, T_{CI}, are shown in Figures 2.6 and 2.7. The corresponding summary statistics are provided in the Table 2.6.

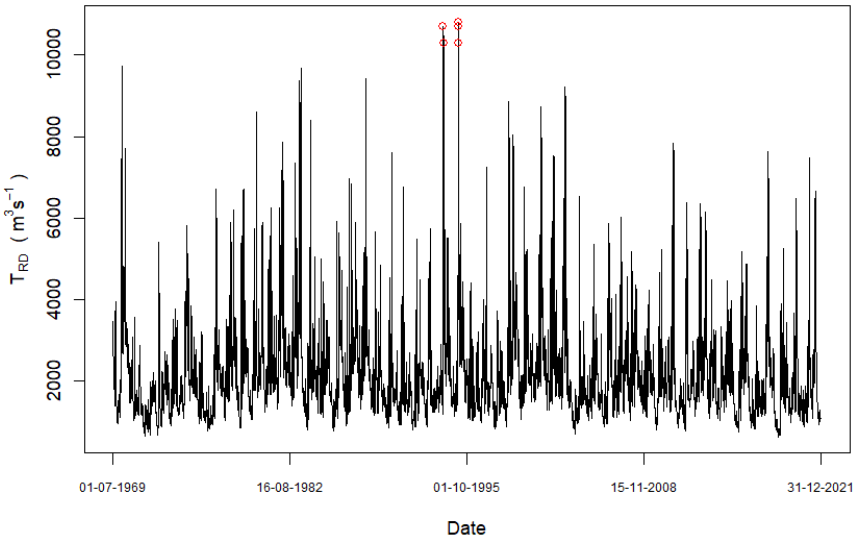


Figure 2.6: The daily river discharge time series with the five most extreme observations marked by the red circles.

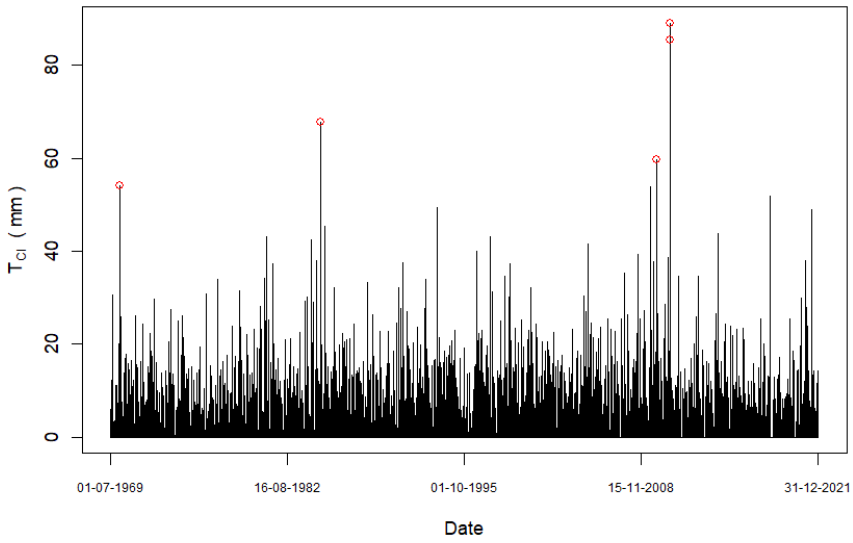


Figure 2.7: The daily collective impact time series with the five most extreme observations marked by the red circles.

Time Series \ Statistics	Minimum	Mean	Maximum	Standard Deviation	Median
T_{RD} (m^3/s)	641	2153	10800	1123	1860
T_{CI} (mm)	0.00	1.66	89.03	4.17	0.00

Table 2.6: Summary statistics of the time series T_{RD} and T_{CI}.

The red points in Figure 2.6 depict the first five most extreme observations, which are listed in panel (a) of Table 2.7 along with their corresponding dates and T_{CI} values. It's important to note that additional extreme observations could have been included if desired. The dates correspond to events of severe floods that caused significant damages along the stretch of the river Rhine. In fact, these events have led to some of the improvements that were later implemented in flood disaster management and preparedness policies [4]. This exercise could also be replicated for the T_{CI} time series, as shown in Figure 2.7 and in panel (b) of Table 2.7. Unlike the events listed in panel (a), the events in panel (b) did not lead to floods. The T_{RD} values in panel (b) of Table 2.7 are in fact only slightly above average which means that the flood mitigation measures could handle the extreme runoff that is attributed to T_{CI}.

Date \ Time Series	T_{RD}	T_{CI}
31/01/1995	10800.00	0.00
24/12/1993	10700.00	7.70
30/01/1995	10700.00	0.00
29/01/1995	10300.00	12.70
25/12/1993	10300.00	0.00

(a)

Date \ Time Series	T_{CI}	T_{RD}
27/12/2010	89.03	4480.00
22/12/2010	85.43	2730.00
22/01/1985	67.76	1170.00
22/12/2009	59.73	1630.00
18/02/1970	54.20	4020.00

(b)

Table 2.7: The five highest observations of the time series T_{RD} shown in the left panel (a) and the five highest observations of the time series T_{CI} shown in the right panel (b).

To gain a more comprehensive understanding of the behavioral changes in T_{RD} and T_{CI} over time, it would be beneficial to examine a hydrograph. A hydrograph is a chart that demonstrates the variation in the river discharge over time as well as the variation in collective impact over the same time period. According to [40], the 1995

flood was caused by the accumulation of rain and the gradual melting of snow from previous days. In fact, the delayed impact of T_{CI} is demonstrated in the corresponding hydrograph shown in Figure B.12 and panel (a) of Table 2.7. The delay in the impact of T_{CI} on T_{RD} is also shown for the 1993 floods, as shown by the hydrograph shown in Figure B.13. An estimate of this delay can be obtained by observing the lag time between the peak discharge and the peak of the combined impact of snow melted water equivalent and precipitation excess. To be able to do so, the lag time for the first 800 ordered river discharge values were considered, and from the average of these lag times it was concluded that on average the lag time is of 7 days. Hence, as a second pre-processing step, the delay effect is introduced in the new time series T_{RS} by applying a rolling sum (RS) on T_{CI} with a window of 7 days. This time series is presented in Figure B.14. The corresponding summary statistics are presented in Table 2.8.

Time Series \ Statistics	Minimum	Mean	Maximum	Standard Deviation	Median
T_{RS} (mm)	0.00	11.60	296.74	14.93	6.95

Table 2.8: Summary statistics of the time series T_{RS}.

An important aspect to consider when investigating the joint behaviour of the target variables is to assess the dependence structure, especially at the extreme ends. As the variables under consideration were specifically selected to describe the flood regime, the Pearson's correlation coefficient, the Kendall's tau coefficient, and the upper tail dependence measure (the latter is described in detail in Chapter 4) are used to get an estimate of the degree of the general and extreme dependence between T_{RD} and T_{RS}. From the matrix of plots displayed in Figure 2.8, it can be deduced that there is a positive dependence both in the general structure of the variables and at the tails.

However, this type of structure is regarded as relatively weak as the measures are close to zero. Motivated by these measures, it was found that if the resolution of the time series T_{RD} and T_{RS} is down-scaled from a daily time scale to a monthly time scale, the dependence structure rank is improved from relatively weak to a medium dependence structure. Subsequently, the third pre-processing step was to find the monthly mean

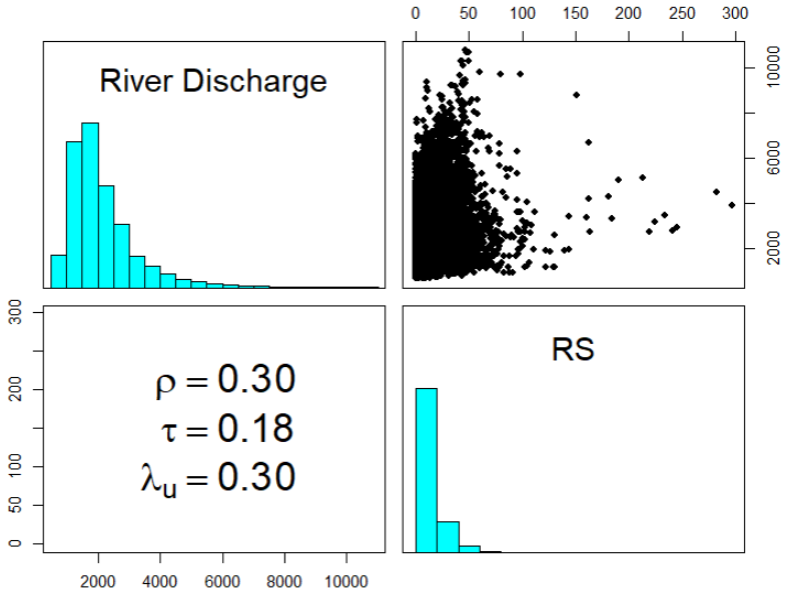


Figure 2.8: The scatter and histogram plots of the RD and RS data. The lower panel shows the values of the dependence metrics covered in Section 4.1.2.

of the T_{RS} time series such that the monthly rolling sum time series T_{MRS} was created. Similarly, the monthly mean was found for the T_{RD} time series such that the monthly river discharge time series T_{MRD} was constructed. Despite the improvement is not that substantial, the new dependence structure is preferable, as can be viewed from the matrix of plots displayed in Figure 2.9. In addition, as shown in Figures B.15 (a) and B.15 (b), the strength of the dependence measures are at their peak while maintaining significance, significance shown by the black dot, when no lags are considered between the variables. Hence, the final target time series upon which the case study will focuses on, are shown in Figures 2.11 and 2.10 respectively, with the corresponding summary statistics shown in Table 2.9.

Statistics / Time Series	Minimum	Mean	Maximum	Standard Deviation	Median
T_{MRD} (m^3/s)	727.70	2155.90	5869.60	902.25	1949.00
T_{MRS} (mm)	0.00	11.60	95.76	8.66	10.14

Table 2.9: Summary statistics of the time series T_{MRD} and T_{MRS}.

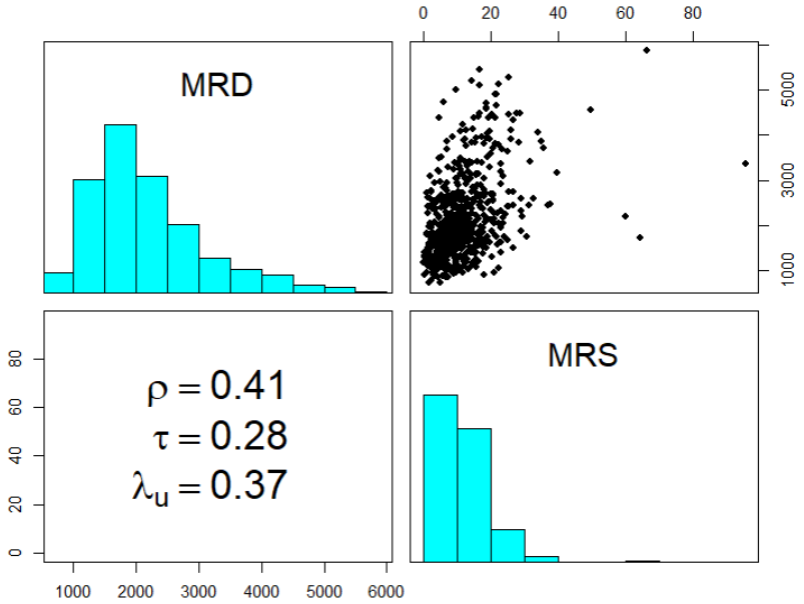


Figure 2.9: The scatter and histogram plots of the MRD and MRS data. The lower panel shows the values of the dependence metrics covered in Section 4.1.2.

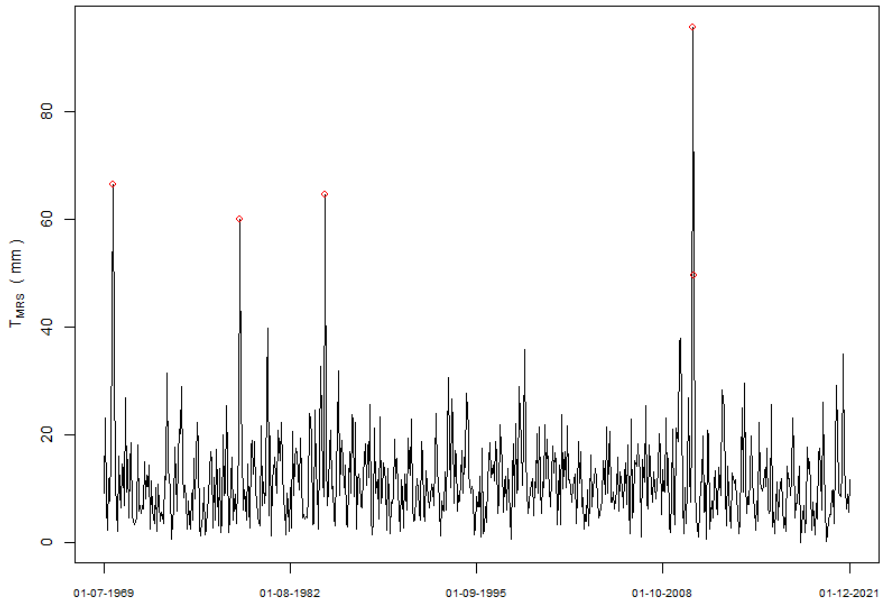


Figure 2.10: The monthly rolling sum time series with the five most extreme observations marked by the red circles.

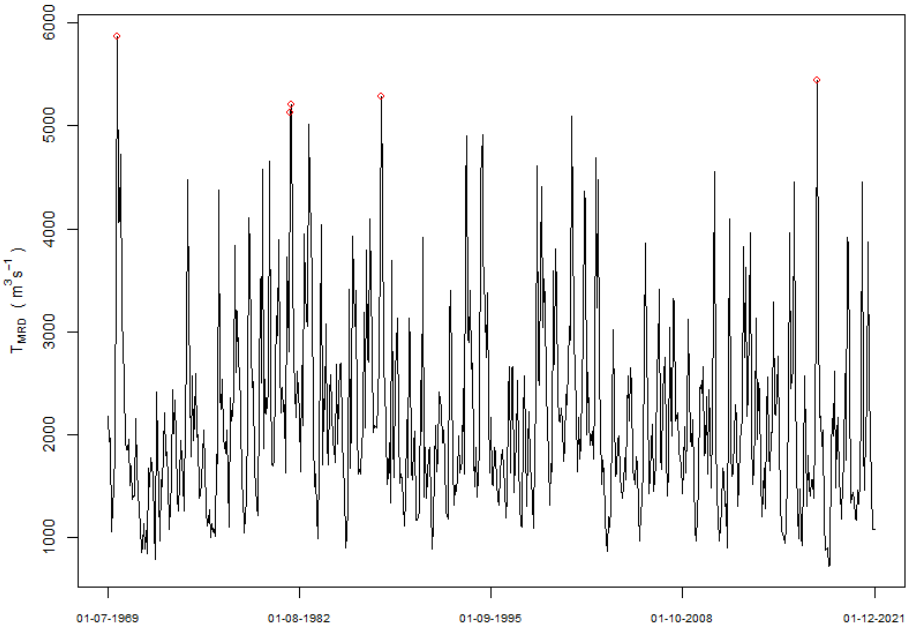


Figure 2.11: The monthly river discharge time series with the five most extreme observations marked by the red circles.

The red points in Figures 2.11 and 2.10 represent the five most extreme observations for T_{MRD} and T_{MRS}, respectively, which are listed in panels (a) and (b) of Table 2.10. In contrast to Table 2.7, the panels in this table contain a common event that corresponds to an actual flood that happened during the corresponding period. In [4], it is stated that floods frequently occur during the winter and early spring in the lower Rhine region, where Düsseldorf is located. Although this is already partially underlined in Table 2.10, 88.68% of the annual maxima occurred between November and March. The reason is that the winter season in Düsseldorf is characterized by a substantial amount of precipitation and river discharge, as well as low temperatures and few sunshine hours, as can be seen in Figures B.6, B.7 and B.16.

Oftentimes, the precipitation that falls in late autumn and winter accumulates as snow. This snow then melts and drains in the following spring and summer. In fact, according to [40], the flow regime of the Rhine river in the summer months is governed by snowmelt and precipitation runoff from the ALPS. During the winter months, it is governed by

Date \ Time Series	T_{MRD}	T_{MRS}
02/1970	**5869.64**	**66.39**
01/2018	5438.39	16.67
03/1988	5278.07	25.53
01/1982	5205.48	14.41
12/1981	5129.56	22.23

(a)

Date \ Time Series	T_{MRS}	T_{MRD}
12/2010	95.76	3372.26
02/1970	**66.39**	**5869.64**
01/1985	64.60	1736.13
01/1979	60.05	2202.26
01/2011	49.67	4550.00

(b)

Table 2.10: The five highest observations of the time series T_{MRD} shown in the left panel (a) and of the time series T_{MRS} shown in the right panel (b).

the precipitation runoff from the uplands. For a comprehensive analysis, the cause-and-effect dynamics of floods in a water cycle need to be within a common time frame. Thus, it is more important to opt for the hydrological year (water year) rather than the normal calendar year. The selection would ensure the highest correlation between the river discharge and the combination of precipitation excess and snowmelt. The water year does not have a standard format, but in general, it constitutes of a 12-month period that begins from the month of the lowest average river discharge [160]. From Figure B.16, it can be deduced that the month with the lowest river discharge throughout the whole study period is October. So, for the rest of this dissertation, a year would refer to the period starting from 1 October of one year and ending on 30 September of the subsequent year. Thus, the fourth pre-processing step consists of rearranging the time horizon of the time series T_{MRD} and T_{MRS} to start on October 1, 1969 and end on September 30, 2020. The flowchart shown in Figure 2.12 details a summary of the pre-processing steps performed.

2.3.2 Motivation for Partitioning the Time Horizon

A recurrent requirement for all models presented in the following chapters is that the underlying observations are independently and identically distributed (IID). In this section, the two target time series will be analyzed to confirm that this assumption is met. In order to ensure that the observations in each time series are identically

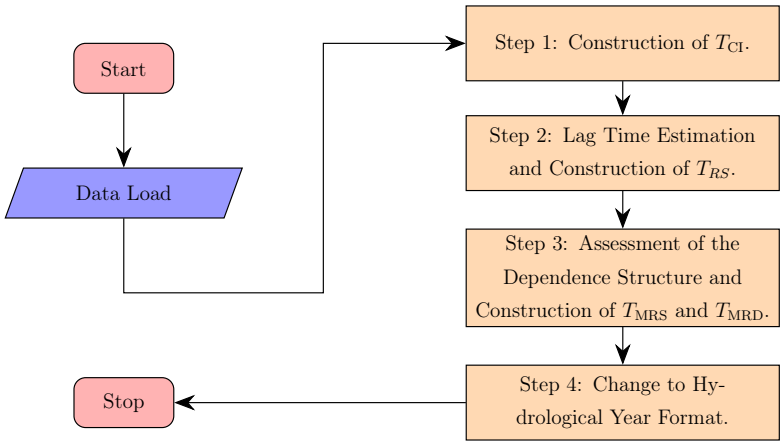


Figure 2.12: Flowchart of the data treatment process.

distributed, it is necessary for both time series to be stationary. The appropriate manner to assess the stationarity of a time series is by means of hypothesis testing. There are various methods available, but in this dissertation the Phillips-Perron (PP) test will be used as it is robust against heteroscedasticity and performs relatively well for sample sizes greater than 45 observations [1]. The hypotheses for the PP test are:

H_0: the time series contains a unit root;

H_1: the time series is trend stationary.

Table 2.11 shows that there is enough evidence to reject the null hypothesis for both series, so we can deem both T_{MRD} and T_{MRS} to be trend stationary.

	Statistic Value	P-Value
MRD	-268.32	< 0.01
MRS	-429.55	< 0.01

Table 2.11: Phillips-Perron test results.

Figures 2.10 and 2.11 show that the two target time series do not exhibit any visible trend. Testing whether a trend is present in the target time series is also carried out.

As it appears that there may be a hint of a seasonal cycle indicated in Figures B.21 and B.22, the extended seasonal version of the Mann-Kendall test will be considered as it is highly robust, comparatively powerful, and is frequently the suggested approach in such instances [153]. Hence, the hypotheses for the Seasonal Mann-Kendall test are

H_0: there is no monotone trend in the data,

H_1: there is a monotone trend in the data.

From Table 2.12, it can be remarked that the tests found that there is no evidence to reject the null hypothesis for both time series, so both time series T_{MRD} and T_{MRS} can be deemed not containing a significant trend.

	Statistic Value	P-Value
MRD	-0.03	0.28
MRS	< 0	0.86

Table 2.12: Seasonal Mann-Kendall test results.

For both time series, however, a seasonal cycle is indicated in Figures B.21 and B.22. To further study this pattern, techniques such as the autocorrelation function (ACF) and the periodogram plot can be used to investigate the seasonal pattern for each time series. In addition, techniques such as the cross correlation function (CCF) and the coherence spectrum plot can also shed some light on the inter-relationship between the two time series. The correlogram plots for both time series are presented on the major diagonal of the matrix plot shown in Figure B.17. These plots reveal a sinusoidal pattern that repeats every 12 lags (each lag representing one month), indicating a seasonal component with a period of 12 months. This observation is consistent with the highest peak found in the periodogram plots shown in Figures B.18 and B.19, where each peak occurred at the frequency of 0.08, approximately equal to 1/12. Moreover, the cross-correlation function plots presented on the minor diagonal of the matrix plot in Figure B.17 suggest that both time series contain seasonal components due to the sinusoidal cycle pattern. The plots also show that the time series are most synchronized when the

lag is zero, indicating that neither time series leads the other. In fact, the frequency obtained from the periodogram plots is consistent between the two time series, as it is the highest peak in the coherence spectrum plot shown in Figure B.20. However, the seasonal property is not in line with the identically distributed assumption that will be considered by the models to be covered in the subsequent chapters. To address this issue, the data is split into two distinct time horizons:

P1: the first time horizon contains observations from years $1969 - 2020$, from October till March;

P2: the second time horizon contains observations from years $1969 - 2020$, from April till September.

The idea is to split the hydrological year in half. The summary statistics for T_{MRD} and T_{MRS} for both time horizons are presented in Table 2.13.

	Statistics / Variable	Minimum	Mean	Maximum	Standard Deviation	Median
P1	T_{MRD}	727.70	2301.60	5869.60	1029.41	2068.20
	T_{MRS}	0.49	14.18	95.76	10.04	12.35
P2	T_{MRD}	868.90	2021.70	5011.30	732.15	1874.10
	T_{MRS}	0.00	9.03	34.98	6.07	7.69

Table 2.13: Summary statistics of T_{MRD} and T_{MRS} time series for each time horizon.

Performing the same two tests that were done earlier in this section on the two distinct time horizons for each time series resulted in the same conclusions as those obtained for the full dataset. Additionally, when the seasonal pattern analysis was re-performed, Figures B.23 - B.30 were produced. From these figures, it can be seen that by splitting the full time horizon into two, the high seasonality component is less evident. Furthermore, the ACF plots show that there is still some temporal dependence between the observations. In this dissertation, we used the Durbin-Watson test with the following hypotheses to test if the dependence between observations in both time series for the two time horizons is significant:

H_0: there is no autocorrelation in the time series,

H_1: there is autocorrelation in the time series.

Table 2.14 shows that only in the case of T_{MRS} during the P2 time horizon, there is not enough evidence to reject the null hypothesis. Therefore, for the other three instances, we can conclude that there is temporal dependence between the observations.

	Time Horizon	Statistic Value	P-Value
T_{MRD}	P1	1.24	< 0.01
	P2	0.98	< 0.01
T_{MRS}	P1	1.54	< 0.01
	P2	1.82	0.11

Table 2.14: Durbin-Watson test results.

The results presented in Table 2.14 are commonly observed in practice, and the usual approach when applying models based on EVT is to classify an event as extreme if the time elapsed between that event and the previous extreme event is more than a pre-specified minimum separation interval. Typically, this minimum separation interval would be taken as a month, which is the case in our study. Based on this analysis, it can be deemed that the IID assumption is reasonable for both time series over the two time horizons. As a result, the analysis in Chapters 3 and 4 is conducted separately using the time series for the P1 and P2 time horizons.

Chapter 3

Univariate Analysis

Univariate extreme value theory (UEVT) is a branch of statistics which provides techniques that evaluate the statistical behavior of univariate extreme observations. The principal objective of this chapter is to introduce the methodology behind UEVT's fundamental approaches, namely the Block Maxima (BM) method outlined in Section 3.1, the K Largest Order Statistic (KLOS) method outlined in Section 3.2, and the Peaks Over Threshold (POT) method outlined in Section 3.3. These methods rely on asymptotic results that assume that the observations are independent and identically distributed (IID). These models are more practical than empirical methods for preliminary analysis.

3.1 Block Maxima Method

The focus of this section is to present the theoretical results produced by Fisher and Tippett [47], Gnedenko [58] and the corresponding results that will serve as basis for the introduction of the first fundamental method of UEVT, that is known as the Block Maxima. The starting point will be the development of the characterization of the class of possible limiting distributions of the normalized sample maxima, which is parallel to the theoretical development of the Central Limit Theorem (CLT). The BM approach and supporting techniques are then established in Sections 3.1.1 and 3.1.2. These are followed by the application to the Düsseldorf case study in Section 3.1.3.

3.1.1 Modelling the Largest Order Statistic

Although the classical UEVT is structured using similar concepts as for the Central Limit Theorem, extreme analysis is interested in the characterization of observations found at the tail of the distribution. Therefore, the statistical behaviour of the asymptotic distribution of the sample maximum or sample minima is more in line with the set objective. As the framework of the UEVT is probabilistic in nature, a sequence of univariate IID random variables $\{X_i\}_{i\in\mathbb{N}}$ with a common underlying distribution function D is considered. Given the nature of the case study, the theory presented in this dissertation will revolve around the asymptotic results that are based on the upper right tail. Let M_n represent the sample maximum over the first n observations (n^{th} order statistic), i.e.

$$M_n = \max\{X_1, X_2, \ldots, X_n\}. \tag{3.1}$$

As an initial strategy, we could attempt to obtain the exact distribution of M_n. This is achieved by considering the fact that the random variables X_i are independent with a common distribution D for all values of n

$$\mathbb{P}(M_n \leq z) = \mathbb{P}(X_1 \leq z, \ldots, X_n \leq z) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq z) = (D(z))^n. \tag{3.2}$$

However, in many instances, the underlying distribution D would be undefined, making (3.2) impractical. The first instinctive approach is to empirically estimate the underlying distribution. Nevertheless, a slight inaccuracy in the empirical estimate would result in a significant compound error of the (3.2) estimate. Hence, the preferred approach is to follow the same procedure upon which the CLT is based, that is, to attain a family of distributions that asymptotically models (3.2). Subsequently, let us define the upper limit of a distribution and then consider the asymptotic behaviour of (3.2).

Definition 1 (Upper limit of a distribution)**.** *The upper limit (also known as the right endpoint) of the support of the distribution function D represents the greatest value that a random variable generated from the distribution D can obtain, and it is defined as follows*

$$z^* = \sup\{z \mid D(z) < 1\} \in (-\infty, \infty].$$

Consequently, when the limit (as $n \to \infty$) of (3.2) is considered, it can be observed that as a result of the non-decreasing property of the distribution function D and the fact that z^* in essence is the smallest value of z such that $D(z) = 1$, then it is follows that (3.2) converges to zero for $z < z^*$ and converges to one when $z \geq z^*$. Alternatively stated,

$$\mathbb{P}(M_n \leq z) = (D(z))^n \xrightarrow{\mathbf{d}} \mathbb{I}_{\{z \geq z^*\}} \qquad \text{as } n \to \infty, \tag{3.3}$$

where $\mathbb{I}_A$ denotes the indicator function of the event A. Subsequently, it can be deduced that

$$\begin{aligned} \mathbb{P}\left(|M_n - z^*| \leq \epsilon\right) &= \mathbb{P}(z^* - \epsilon \leq M_n \leq z^* + \epsilon) \\ &= \mathbb{P}(M_n \leq z^* + \epsilon) - \mathbb{P}(M_n \leq z^* - \epsilon) \\ &= (D(z^* + \epsilon))^n - (D(z^* - \epsilon))^n \xrightarrow{\mathbf{p}} 1 \end{aligned}$$

as $n \to \infty$. In view of this and the non-decreasing property of the sample maximum sequence M_n, it can be concluded that the sample maximum sequence converges almost surely to the right endpoint i.e.

$$M_n \xrightarrow{\mathbf{a.s.}} z^* \qquad \text{as } n \to \infty.$$

Hence, from the above, it can be observed that the asymptotic distribution of the sample maximum degenerates to a point mass on z^*. Therefore, a non-degenerate asymptotic distribution is crucial to enable further study about the behaviour of the extremal observations. Fisher & Tippett [47] and Gnedenko [58] followed the same approach that was considered for the CLT. As a consequence, through Theorem 1, it was shown that for some appropriate sequence of constants $\{a_n > 0\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, the linear normalization of the sample maximum that is defined as follows

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{3.4}$$

provided the required stability to determine the three possible asymptotical non-degenerate distributions types (See Definition A.1 in Appendix A) of M_n^*, which are unified into a single distribution introduced by Jenkinson [86]. It should be remarked that if the normalization of the sample maximum is only based on linear transformations, then

there are cases of underlying distribution functions where no adequate normalizing constant sequences can be constructed such that the normalized sample maxima have a degenerate limit. Subsequently, there are other technical results which are based on other forms of normalization techniques [42]. However, in this dissertation only linear normalization is considered as it provides sufficient theory to model and acquire results on the extreme observations of the case study.

Theorem 1. *Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of IID random variables with a common distribution D and let $M_n = \max\{X_1, \ldots, X_n\}$ denote the sample maximum. If there exists sequences of constants $a_n > 0$ and b_n such that*

$$\mathbb{P}(M_n^* \leq z) = D^n(a_n z + b_n) \xrightarrow{\mathbf{d}} G(z) \qquad \text{as } n \to \infty \tag{3.5}$$

for a non-degenerate distribution function G, then G is of the same type as the GEV distribution defined as follows

$$G_\xi(z) = \begin{cases} \exp\left\{-(1+\xi z)_+^{\frac{-1}{\xi}}\right\} & \xi \neq 0, \\ \exp\{-e^{-z}\} & \xi = 0, \end{cases} \tag{3.6}$$

where $(z)_+ = \max(z, 0)$ and $\xi \in \mathbb{R}$. The GEV distribution is known as the extreme value distribution.

Proof. The proof of this theorem is included in Appendix A.1. □

The shape parameter ξ of (3.6) is referred to as the extreme value index (EVI) and the main purpose of this parameter is to measure the weight of the tail structure that the asymptotic distribution possesses. The case $\xi = 0$ corresponds to an extreme value distribution with a light tail as it has an exponential rate of decay. On the other hand, the case $\xi > 0$ refers to an extremal distribution with a heavy tail as it has a polynomial rate of decay with an infinite upper endpoint, whilst the case $\xi < 0$ refers to an extreme value distribution with a short tail as it has a finite upper limit $(-1/\xi)$. These three cases are illustrated in Figure $B.31$ in Appendix B. The statement of Theorem 1 explicitly recognizes the fact that the normalizing sequence of constants plays a significant role in the weak convergence of normalized sample maxima. Let

us assume that, for some large sample size n, the extreme value distribution is an appropriate approximation to the distribution of the normalized maxima

$$\mathbb{P}\left(M_n^* \leq z\right) \approx G_\xi(z).$$

Then, by letting $\overline{z} = a_n z + b_n$ it can be inferred that

$$\mathbb{P}\left(M_n \leq \overline{z}\right) \approx G_\xi\left(\frac{\overline{z} - b_n}{a_n}\right) = G_{\xi,a_n,b_n}(z).$$

However, in practical applications the possible normalizing sequence of constants (a_n and b_n) are difficult to determine as the underlying distribution D is unknown. Hence, the normalizing constant sequences are replaced by the scale ($\sigma > 0$) and the location ($\mu \in \mathbb{R}$) parameters giving the three-parameter GEV distribution function which follows

$$G_{\xi,\sigma,\mu}(z) = G_\xi\left(\frac{z-\mu}{\sigma}\right) = \begin{cases} \exp\left\{-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)_+^{\frac{-1}{\xi}}\right\} & \xi \neq 0, \\ \exp\left\{-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right\} & \xi = 0, \end{cases} \tag{3.7}$$

and it is defined on the adjusted support $\{z \mid 1 + \xi(z-\mu)/\sigma > 0\}$. This distribution is of paramount importance in the block maxima (BM) method. The concept of the BM approach is to divide the nm observations obtained on a variable of interest into m blocks, where each block is treated as if it was a sample of size n. Then the maximum from each block is obtained to form the set $\{M_{ni} \mid 1 \leq i \leq m\}$. As a result of Theorem 1 and the preceding arguments, the three-parameter GEV distribution is fitted to the set of maxima. This process is outlined in Figure 3.1.

The effectiveness of the BM method heavily relies on the assumption that the block size is appropriate. Implementing the BM method requires the estimation of the parameters of the GEV distribution. As a result, it is crucial to find a balance between estimation bias and variance in the estimates. Considering that if the block size is relatively large then it will result into fewer maxima leading to a larger variance. Meanwhile, if the block size is relatively small, this will result into insufficient number of observations in each block. Thus, the approximation assumption of the GEV distribution would no longer be valid leading to bias in the parameter estimates [26]. The block size is often determined based on the length of the available data. A typical choice is the length depicting one year [26].

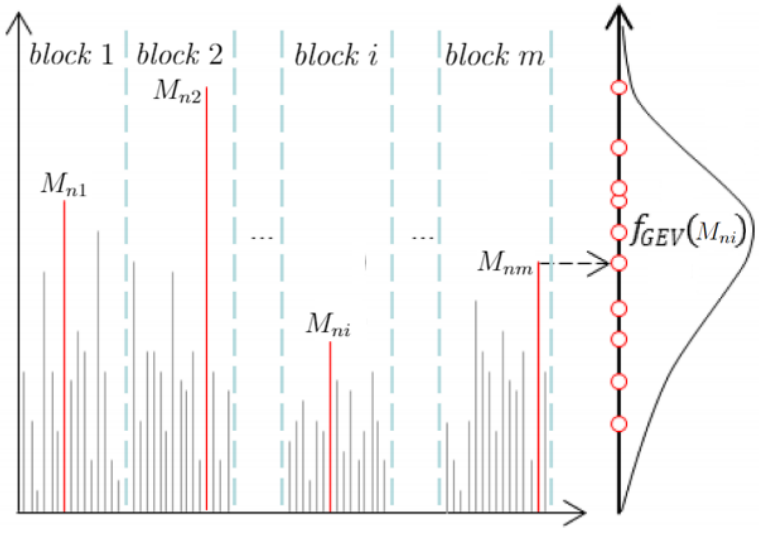


Figure 3.1: BM method where M_{ni} is the block maxima of the i^{th} block.

3.1.2 Inference on the GEV Distribution

In principle, the framework of the BM method allows us to assess and model the extreme observations through the GEV distribution. Subsequently, inferences and predictions based on the fitted extreme distribution could be obtained. Hence, in the following sections, a technique to estimate the unknown parameters of the GEV distribution is presented. A statistic based on the GEV distribution will then be presented with the aim to determine statistical properties of the next extreme observations. These techniques will also be used in the case study.

3.1.2.1 Parameter Estimation

There are several methods used in practical applications that provide estimates to the unknown parameters of the three-parameter GEV distribution. For instance, the maximum likelihood estimation method (MLE) [26], the probability weighted moments method [39] and the Bayesian method [12] are the most commonly used. Each method has its own strengths and weaknesses, but according to Lettenmaier and Burges [103], the MLE approach has an edge over the others as it is relatively simple to implement and consistently performs well. Furthermore, Coles [26] also noted that the MLE method

adapts well to changes in model structure. Additionally, given that the case study is concerned with extreme observations in the field of climate and hydrology, MLE is widely used in this field as a standard estimation technique [110, 120, 122]. Hence, in this section we will focus on the ML estimation procedure for the GEV distribution.

Let us assume that the block size (n) is adequate and as the blocks are disjoint, then the sample of maxima $\widetilde{\boldsymbol{M}} = \{M_{n1}, \ldots, M_{nm}\}$ achieved from the m blocks are IID of each other. Let the vector of parameters of the GEV distribution be denoted by $\boldsymbol{\theta} = (\xi, \sigma, \mu)'$. Then, following the general steps of the MLE procedure as demonstrated in [26], the log-likelihood function when $\xi \neq 0$ is given by

$$l(\boldsymbol{\theta}; \widetilde{\boldsymbol{M}}) = -m \ln(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln\left(1 + \xi\left(\frac{M_{ni} - \mu}{\sigma}\right)\right) - \sum_{i=1}^{m} \left(1 + \xi\left(\frac{M_{ni} - \mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}, \tag{3.8}$$

provided that $\sigma > 0$ and $1 + \xi((M_{ni} - \mu)/\sigma) > 0$ for $i = 1, \ldots, m$. On the other hand, for $\xi = 0$ the log-likelihood reduces to

$$l(\boldsymbol{\theta}; \widetilde{\boldsymbol{M}}) = -m \ln(\sigma) - \sum_{i=1}^{m} \left(\frac{M_{ni} - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp\left\{-\left(\frac{M_{ni} - \mu}{\sigma}\right)\right\}. \tag{3.9}$$

The MLE $\widehat{\boldsymbol{\theta}}$ is obtained by maximizing the equations (3.8) and (3.9) with respect to the parameter $\boldsymbol{\theta}$ over the appropriate parameter space Θ:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\, l(\boldsymbol{\theta}; \widetilde{\boldsymbol{M}}).$$

There is no explicit solution to the partial derivatives of (3.8) and (3.9) when equated to zero. In practice, the maximization can be done by non-linear optimization algorithms using variants of the Newton-Raphson algorithm [119, 120] amongst other algorithms [74]. Furthermore, the end-points of the support of the GEV distribution depend on the unknown parameters values. As a result, the typical regularity conditions are not always applicable. Smith [148] demonstrated that

(1) when $-1 < \xi < -0.5$ the MLE exists but does not have the standard asymptotic properties,

(2) when $\xi < -1$ the MLE does not exist,

(3) when $\xi > -0.5$ the MLE is regular and has all the usual asymptotic properties (asymptotic efficiency, consistency and asymptotic normality). According to Theorem $A.1$ in Appendix A:

$$\sqrt{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{\mathbf{d}} \mathcal{N}_3\left(\mathbf{0},\, \mathbb{I}_E(\boldsymbol{\theta})^{-1}\right), \tag{3.10}$$

where $\mathbb{I}_E(\boldsymbol{\theta})$ is the Fisher information matrix. In the case of the GEV distribution, this was fully derived in [120].

As previously stated, the case study in this dissertation centers on climate and hydrological observations. It is well established that such observations often result in extreme value distributions with heavy tails ($\xi > 0$), as referenced in [27]. Hence, our study aligns with scenario (3) and as a result, this does not present any of the theoretical difficulties mentioned previously.

3.1.2.2 Confidence Intervals

In practice, there are two main techniques for determining the confidence intervals of parameter estimates: the profile likelihood method and the Wald method. The Wald method is a popular choice because it is easy to apply, whereas the profile likelihood method is more intricate to apply. The Wald method is based on estimating the log likelihood function using a quadratic function, and it particularly struggles when the likelihood function is not symmetric about the MLE [107]. However, as shown in [32, p. 321], the confidence intervals generated by this method are asymptotically equivalent to those generated by the profile likelihood method. Since the log likelihood functions in equations (3.8) and (3.9) are not close to a quadratic form and the sample size considered in the case study is small, it was decided to use the profile likelihood method [31]. Suppose that the profile likelihood function for ξ is given by

$$l_p(\xi) = \max_{\sigma,\mu|\xi} \left(l\left(\boldsymbol{\theta}; \widetilde{\boldsymbol{M}}\right)\right), \tag{3.11}$$

The null hypothesis $H_0 : \xi = \xi_0$ is tested against the alternative hypothesis $H_1 : \xi \neq \xi_0$ with the profile likelihood ratio statistic given by

$$\Gamma = \frac{l_p(\xi_0)}{l_p(\widehat{\xi})},$$

which corresponds to the classical likelihood ratio statistic. According to Theorem $A.2$ in Appendix A the deviance function is

$$D_p(\xi_0) = -2\ln(\Gamma) = 2(l_p(\widehat{\xi}) - l_p(\xi_0)) \sim \chi^2_1.$$

Consequently, the null hypothesis is rejected at a significance level α if $-2\ln(\Gamma) > (\chi^2_1)^{\leftarrow}(1-\alpha)$ (see Definition A.2). Hence, the $100(1-\alpha)\%$ profile likelihood based confidence interval for ξ consists of all possible ξ_0 such that the null hypothesis is not rejected

$$C_\xi = \left\{ \xi_0 : -2\ln\left(\frac{l_p(\xi_0)}{l_p(\widehat{\xi})}\right) \leq (\chi^2_1)^{\leftarrow}(1-\alpha) \right\},$$

or equivalently

$$C_\xi = \left\{ \xi_0 : \ln(l_p(\xi_0) \geq \ln(l_p(\widehat{\xi})) - \frac{(\chi^2_1)^{\leftarrow}(1-\alpha)}{2}) \right\}.$$

In general, the profile likelihood confidence intervals are asymmetric about the estimated parameter [26]. These intervals can be visualized by plotting the curve of $\ln(l_p(\xi))$ against ξ that is known as the profile log-likelihood curve for ξ. The profile likelihood confidence intervals for the remaining parameters are derived similarly, retaining the same form but adjusted for each specific parameter.

3.1.2.3 Return Level and Return Period

Extreme value analysis is crucial in understanding the likelihood and magnitude of future extreme events. The return period and return level, derived from extreme value models, provide valuable insight on how frequent and intense these events are expected to be. These concepts are widely used in various fields, particularly in determining design standards as a means of mitigating the risk associated with extreme events. The aim of this section is to introduce these notions, which will also be used in the case study.

Suppose that p is the probability of the event $\{M_n > z_p\}$, where z_p is the corresponding quantile. Let the random variable M represent the number of periods (blocks) until the first instance that this exceedance occurs (referred to as the waiting time). Therefore, considering the following configuration

Blocks	1	2	3	$\cdots$	$m-1$	m
Exceeded	No	No	No	$\cdots$	No	Yes
Probability	$1-p$	$1-p$	$1-p$	$\cdots$	$1-p$	p

The random variable M can be modelled as a geometric random variable. Therefore, the expected value of the waiting time M is simply

$$\mathbb{E}(M) = \sum_{m=1}^{\infty} mp(1-p)^{m-1} = \frac{1}{p}.$$

As a result, for a stationary time series, it can be inferred that the return level z_p is anticipated to be exceeded on average every $1/p$ years, with $1/p$ being referred to as the return period. In general, the return level z_p is referred to as the "1 in $1/p$" year event and it is calculated through the quantile function of the GEV distribution as defined in the following definition.

Definition 2 (Return Level & Return Period)**.** *Let $G_{\boldsymbol{\theta}}$ represent the three-parameter GEV distribution function of the block maxima of size n.*

(*i*) *The return level is the $(1-p)$-quantile given by*

$$z_p = G_{\boldsymbol{\theta}}^{-1}(1-p) = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\ln(1-p)\}^{-\xi}\right] & \text{for } \xi \neq 0, \\ \mu - \sigma \ln(-\ln(1-p)) & \text{for } \xi = 0 \end{cases}.$$

Hence, the return level represents the level which is expected to be exceeded once every $1/p$ blocks.

(*ii*) *The corresponding return period $(1/p)$ is the expected waiting time until the return level z_p is next exceeded and it is given by*

$$p = 1 - G_{\boldsymbol{\theta}}(z_p).$$

So, once the three-parameter GEV distribution is fitted to the sample data, the ML parameter estimates are substituted in the above equations of Definition 2 to obtain the estimated return period and return level. The confidence intervals for the estimated return level and return period can then be calculated using the profile likelihood method, but an additional step is needed. The GEV distribution is reparameterized by defining $\psi = G_{\xi}^{-1}(1-p))$, which changes the unknown parameter vector of the GEV distribution from $\boldsymbol{\theta} = (\xi, \sigma, \mu)'$ to $\boldsymbol{\theta} = (\xi, \sigma, \psi)'$ (as seen in Definition 2 (i)). With this change, the profile likelihood confidence intervals for z_p can be calculated using the method outlined in Section 3.1.2.1.

3.1.3 Düsseldorf Case Study: BM Analysis

In the previous sections, the theoretical framework to model univariate extreme observations using the Block Maxima approach was established. The aim of this section is to perform the BM method on the Düsseldorf data introduced in Chapter 2. In other words, the GEV distribution is fitted to each univariate time series data for the two time horizons (P1 and P2). The results presented in this section were obtained by following the steps outlined in the following flowchart.

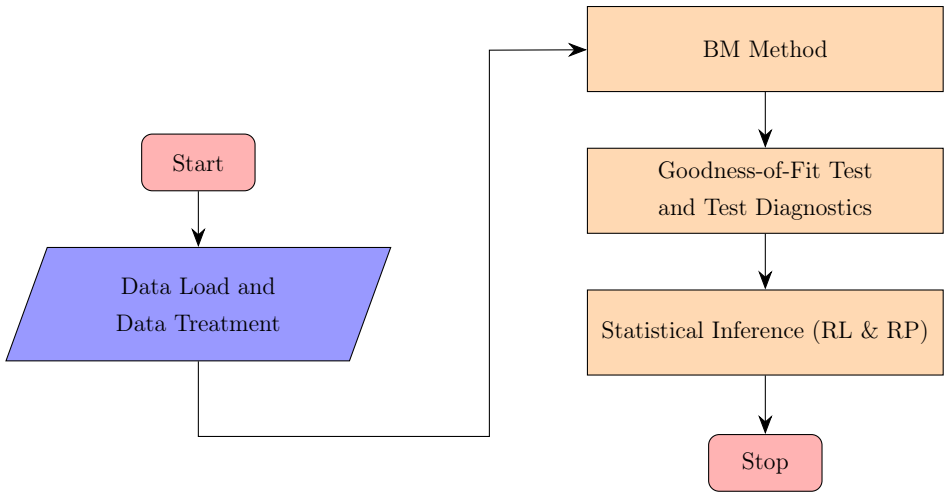


Figure 3.2: Flowchart for the BM method process.

This section of the main dissertation text only showcases the results of the extreme value analysis carried out on the time series T_{MRD} during the P1 time horizon. The same analytical approach was taken in the examination of T_{MRD} during the P2 time horizon and T_{MRS} in both time horizons. Tables B.1, B.2, and B.3 in Appendix B.2 contain the outcomes of these analyses. Accordingly, after the Düsseldorf dataset was loaded and pre-processed, the maximum values of MRD for each hydrological year were determined (see the observations marked in red in Figure 3.3). Then the three-parameter GEV distribution (3.7) was fitted to the hydrological annual maxima of the monthly river discharge data, and Table 3.1 shows the ML estimates of the three parameters, along with their approximate 95% confidence intervals. The confidence intervals of Table 3.1 were obtained using the profile likelihood method, and the corresponding profile likelihood functions are shown in Figure B.32.

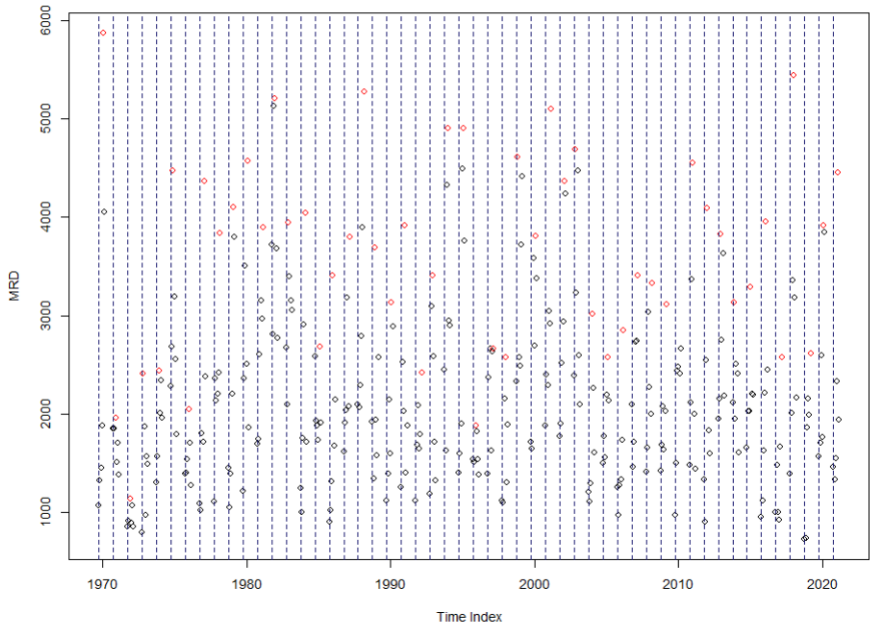


Figure 3.3: The MRD hydrological annual maxima observations are shown in red, while the hydrological annual blocks are shown in blue for the P1 time horizon in $1969-2020$.

The corresponding Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are 873.61 and 879.47, respectively. The Anderson-Darling (AD) test and diagnostic plots (see Figure 3.4) are used to evaluate the fit of the GEV distribution.

Parameter	95% lower CI	Estimated	95% upper CI
ξ	-0.54	-0.33	-0.03
σ	803.84	1055.62	1444.59
μ	2853.56	3258.75	3691.14

Table 3.1: GEV parameter estimates and corresponding confidence intervals.

The p-value of the AD test is found to be $0.76 > 0.05$, indicating that the sample follows the estimated GEV distribution. The points in the probability and quantile plots shown in the first row of Figure 3.4, are in fact observed to be relatively close to the bisector, except for a slight departure from linearity at the extreme ends of the tail of the quantile plot. The density curve is seen to accurately represent the data's histogram, and the empirical return levels lay very close to the model return level curve. This also shows that the GEV distribution gives a reasonable fit to the extremes of T_{MRD} in P1.

The final step in the BM procedure is to obtain the return level and the corresponding return period. According to standard flood protection guidelines of the North Rhine-Westphalia state, a return period of 200-500 years is typically considered for risk analysis [73]. Additionally, European Union regulations require the assessment of flood mitigation using two scenarios with return periods of 10 and 100 years, as well as an extreme scenario with higher return period [29]. Since the accuracy of the return levels decreases with increasing return periods, as also evidenced by the widening of the confidence intervals in Figure 3.5, it is suggested to consider flood scenarios with return periods of $20, 100, 200$ years. The return level estimates are shown in Figure 3.5, the values achieved are listed in Table 3.2 along with their 95% corresponding confidence intervals, using the profile likelihood method.

Return Periods	95% lower CI	Return Levels	95% upper CI
1 in 20 year event	4984.24	5269.34	5838.01
1 in 100 year event	5484.21	5777.78	6847.66
1 in 200 year event	5600.00	5925.04	7228.14

Table 3.2: The BM return levels estimates and the corresponding confidence intervals.

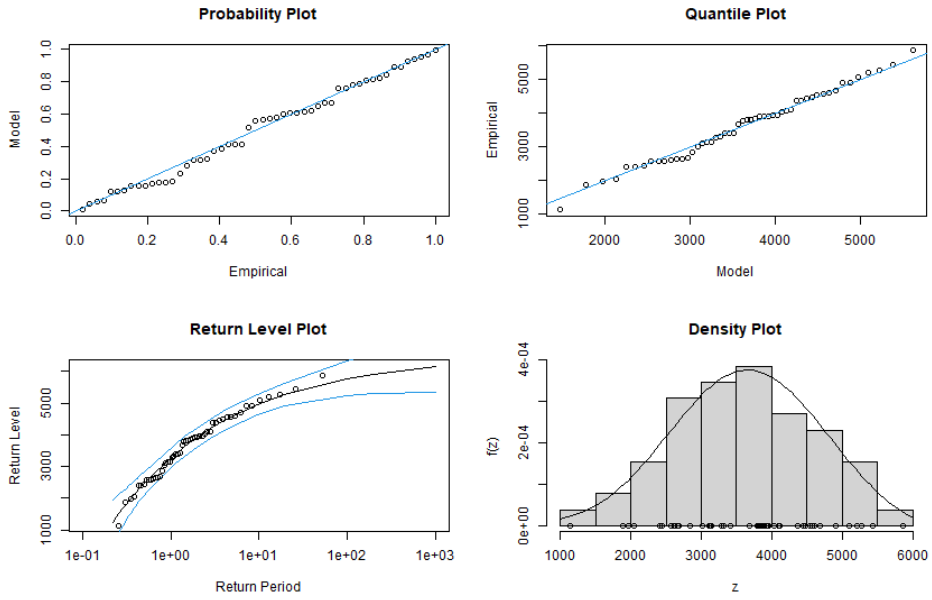


Figure 3.4: The diagnostic plots of the GEV distribution fit on the hydrological annual maxima observations of the T_{MRD} time series in the P1 time horizon.

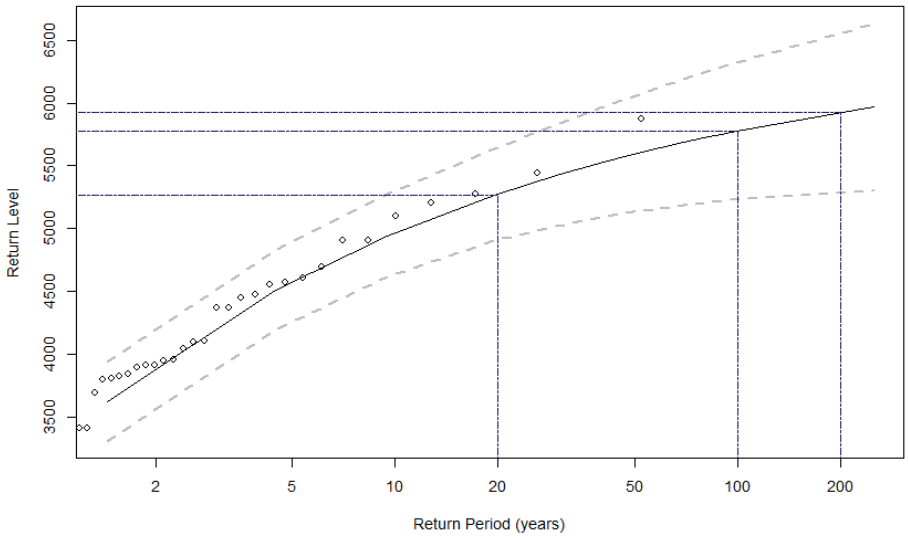


Figure 3.5: The BM return level curve, with the 95% profile likelihood confidence intervals represented by grey dotted lines, and the empirical return levels denoted by black circles, is shown for the time series T_{MRD} in the P1 time horizon. The estimated flood scenarios presented in Table 3.2 are displayed by blue dotted lines.

The corresponding profile likelihood functions are shown in Figure B.33. The estimate of the 1 in 20 year event is $z_{0.05} \approx 5269.34\,\mathrm{m}^3/\mathrm{s}$, which is interpreted as the level that is expected to be exceeded, on average, once every 20 years (blocks). The same interpretation applies to the other return periods. Accordingly, if the Düsseldorf flood protection is based on the highest values of Table 2.10, then the univariate scenarios in Table B.2 would not cause a flood, except for the 1 in 200 year event of T_{MRD} for P1 and P2 time horizons and the 1 in 200 year event of T_{MRS} for P1 time horizon. Furthermore, the results presented in Tables B.1, B.2, and B.3 demonstrate that when using the P2 time horizon, the extreme value distribution is reduced to a Gumbel distribution ($\xi = 0$) for both T_{MRD} and T_{MRS}, but the model fit was better when compared to the P1 time horizon. As a reminder, the shape parameter (ξ) represents the weight of the tail structure, which also impacts the structure of the return level curve. As a result, the return levels of T_{MRD} in P2 and T_{MRS} in P1 had a higher rate of increase compared to their corresponding counterparts. The results obtained for this section are produced by the R code uploaded in the GitHub repository [157].

3.2 K Largest Order Statistic Method

In the previous section, the univariate extreme observations were modeled using the Block Maxima approach. Even though this model provided a framework in which extreme observations could be modeled, the approach has a major drawback that might yield inaccurate statistical results. The BM method is a wasteful approach, as it only considers the maximum in each block and disregards other extreme events that occur in the same block. For instance, from Figure 3.1, it can be observed that there are blocks that contain extreme events other than the maximum of the block, which were ignored when using the BM approach. However, these extremes might hold crucial information regarding the behaviour of such events that has been disregarded. Consequently, a framework that mitigates this drawback by modeling these additional extreme observations is also considered. Throughout the following sections, we will present the K Largest Order Statistics (KLOS) method. This method provides a framework which considers multiple extreme observations from each block. The objective of the following

section is to introduce the generalized theoretical result produced by Weissman [161] and corresponding results that will serve as basis for the introduction of the KLOS method. The material covered in this section will then lead to the application of the method as a case study.

3.2.1 Modelling the K Largest Order Statistics

Let us consider the univariate IID random variables $\{X_1, \ldots, X_n\}$ with common underlying distribution function D. Let $M_n^{(k)}$ denote the k^{th} largest order statistic (see Definition $A.6$ in Appendix A.2) from the n random variables. Based on the relation between k and n, Arnold et al. [5] categorized the behaviour of $M_n^{(k)}$ into the following three types:

- Central Order Statistics: when $k = [np] + 1$ where $p \in (0, 1)$.
- Intermediate Order Statistics: when $k \to \infty$ but $k/n \to 1$ or 0.
- Extreme Order Statistics: when k is fixed and $n \to \infty$.

The distinct attributes of the categories mentioned above will lead to different asymptotic distributions. However, in this section the extreme order statistics will be considered, as it is the most similar to what was considered in the previous section and it gives the required statistical theory to model and acquire results on the extreme observations of the case study. On the basis of Theorem 1 covered in Section 3.1, it was concluded that the asymptotic non-degenerate distribution of the rescaled sample maximum is the GEV distribution. In view of this result, it is to be expected that the distribution of the rescaled extreme order statistic could be related to the GEV distribution. Through the following theorem, the GEV distribution is in fact generalized to a non-degenerate asymptotic distribution that models an arbitrary rescaled extreme order statistic.

Theorem 2. *If there exist sequences of constants* $\{a_n > 0\}_{n\in\mathbb{N}}$ *and* $\{b_n\}_{n\in\mathbb{N}}$ *such that*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \xrightarrow{d} G(z) \tag{3.12}$$

for some non-degenerate distribution G, so that G is of the same type as the GEV distribution given by (3.6)*, then, for each fixed $k \in \mathbb{N}$ it follows that*

$$\mathbb{P}\left(\frac{M_n^{(k)} - b_n}{a_n} \leq z\right) \xrightarrow{d} G^{(k)}(z) = G(z) \sum_{s=0}^{k-1} \frac{(-\ln G(z))^s}{s\,!}, \tag{3.13}$$

where $G(z) > 0$ and $G^{(k)}(z) = 0$ for all z that have $G(z) = 0$. Conversely, if for some fixed k,

$$\mathbb{P}\left(\frac{M_n^{(k)} - b_n}{a_n} \leq z\right) \xrightarrow{d} H(z) \tag{3.14}$$

for some non-degenerate distribution function H, then, $H(z)$ must be of the same form as $G^{(k)}(z)$ where (3.12) *holds with the same G, a_n and b_n such that* (3.13) *holds for all k.*

Proof. The proof of this theorem is included in Appendix A.1. □

Essentially, this theorem is stating that if the extreme order statistic $M_n^{(k)}$ is rescaled using the same normalizing constant sequences such that the limit (3.5) holds, then the asymptotic non-degenerate distribution is of the type (3.13). Similarly to the standard GEV distribution, the normalizing constant sequences are difficult to determine in practical applications. Hence, the normalizing constant sequences are absorbed by replacing the scale ($\sigma > 0$) and the location ($\mu \in \mathbb{R}$) such that from here on the three-parameter family that is determined as follows

$$G_{\xi,\sigma,\mu}^{(k)}(z) = G^{(k)}\left(\frac{z-\mu}{\sigma}\right) = G_{\xi,\sigma,\mu}(z) \sum_{s=0}^{k-1} \frac{(-\ln G_{\xi,\sigma,\mu}(z))^s}{s\,!}. \tag{3.15}$$

will be used on the adjusted support $\{z \,|\, 1+\xi(z-\mu)/\sigma > 0\}$. Subsequently, from (3.15) we can see that ξ is independent of k, so the parameters of $G_{\xi,\sigma,\mu}^{(k)}$ correspond to the parameters of $G_{\xi,\sigma,\mu}$. As a result, the three distinct tail structure cases of $G_{\xi,\sigma,\mu}^{(k)}$ emerge and are illustrated in Figure B.49 for $k = 3$. However, Theorem 2 alone is not sufficient for modeling the joint behaviour of multiple extreme values. A framework in which the joint behaviour of the several large extreme order statistics was thus developed and will be presented next. Accordingly, let the following vector represent the top k ($k \geq 2$) extreme statistics

$$\boldsymbol{M}_n^{(k)} = \left(M_n^{(1)}, \ldots, M_n^{(k)}\right)'.$$

Theorem 2 gives the asymptotic distribution of the components of $\boldsymbol{M}_n^{(k)}$. However, due to the ordering of the observations, the components of $\boldsymbol{M}_n^{(k)}$ are inherently dependent. For example, $M_n^{(2)}$ is always smaller than $M_n^{(1)}$. Thus, the asymptotic distribution of each component impacts the asymptotic distribution of the others. Therefore, Theorem 2 cannot be used directly to model $\boldsymbol{M}_n^{(k)}$. Using similar arguments as in Theorem 2, the convergence of the joint distribution of the top k rescaled extreme order statistics is found instead by using the following theorem. This asymptotic joint distribution is referred to as the standard GEV_k distribution.

Theorem 3 (Weissman)**.** *If there exist sequences of constants* $\{a_n > 0\}_{n\in\mathbb{N}}$ *and* $\{b_n\}_{n\in\mathbb{N}}$ *such that*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \xrightarrow{d} G_\xi(z) \qquad \text{as } n \to \infty \tag{3.16}$$

for some non-degenerate distribution G_ξ*, then, for some fixed* k *it follows that the top* k *order statistics vector:*

$$\tilde{\boldsymbol{M}}_n^{(k)} = \left(\frac{M_n^{(1)} - b_n}{a_n}, \ldots, \frac{M_n^{(k)} - b_n}{a_n}\right) = \left(\tilde{M}_n^{(1)}, \cdots, \tilde{M}_n^{(k)}\right) \tag{3.17}$$

has the following limiting joint distribution:

$$\begin{aligned}&\mathbb{P}\left(\tilde{M}_n^{(1)} < z^{(1)}, \cdots, \tilde{M}_n^{(k)} < z^{(k)}\right)\\ &\quad\to \sum_{s_1=0}^{1}\sum_{s_2=0}^{2-s_1}\cdots\sum_{s_{k-1}=0}^{k-1-s_1-\cdots-s_{k-2}} \frac{(\gamma_2-\gamma_1)^{s_1}}{s_1\,!}\cdots\frac{(\gamma_k-\gamma_{k-1})^{s_{k-1}}}{s_{k-1}\,!}\cdot e^{-\gamma_k},\end{aligned} \tag{3.18}$$

where $\gamma_i = -\ln\left(G_\xi\left(z^{(i)}\right)\right)$ *and* $z^{(k)} \leq z^{(k-1)} \leq \cdots \leq z^{(1)}$.

Proof. The proof of this theorem is included in Appendix A.1. □

As was explained previously, for practical reasons, the normalizing constant sequences are replaced by the scale ($\sigma > 0$) and the location ($\mu \in \mathbb{R}$) parameters. Now, although the general form of the cumulative joint distribution (3.18) is justified by the probabilistic theoretical result, it is quite cumbersome to work with in practical applications. Hence, the joint probability density function is used more and it is expressed as follows

$$g\left(z^{(1)},\ldots,z^{(k)}\right)=\begin{cases} e^{-e^{-\left(\frac{z^{(k)}-\mu}{\sigma}\right)}} \times \prod_{j=1}^{k} \sigma^{-1} e^{-\left(\frac{z^{(j)}-\mu}{\sigma}\right)} & \xi = 0, \\ e^{-\left[1+\xi\left(\frac{z^{(k)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}} \times \prod_{j=1}^{k} \sigma^{-1} \left[1+\xi\left(\frac{z^{(j)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} & \xi \neq 0, \end{cases} \tag{3.19}$$

defined on $\{z^{(j)} \mid 1+\xi(z^{(j)}-\mu)/\sigma > 0\}$ for $j = 1,\ldots,k$. The full lengthy derivation can be found in [123]. Theorem 3 and related results covered in this dissertation are incorporated in the KLOS method. The KLOS method is based on the same principles as the BM method that is outlined in Section 3.1, with a slight adjustment. The concept of the KLOS approach is to divide the nm observations obtained on a variable of interest into m blocks, where each block is treated as if it was a sample of size n. Then the top k observations from each block are obtained to form the set $\left\{\boldsymbol{M}_{ni}^{(k)} = \left(M_{ni}^{(1)},\ldots,M_{ni}^{(k)}\right)' \mid 1 \leq i \leq m\right\}$. Assuming that the blocks are disjoint, Theorem 3 suggests fitting the GEV_k distribution on $\boldsymbol{M}_{ni}^{(k)}$. Figure 3.6 summarizes this process, wherein the top 3 extreme order statistics are collected from each block.

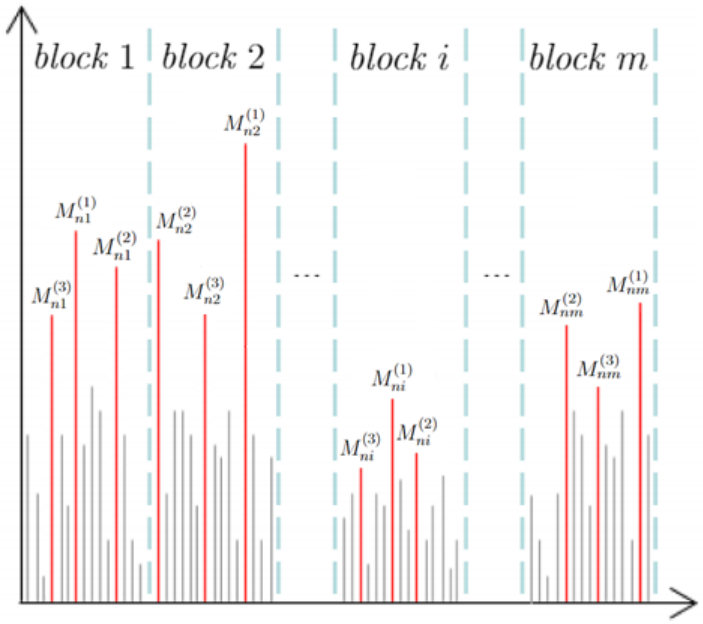


Figure 3.6: KLOS method with $M_{ni}^{(k)}$ as the k^{th} order statistic of block i.

Similar to the BM approach, the KLOS method is based on the block size assumption, but with the added challenge of selecting the appropriate value of k. This leads to a trade-off between parameter estimation bias and variance in the estimates. If a significantly small value of k is selected, then, few extreme observations would be chosen, which could lead to a higher variance. In [147], it is stated that as k increases the rate of convergence to the asymptotic distribution decreases significantly. Hence, if k is selected to be significantly large, there is a risk that the approximation of the data by the GEV_k distribution will no longer holds, leading to bias. There is no specific guidance in the literature for selecting the appropriate value of k, but the standard practice is to choose the largest value that maintains adequate model diagnostics. Nevertheless, [7] suggests that the value of k should be relatively small compared to the block size.

3.2.2 Inference on the GEV_k Distribution

Using sample data, the KLOS method provides a framework in which the largest k extreme order sample statistics from each time period are modeled through the GEV_k distribution. The same technique used for the BM method to estimate the unknown parameters of the GEV distribution will be implemented for the GEV_k distribution. The return level and return period concepts that were considered for the BM method will also be used in this section to provide information on future events. These techniques will be heavily used in the case study.

3.2.2.1 Parameter Estimation

As mentioned for the BM approach, there are various methods which can be used to estimate the unknown parameters of the GEV_k distribution. Considering the reasons listed for the BM method and the ability to compare the results between the two approaches, the MLE approach will also be considered for the KLOS method. Suppose that the block size and k are adequate, then the sample of top k extreme statistics vectors $\mathbb{M}^{(k)} = \left\{ \boldsymbol{M}_{n1}^{(k)}, \ldots, \boldsymbol{M}_{nm}^{(k)} \right\}$ is achieved from the m blocks. Let $\boldsymbol{\theta} = (\xi, \sigma, \mu)'$ represent the vector of parameters of the GEV_k distribution with the corresponding density function (3.19). Then, the log-likelihood function for $\xi \neq 0$ is given by

$$l\left(\boldsymbol{\theta}; \mathbb{M}^{(k)}\right) = \sum_{i=1}^{m} -k\ln(\sigma) - \left[1+\xi\left(\frac{M_{ni}^{(k)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} - \left(\frac{1}{\xi}+1\right)\sum_{j=1}^{k}\ln\left(1+\xi\left(\frac{M_{ni}^{(j)}-\mu}{\sigma}\right)\right) \tag{3.20}$$

providing that $1+\xi\left(M_{ni}^{(j)}-\mu\right)/\sigma > 0$ for $j =\in \{1,\ldots,m\}$, otherwise the log-likelihood function is equal to zero. Alternatively, for $\xi = 0$ the log-likelihood function reduces to

$$l\left(\boldsymbol{\theta}; \mathbb{M}^{(k)}\right) = \sum_{i=1}^{m} -k\ln(\sigma) - e^{-\left(\frac{M_{ni}^{(k)}-\mu}{\sigma}\right)} - \sum_{j=1}^{k}\frac{M_{ni}^{(j)}-\mu}{\sigma}. \tag{3.21}$$

The MLE $\widehat{\boldsymbol{\theta}}$ is obtained by maximizing the log-likelihood function with respect to the parameter $\boldsymbol{\theta}$ over the appropriate parameter space. As was the case for the BM approach, there is no analytical solution, so estimates are acquired by numerical optimization algorithms. Since the parameter $\boldsymbol{\theta}$ corresponds to the parameters of the GEV distribution, the limitations of the MLE for the GEV distribution found by Smith [148] will also hold for the GEV_k distribution. In fact, the general MLE properties hold for $\xi > -0.5$ where Tawn [151] derived the standard errors and the Fisher information matrix using a similar version of (3.10). Finally, the profile likelihood method presented for the BM method can also be applied to obtain the confidence intervals for the parameters of the GEV_k distribution.

3.2.2.2 Return Level and Return Period

The return level and the corresponding return period served as essential risk measure metrics for the BM method case study. It is reasonable to develop such metrics and integrate them into the KLOS approach. The return level metric refers to the quantile of the asymptotic distribution, as specified in Definition 2. However, as already stated, the GEV_k distribution (3.18) is complicated to work with. If k is equal to 1, the log-likelihood functions (3.20) and (3.21) are equivalent to the log-likelihood functions (3.8) and (3.9) that were used for the BM method. Therefore, based on the structure of the log-likelihood function, it is clear that the log-likelihood functions derived for the KLOS method provide parameter estimates that align with the GEV distribution,

but also allow for the inclusion of additional extreme observations. As a result, the estimated parameters have the same interpretation, but with improved accuracy from the additional information included [26]. Thus, the return level, return period and their respective confidence intervals are calculated as defined for the BM method but using the parameter estimates from the KLOS method.

3.2.3 Düsseldorf Case Study: KLOS Analysis

As Figure 3.3 demonstrates, there are instances where values in one block are not considered to be extreme in that block (black circle), but corresponding values in another block are considered as maxima/extreme (red circle). To gain a complete understanding of the dynamics of extreme events, it is thus essential to consider these types of occurrences in our investigation. In this section, we aim to present the results obtained from applying the KLOS method presented in this section to the Düsseldorf data introduced in Chapter 2. Specifically, the GEV_k will be fitted to each univariate time series data for both the P1 and P2 time horizons. The process will be similar to what was done for the BM case (as outlined in Figure 3.2), but the KLOS method will be used instead.

This section of the main dissertation text only showcases the results of the extreme value analysis carried out on the time series T_{MRD} during the P1 time horizon. The same analytical approach was taken in the examination of T_{MRD} during the P2 time horizon and T_{MRS} in both time horizons. Tables B.4, B.5, and B.6 in Appendix B.2. Accordingly, after loading and pre-processing the data from Düsseldorf, it was noted that the KLOS model can be applied for any value of $k \in \{1, \ldots 6\}$ if the block size is selected to represent a year (this choice aligns with the structural framework used in the BM approach). The justification for using the KLOS method was investigated by Tawn in [151]. This justification was based on three main assumptions:

(1) Observations within and across blocks are roughly independent,

(2) The distribution of the block maxima follows the GEV_1 distribution,

(3) The block size is relatively large when compared to the choice of k.

Taking into account that each observation represents the monthly average, the first assumption is satisfied. Additionally, the second assumption is implicitly acknowledged in the first case study. Although the third assumption for our case is not fulfilled, we will demonstrate at a later stage that it is not a significant issue. Subsequently, the MLE and their corresponding estimated standard errors for different values of k are given in Table 3.3.

k	AIC	BIC	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\xi}$
1	873.61	879.47	3258.75 (158.74)	1055.62 (114.28)	−0.33 (0.09)
2	1647.24	2380.31	3297.70 (125.84)	919.05 (63.40)	−0.22 (0.09)
3	2380.31	2380.05	3250.37 (115.10)	913.22 (61.49)	−0.15 (0.08)
4	3075.15	3074.89	3209.57 (110.08)	915.77 (65.24)	−0.10 (0.07)
5	3765.75	3765.50	3302.32 (118.16)	977.75 (83.12)	−0.10 (0.07)
6	4431.99	4431.73	3214.36 (109.88)	949.46 (69.57)	−0.11 (0.06)

Table 3.3: GEV_k parameter estimates and estimated standard errors (in parentheses).

It is expected that as k increases, the estimated standard errors of the model will decrease. This is related to the increase in parameter estimation precision. Table 3.3 and Figure 3.7 suggest that there is instability in the estimated standard errors for $\hat{\mu}$ and $\hat{\sigma}$ when $k \geq 4$, while the estimated standard errors for $\hat{\xi}$ continue to decrease with increasing k. The framework of the KLOS method is set up in such a way that if the model is appropriate for a certain k, then the parameter estimates of the KLOS models fitted on fewer order statistics should be stable. However, the instability in the location and scale estimates when $k \geq 4$ raises doubts about the validity of the model for the corresponding k. Another possible way to assess the model fit for different k is to consider the return level curves. Recall from Section 3.2.2.2, that these are created in the same way as they are for the BM method, but using the ML estimates obtained using the KLOS model. Figures B.50 and B.51 demonstrate the return level curves for $k \in \{1, \ldots, 6\}$. Based on the analysis of the least displacement between the empirical return levels and the modeled return level curves, as well as the corresponding profile likelihood confidence intervals, it has been determined that the most suitable KLOS model for the time series T_{MRD} during the P1 time horizon is when $k = 3$. This decision

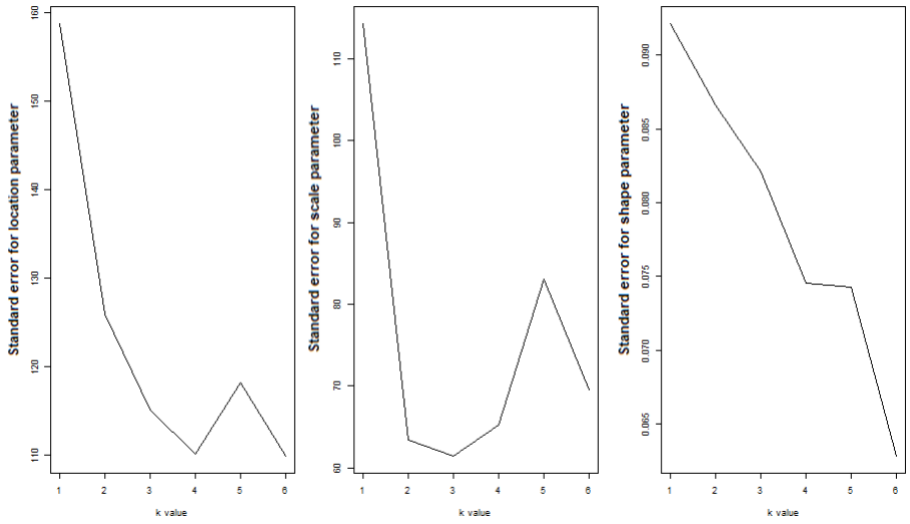


Figure 3.7: The MLE standard errors for the GEV_k distribution fit on the top k hydrological annual extreme observations of the time series T_{MRD} in the P1 time horizon.

is reinforced by the assessment of the fit of the k^{th} order statistic distribution (3.15) for $k \in \{1, 2, 3\}$, as outlined in Theorem 2, which are illustrated in Figures B.52 and B.53. The extreme observations selected are marked in red in Figure B.54. Table 3.4 shows the confidence intervals calculated using the profile likelihood method, and the corresponding profile likelihood functions are illustrated in Figure B.55.

Parameter	95% lower CI	Estimate	95% upper CI
ξ	-0.29	-0.15	0.03
σ	809.53	913.22	1063.38
μ	3032.40	3250.37	3484.52

Table 3.4: GEV_3 parameter estimates and corresponding confidence intervals.

Despite the negative value of $\widehat{\xi}$, the 95% upper confidence bound extends just above zero. This suggests that it would be worth testing whether the two-parameter model (GEV_3 with ξ fixed to zero) is more appropriate than the three-parameter model. Following the same arguments that were used in Section 3.1.3 to assess the fit of the proposed model, from the diagnostic plots of the two model fits shown in Figures 3.8 and

B.56 it was determined that the three-parameter GEV_3 distribution is more adequate to model the time series T_{MRD} in the P1 time horizon. In fact, the probability and quantile plots shown in Figure 3.8 demonstrate that the empirical points are relatively close to the bisector, with minimal deviation from linearity at the lower end of the tail of the quantile plot. Additionally, the density curve is seen to closely match the data's histogram, and the empirical return levels are close to the model's return level curve, with narrower confidence intervals. Additionally, an AD test was conducted to check that the observations come from the three-parameter GEV_3 distribution, and the resulting p-value is 0.06 which suggests that the null hypothesis cannot be rejected.

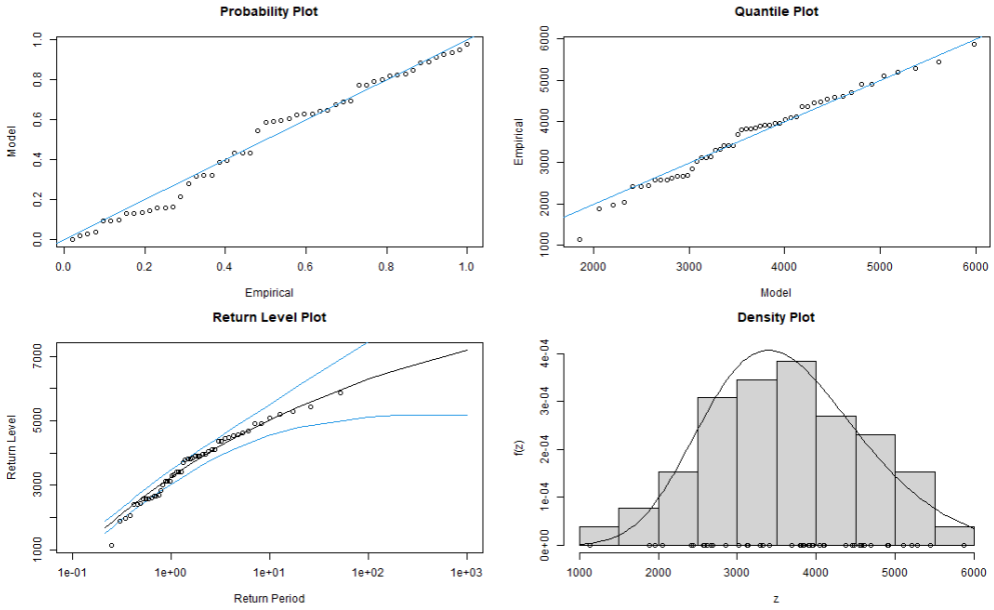


Figure 3.8: The diagnostic plots of the GEV_3 distribution fit on the k largest hydrological annual extreme observations of the time series T_{MRD} in the P1 time horizon.

Figure 3.9 shows the return level estimates, which are listed in Table 3.5 along with their 95% corresponding confidence intervals, calculated using the profile likelihood method. Additionally, the profile likelihood functions are presented in Figure B.57. Subsequently, it is observed that the values for T_{MRD} in Table B.6 are marginally higher than the values found in Table B.2. Hence, if the flood protections were based on the highest values found in panel (a) of Table 2.10, the univariate scenarios outlined in

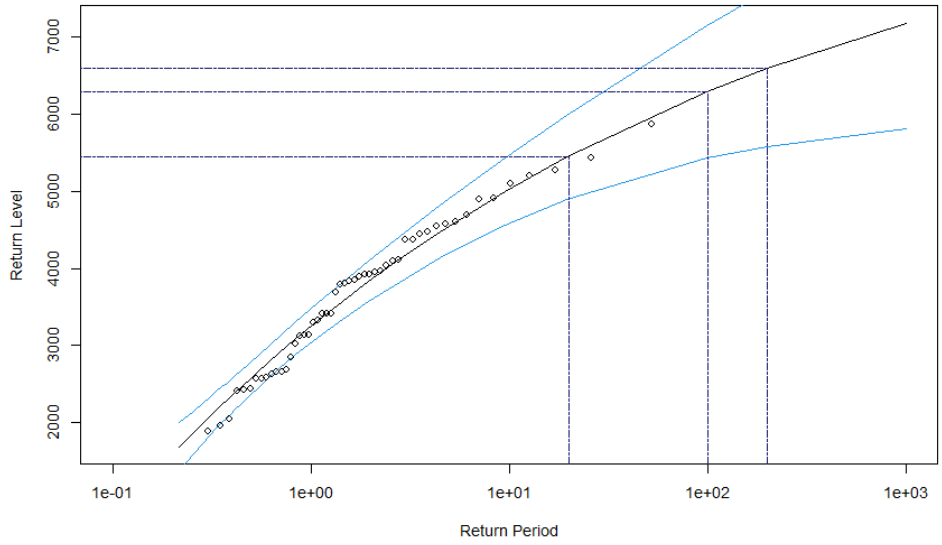


Figure 3.9: The KLOS return level curve, with the 95% profile likelihood confidence intervals (light blue dotted lines) and the empirical return levels (black circles). The estimate flood scenarios presented in Table 3.5 are displayed by the blue dotted lines.

Return Periods	95% lower CI	Return Levels	95% upper CI
1 in 20 year event	4973.91	5442.17	6425.93
1 in 100 year event	5525.87	6290.67	8200.14
1 in 200 year event	5689.13	6594.52	8983.38

Table 3.5: The KLOS return levels estimates and the corresponding confidence intervals.

Table B.6 would result in a flood, except for the 1 in 20 year events of T_{MRD} in P1 and P2. Conversely, the values for T_{MRS} in Table B.6 are approximately the values found in Table B.2. Similarly, if the flood protections were based on the highest values found in panel (b) of Table 2.10, the univariate scenarios outlined in Table B.6 would not result in a flood. However, a similar rate of increase in the return level value from one scenario to the other can be viewed in both tables. The AIC and BIC values (see Tables B.3 and B.4) and the corresponding diagnostic plots demonstrate that the best fit was always provided by the BM method. The results obtained for this section are produced by the R code uploaded in the GitHub repository [157].

3.3 Peak over Threshold Method

The previous section introduced the K Largest Order Statistic (KLOS) approach, an extension of the Block Maxima (BM) approach that allows for modeling additional univariate extreme observations. However, the KLOS model also has some drawbacks, such as the uneven distribution of extreme observations among blocks. Alternative methodologies that do not employ the blocking procedure can be utilized to overcome these limitations. In fact, an alternative method known as the Peak over Threshold (POT) method that models threshold exceedances is preferred over the previous models as it provides a more comprehensive description of the upper tail of the underlying distribution [18]. Hence, the target of this section is to introduce the theoretical result produced by Balkema & De Haan [8] and Pickands [79] and the corresponding results that form the basis of the POT method. The material covered in this section will be applied to the Düsseldorf data as a case study.

3.3.1 Modelling Threshold Excesses

Let us consider the univariate IID random variables $\{X_i\}_{i\in\mathbb{N}}$ with common underlying distribution function D. It is natural to identify an event as extreme when a given threshold ($u \in \mathbb{R}$) is exceeded $\{X_i > u\}$. Thus, a framework built around the behaviour of the magnitude of the exceedances over a given arbitrary threshold $\{X_i - u\}$ is essential. This type of extremal behaviour is modeled by the excess distribution function.

Definition 3 (Excess distribution over threshold u)**.** *Let Z be a random variable with a distribution function D. For a fixed $u < z^*$ and $0 \leq z \leq z^* - u$ the excess distribution over the threshold u is defined as follows*

$$\begin{aligned} D_u(z) = \mathbb{P}(Z \leq u + z | Z > u) &= \frac{\mathbb{P}(\{Z \leq u + z\} \cap \{Z > u\})}{\mathbb{P}(Z > u)} \\ &= \frac{\mathbb{P}(Z \leq u + z) - \mathbb{P}(Z > u)}{\mathbb{P}(Z > u)} \\ &= \frac{D(u + z) - D(u)}{1 - D(u)}. \end{aligned} \tag{3.22}$$

In general, the underlying distribution D is undefined. Therefore, the excess distribution (3.22) is not useful in practical applications. To address this limitation, similar to the approach developed for the BM method, we consider the asymptotic behavior of (3.22) for high threshold values, as presented in the following theorem.

Theorem 4. *The distribution function D belongs to $\mathcal{D}(G_\xi)$ (see Definition A.7) if and only if there exists a positive and measurable function σ such that*

$$\lim_{u \uparrow z^*} \frac{1 - D(u + z\sigma(u))}{1 - D(u)} = \begin{cases} (1 + \xi z)^{\frac{-1}{\xi}} & \xi \neq 0, \\ e^{-z} & \xi = 0. \end{cases} = 1 - H_\xi(z), \tag{3.23}$$

where $H_\xi(z)$ is known as the Generalized Pareto (GP) distribution defined on $\{z \mid z > 0$ and $1 + \xi z > 0\}$.

Proof. The proof of this theorem is included in Appendix A.1. □

The GP distribution mentioned in Theorem 4 is at the heart of the POT method, just as the GEV distribution is for the BM method, and the following theorem shows that, for increasing thresholds, the excess distribution converges uniformly to the GP distribution, provided that the conditions in Theorem 1 hold. This limiting theorem was established by Balkema & De Haan [8] and Pickands [79].

Theorem 5 (Pickands–Balkema–De Haan)**.** *Let $\{X_1, X_2, \ldots, X_n\}$ be a sequence of IID random variables with a common distribution D. Suppose there is a positive-measurable function $\sigma(u)$ then*

$$\lim_{u \uparrow z^*} \sup_{0 < z < z^* - u} \left| D_u(z) - H_\xi\left(\frac{z}{\sigma(u)}\right) \right| = 0 \tag{3.24}$$

if and only if $D \in \mathcal{D}(G_\xi)$ for every $\xi \in \mathbb{R}$.

Proof. The proof of this theorem is included in Appendix A.1. □

In practice, the positive-measurable function $\sigma(u)$ is unknown. Therefore, similar to what was done for the normalizing sequences of the GEV and GEV_k distribution functions, $\sigma(u)$ is replaced by the scale parameter ($\sigma > 0$). As a result, the two-parameter family is determined as follows:

$$H_{\xi,\sigma}(z) = H_{\xi}\left(\frac{z}{\sigma}\right) = \begin{cases} 1 - \left(1 + \xi\left(\frac{z}{\sigma}\right)\right)^{\frac{-1}{\xi}} & \xi \neq 0, \\ 1 - e^{-\left(\frac{z}{\sigma}\right)} & \xi = 0, \end{cases} \tag{3.25}$$

and defined on the adjusted support $\{z \mid z > 0 \text{ and } 1 + \xi z/\sigma > 0\}$. It is known that the GP distribution function can be directly obtained from the GEV distribution function, through the following relationship.

$$H_{\xi}(z) = 1 + \ln(G_{\xi}(z)), \qquad \frac{1}{e} \leq G_{\xi}(z) \leq 1, \tag{3.26}$$

where ξ is equivalent for both distributions. This means that the GP distribution also exhibits the same three cases of tail structure, as illustrated in Figure $B.87$. By considering these theoretical results, we can implement the POT method. Let us consider a specific sample $(x_1, \ldots, x_n)$ of IID observations, but the following arguments apply to other possible samples as well. Suppose that for some ξ, the underlying distribution function $D(x)$ belongs to the MDA (see Definition A.7) of the GEV distribution. The observations that exceed the high threshold value are identified and relabeled $\{\tilde{x}_j \mid x_i > u\}_{j \in \{1,\ldots,N_u\}}$ according to the total number of extreme observations represented by $N_u = \#\{i \mid i \in \{1, \ldots, n\} \,\&\, x_i > u\}$. Then the sample of corresponding threshold excesses is determined as follows

$$\boldsymbol{y} = \{y_j \mid y_j = \tilde{x}_j - u\}_{j \in \{1,\ldots,N_u\}}. \tag{3.27}$$

As per Theorem 5, for a sufficiently high threshold, the two-parameter GP distribution is an adequate approximation to model the sample of threshold excesses (3.27). This process is outlined in Figure 3.10, and it is clear that the POT method exploits more information from the observations when compared to the previous models. The implementation of the POT method depends on the crucial assumption that the threshold value is adequate. Similar to the block size in the BM method, this choice will lead to a trade-off between variance and bias in the GP distribution parameter estimates. If the threshold value is selected to be too high, there will be fewer extreme observations, resulting in a high variance in the estimators. If the threshold value is selected to be too low, the asymptotic assumptions might not continue to hold, leading to bias in the estimators. Consequently, Tawn and Heffernan [72] recommend that the appropriate threshold value should not produce significant shifts in the estimates if the threshold is

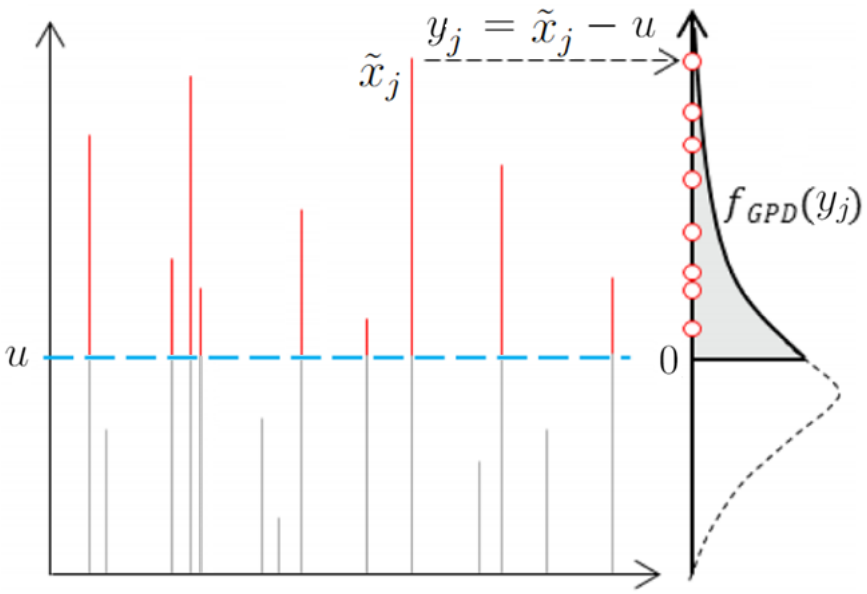


Figure 3.10: POT method with threshold u.

slightly increased. Hence, the standard practice is to select the lowest possible threshold provided that the conditions still hold. Graphical methods used to identify a suitable threshold value are also reviewed in Section 3.3.3.

3.3.2 Inference on the GP Distribution

Essentially, the framework of the POT method allows to assess and model the sample of threshold excesses through the GP distribution for some suitable threshold. Similar to the BM and KLOS methods, techniques to estimate the unknown parameters of the GP distribution will be presented. These techniques will be utilized in the case study.

3.3.2.1 Parameter Estimation

Analogous to the BM and KLOS methods, there are several approaches that are used in practical applications to estimate the unknown parameters of the two-family GP distribution. For the same reasons listed for the BM method, the MLE approach is also considered for the POT method to facilitate comparison of results among the three derived approaches. In the following arguments, the threshold u is assumed to be

the most optimal such that the sample of threshold excesses (3.27) is achieved. Let $\boldsymbol{\theta} = (\xi, \sigma)'$ represent the vector of parameters of the considered distribution. Then, following the general steps of the MLE procedure as demonstrated in [26], the log-likelihood function when $\xi \neq 0$ is given by

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = -N_u \ln(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{N_u} \ln\left(1 + \frac{\xi y_j}{\sigma}\right),$$

provided that $\sigma > 0$ and $(1 + \xi y_j/\sigma) > 0$ for each j. In the case that $\xi = 0$, then it can be deduced that the log-likelihood reduces to

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = -N_u \ln(\sigma) - \sigma^{-1} \sum_{j=1}^{N_u} y_i.$$

The MLE $\widehat{\boldsymbol{\theta}}$ is obtained by maximizing the log-likelihood function with respect to the parameter $\boldsymbol{\theta}$ over the appropriate parameter space. In practice, the maximization is done by numerical non-linear optimization algorithms. In view of (3.26), Smith [148] demonstrated that the MLE limitations for the GEV distribution also hold for the GP distribution. In fact, the fundamental attributes of MLE, namely asymptotic efficiency, consistency, and asymptotic normality, remain valid only when $\xi > -0.5$. Finally, the profile likelihood method that was presented for the BM method can be used to obtain the confidence intervals for the two parameters of the GP distribution.

3.3.2.2 Return Level and Return Periods

The concepts of return level and return period, which were introduced in Definition 2 for the BM method, can also be adopted for the POT method for inferring statistical results on future extreme events through the GP distribution. Suppose that the GP distribution with threshold u is adequate to model the sample of exceedances. Then, it can be deduced from (3.22) that for $z \geq u$ the right tail probability is derived as follows

$$\begin{aligned}
\bar{D}(z) &= \mathbb{P}(Z > u)\mathbb{P}(Z > z \mid Z > u) \\
&= \bar{D}(u)\mathbb{P}(Z - u > z - u \mid Z > u) \\
&= \bar{D}(u)\bar{D}_u(z - u)
\end{aligned}$$

$$= \begin{cases} \bar{D}(u)\left(1+\xi\left(\frac{z-u}{\sigma}\right)\right)^{\frac{-1}{\xi}} & \xi \neq 0, \\ \bar{D}(u)\, e^{-\left(\frac{z-u}{\sigma}\right)} & \xi = 0, \end{cases} \tag{3.28}$$

where $\bar{D}(u)$ is referred to as the portion (rate) of threshold excesses. Thus, the level x_t that is expected to be exceeded once every t observations, is the $(1/t)$-quantile of $\bar{D}(z)$. Hence, the t-observation return level is defined as follows

$$x_t = \begin{cases} u + \frac{\sigma}{\xi}\left((t\bar{D}(u))^{\xi} - 1\right) & \xi \neq 0, \\ u + \sigma \ln(t\bar{D}(u)) & \xi = 0, \end{cases} \tag{3.29}$$

provided that t is large enough so that $x_t > u$. Furthermore, if there are b_t observations per time period (corresponding to the block size in the BM method), then the t-period return level that is expected to be exceeded once every t periods is defined by

$$z_t = \begin{cases} u + \frac{\sigma}{\xi}\left((tb_t\bar{D}(u))^{\xi} - 1\right) & \xi \neq 0, \\ u + \sigma \ln(tb_t\bar{D}(u)) & \xi = 0. \end{cases} \tag{3.30}$$

In order to estimate the return level, the two parameters of the GP distribution and the portion of threshold excesses are required to be estimated. A natural estimate of the portion of threshold excesses is given by the empirical distribution function

$$\hat{\bar{D}}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{x_i > u\}} = \frac{N_u}{n}.$$

Subsequently, the return level can be fully estimated by substituting the above and the ML estimates instead of the unknown parameters of the GP distribution. Furthermore, after a reparameterization of the GP distribution, the profile likelihood procedure can also be applied to obtain the confidence intervals of the t-period return level. Subsequently, consider (3.30) then

$$\sigma = \begin{cases} \frac{(z_t-u)\xi}{tb_t\bar{D}(u)^{\xi}-1} & \xi \neq 0, \\ \frac{z_t-u}{\ln(tb_t\bar{D}(u))} & \xi = 0. \end{cases} \tag{3.31}$$

The profile likelihood confidence interval of the t-period return level is obtained by substituting (3.31) into the log-likelihood function and maximized with respect to ξ and fixed z_t. Then continue to follow the procedure described for the BM method to obtain the required confidence interval.

3.3.3 Threshold Selection Methods

As previously indicated, the selection of the appropriate threshold u is a crucial step for the POT method. For this reason, this section is dedicated to review two graphical methods that will aid in the threshold selection stage of the case study. The diagnostic methods are based on the Mean Residual Life (MRL) plot and the Parameter Stability (PS) plots. The idea behind these approaches is to assess the properties of the model fit for several thresholds through graphical diagnostics.

3.3.3.1 Mean Residual Life Plot

The procedure that is based on the MRL plot centres around the mean of the two-parameter GP distribution function. Let us consider a random variable Y that follows the two-parameter GP distribution (3.25), then

$$\mathbb{E}(Y) = \frac{\sigma}{1-\xi} \tag{3.32}$$

provided that $\xi < 1$ and when $\xi = 1$ the mean is infinite. On the other hand, the mean of the excess distribution (3.22) is determined as follows

$$e(u) = \mathbb{E}(Z - u \mid Z > u), \quad 0 \leq u \leq z^* \tag{3.33}$$

and it is referred to as the mean excess function over the threshold value u. Assuming that the excesses over a high threshold u generated from the sample $X_1, \ldots, X_n$ follow a two-parameter GP distribution, then it can be deduce that

$$D_u(z) = \frac{D(u+z) - D(u)}{1 - D(u)} = \begin{cases} H_{\xi,\,\sigma+\xi u}(u) & \xi \neq 0, \\ H_{\xi,\,\sigma}(u) & \xi = 0. \end{cases}$$

This implies that the excess distribution of GP random variable over the threshold u is again the GP distribution. This shows the lack of memory property when $\xi = 0$ and the stability property when $\xi \neq 0$. Consequently, following the above deductions and (3.32), the mean excess function of the GP distribution is

$$e(u) = \frac{\sigma + \xi u}{1 - \xi}. \tag{3.34}$$

provided that $\xi < 1$. Furthermore, as a result of the threshold stability property deduced from Lemma A.4, if a higher threshold is selected, the approximation of the excess distribution is a GP distribution with the same ξ but different scale parameter that increases linearly with respect to the selected higher threshold. Subsequently, it can be deduced that for any $v > u$, the conditional expectation becomes

$$e(v) = \frac{\sigma + \xi(v-u)}{1-\xi} = \frac{\sigma - \xi u}{1-\xi} + \frac{\xi v}{1-\xi}, \tag{3.35}$$

provided that $\xi < 1$. Hence, by virtue of (3.35) it can be concluded that for $v > u$ the mean excess function is a linear function of v. This property is the fundamental diagnostic criterion for threshold selection using the MRL process. Consequently, consider the sample of threshold excesses $\boldsymbol{Y}$ and the empirical counterpart of the mean excess function (3.33):

$$e_n(u) = \frac{1}{N_u} \sum_{i=1}^{N_u} y_i, \tag{3.36}$$

then the MRL plot is constructed by plotting the locus of points of the possible thresholds against the corresponding empirical mean excess values

$$\{(u, e_n(u)) \mid u < z^*\}.$$

Then, the suitable threshold is the lowest possible threshold value such that the empirical mean excess levels of higher threshold values appear to approximately fall in a straight line. However, the selection of the appropriate threshold based on the linearity approximation is subjective and difficult to interpret. This is because deviance from the straight line might be due to the GP distribution no longer being an optimal distribution to model the excesses, or it might be the result of the high variability of (3.36) due to the sample size. Hence, confidence intervals will turn out to be a valuable tool to assist in the threshold selection process in the case study. Furthermore, to reinforce the threshold selected based on the MRL plot, an alternative method will also be introduced in the next section.

3.3.3.2 Parameter Stability Plot

An alternative approach for threshold selection is based on the Parameter Stability (PS) Plot. This method, although still subjective, has a different way of determining

the appropriate threshold. Tawn and Heffernan [72] proposed to evaluate the consistency of the GP distribution estimated parameters across a range of threshold values.

Let us assume that the GP distribution is appropriate for modeling the sample of threshold excesses $\boldsymbol{y}$ when the threshold is fixed at u. According to the threshold stability property that was deduced in Lemma $A.4$, a sample of threshold excesses over a higher threshold ($v > u$) follows the same distribution with the same shape parameter, but the scale parameter changes with respect to the higher threshold level ($A.34$). By applying a reparameterisation to the scale parameter ($\sigma^* = \sigma - \xi v$), it can be concluded from ($A.34$) that the modified scale parameter will be invariant of the threshold v. Thus, it can be deduced that the parameter estimates of $H_{\xi,\,\sigma^*}(z)$ are constant for all $v > u$, accounting for sampling variability. The PS plots are constructed by plotting the possible threshold against either σ^* or ξ. The appropriate threshold is said to be the lowest value, such that the parameter estimates for higher threshold values stay approximately constant. Similar to the MRL plot, threshold selection is subjective, using confidence intervals based on the profile likelihood method can support the threshold selection process.

3.3.4 Düsseldorf Case Study: POT Analysis

As Figure B.54 demonstrates, there are cases where events in one block are deemed extreme while events of greater magnitude in other blocks are not considered as such. To fully comprehend the dynamics of extreme events, it is essential to take into account these types of occurrences in our analysis. In this section, we present the results obtained from applying the POT model to the Düsseldorf data introduced in Chapter 2. Specifically, the GP distribution will be fitted to each univariate time series data for both the P1 and P2 time horizons. The process will be similar to what was done for the BM case (as outlined in Flowchart 3.2), but the POT method will be used instead. In line with the first and second case studies presented in this chapter, this section of the main dissertation text only showcases the results of the extreme value analysis carried out on the time series T_{MRD} during the P1 time horizon. The same analytical

approach was taken in the examination of T_{MRD} during the P2 time horizon and T_{MRS} in both time horizons. Tables B.7, B.8, and B.9 contain the outcomes of these analyses. Subsequently, after the data has been loaded and pre-processed, the MRL plot and PS plot have been computed and are displayed in Figures 3.11 and 3.12, respectively. These plots are used as diagnostic tools to determine the threshold level.

In fact, the MRL graph in Figure 3.11 shows the relationship between the threshold values and the empirical mean excess values, along with their corresponding 95% confidence intervals. The plot appears to be roughly linear within the range of 2700 to 3200 and 3400 to 3700. However, the plot appears to be curved before the first interval and unstable after the second interval. Based on the number of observations that are in excess of the thresholds in both intervals (shown by the top axis), it was determined that the first region is more adequate. As a result, it is concluded that the optimal threshold level is $u = 2700$, with $N_u = 82$ remaining observations.

Furthermore, Figure 3.12 presents the Parameter Stability plots of $\widehat{\sigma^*}$ and $\widehat{\xi}$ across a range of threshold values, complete with their corresponding 95% confidence intervals. These plots serve as a supplementary technique for verifying the conclusions drawn from the MRL plot. It can be observed that both graphs are stable after $u = 2700$ until the sample becomes too small to fit the asymptotic model, as shown by the increasing confidence bands. This complements the conclusion drawn from the MRL plot. By using the selected threshold, the extreme values that exceed $u = 2700$ are determined and are presented in Figure 3.13. Following Theorem 3, it is reasonable to assume that the sample of excesses follows a two-parameter GP distribution. Subsequently, the corresponding maximum likelihood estimates along with their approximate 95% confidence intervals are shown in Table 3.6.

Parameter	95% lower CI	Estimate	95% upper CI
ξ	-0.70	-0.43	-0.05
σ	936.60	1460.98	2233.56

Table 3.6: GP parameter estimates and corresponding 95% confidence intervals.

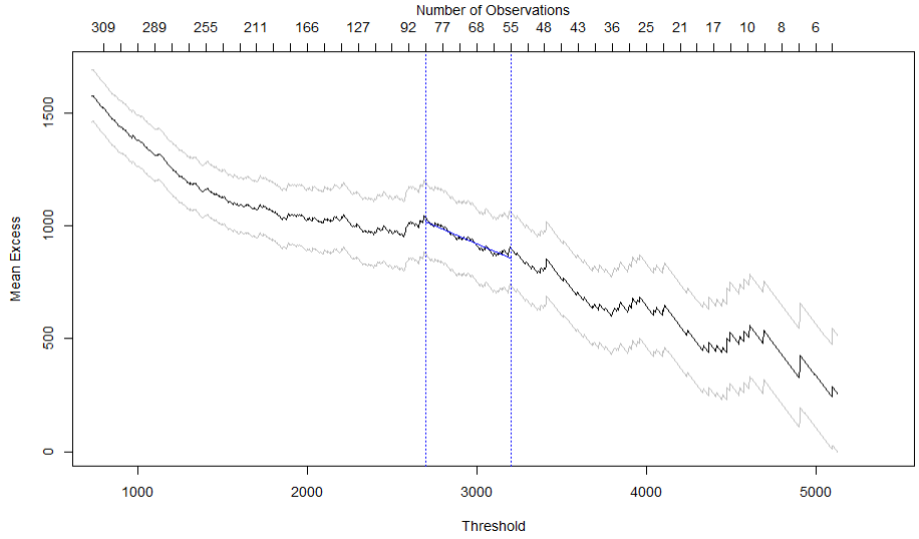


Figure 3.11: The empirical MRL plot for the threshold excesses of the time series T_{MRD} in the P1 time horizon featuring a solid blue line, indicating the plot's linearity after the threshold $u = 2700$, and two vertical dotted blue lines, which mark the corresponding threshold values and simultaneously the remaining number of observations.

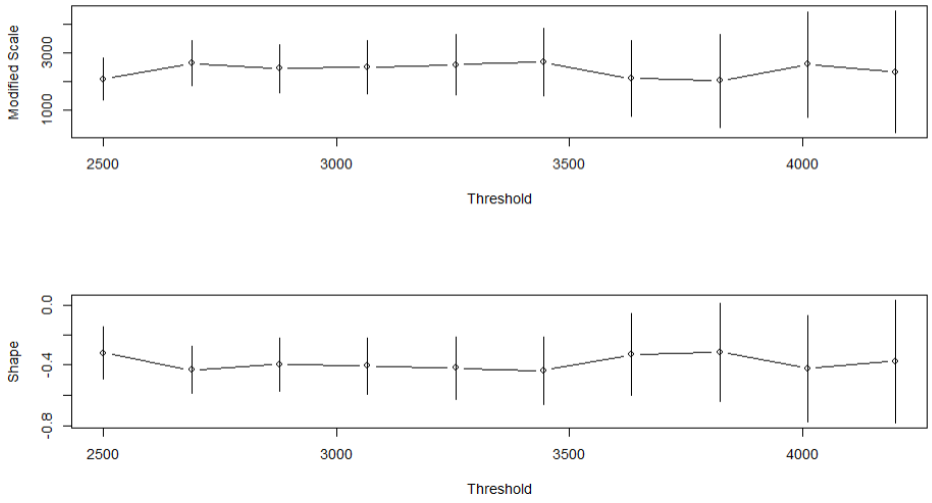


Figure 3.12: The parameter stability plots for the threshold excesses of the extreme observations of the time series T_{MRD} in the P1 time horizon. The black circles are the parameter estimates and the vertical lines represent the corresponding 95% confidence intervals.

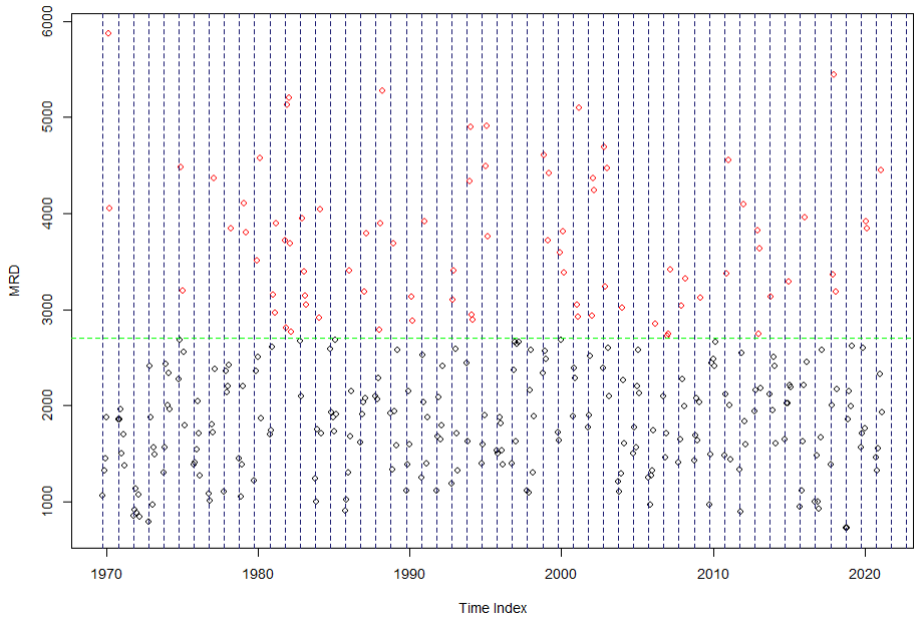


Figure 3.13: The MRD extreme observations that are over the threshold $u = 2700$ (green line) are marked with a red circle and the hydrological annual blocks with blue dotted lines for the P1 time horizon in $1969 - 2020$.

The confidence intervals mentioned in Table 3.6 were obtained using the profile likelihood method and the corresponding profile likelihood functions are shown in Figure B.88. The Anderson-Darling (AD) test was conducted to check that the observations come from the GP distribution, and the resulting p-value id 0.94 which suggests that the null hypothesis cannot be rejected. Following the same arguments that were used in Sections 3.1.3 and 3.2.3 to assess the fit of the proposed model, from the diagnostic plots (see Figure 3.14) it was observed that the points in the probability and quantile plots are relatively close to the bisector, except for a slight departure from linearity at the extreme end of the left tail of the quantile plot. The fitted density curve is seen to accurately represent the empirical density curve, and the empirical return levels lay very close to the model return level curve. Overall, from the diagnostic plots, it can be concluded that the GP distribution is a reasonable fit. Subsequently, the final step in the POT procedure is to assess the return level and the corresponding return period.

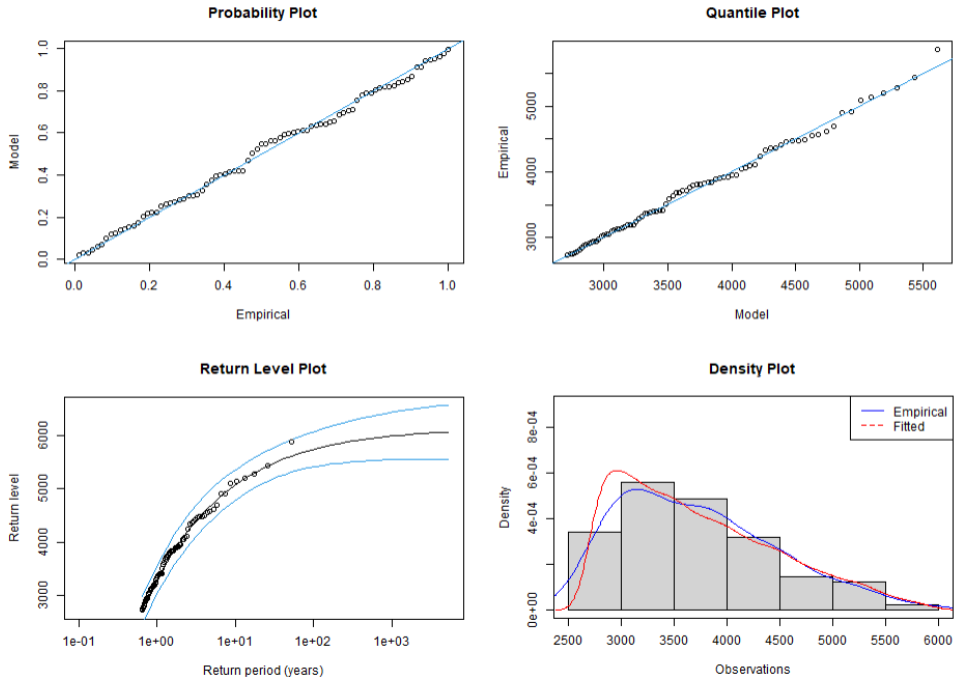


Figure 3.14: The diagnostic plots of the GP distribution fit on the extreme observations threshold excess of the time series T_{MRD} in the P1 time horizon.

The same return period that was considered in Sections 3.1.3 and 3.2.3, will also be considered for this section. As a result, Figure 3.15 shows the return level estimates listed in Table 3.7, along with their 95% corresponding confidence intervals calculated using the profile likelihood method. The corresponding profile likelihood functions are shown in Figure B.89.

Return Periods	95% lower CI	Return Levels	95% upper CI
1 in 20 year event	5087.88	5342.79	5751.97
1 in 100 year event	5500.51	5734.75	6409.60
1 in 200 year event	5608.59	5836.55	6633.84

Table 3.7: The POT return levels estimates and the corresponding confidence intervals.

The thresholds selected for each time series over the two time periods, which are displayed in B.9, are of a comparable level, with the P2 thresholds consistently being the lowest. This is consistent with the summary statistics presented in Table 2.13.

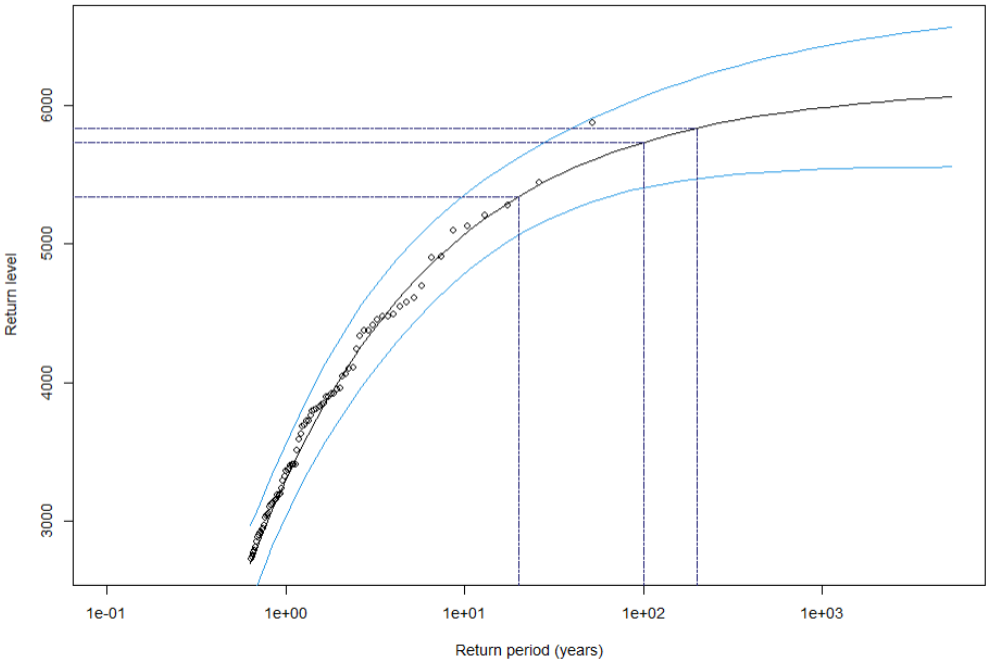


Figure 3.15: The POT return level curve, with the 95% profile likelihood confidence intervals (light blue dotted lines) and the empirical return levels (black circles) for the time series T_{MRD} in the P1 time horizon. The estimate flood scenarios shown in Table 3.7 are displayed by the blue dotted lines.

In general, it is observed that the values for T_{MRD} in Table B.8 are similar to the values found in Tables B.2 and B.6. However, the values for the P2 time horizon are the highest out of the three models. In fact, if the flood protections were based on the highest values found in panel (a) of Table 2.10, the univariate scenarios outlined in Table B.8 would not result in a flood, except for the 1 in 100 and the 1 in 200 year events for the P2 time horizon. Conversely, the values for T_{MRS} in Table B.8 are the highest out of the three models for the P1 time horizon but have similar values for the P2 time horizon. Similarly, if the flood protections were based on the highest values found in panel (b) of Table 2.10, the univariate scenarios outlined in Table B.8 would not result in a flood for the P2 time horizon but the univariate scenarios for the P1 time horizon would result in a flood, except for the 1 in 20 year event.

The AIC and BIC indicators shown in Tables B.3, B.4 and B.9, along with the corresponding diagnostic plots, indicate that the GEV distribution consistently provides the best fit to the extreme observations of the two time series T_{MRD} and T_{MRS} during both periods, except when the GP distribution is fitted to the excess of the extreme observations of the time series T_{MRS} in the P1 time horizon. It is essential to compare the return levels obtained for the two time horizons across the three methods and return periods. Based on the values listed in Tables B.2, B.6 and B.8, the comparison is shown in Figures B.111 and B.112. Consequently, it can be deduced that for the time series T_{MRD}, the return levels obtained for each time horizon for every return period are on the same level. However, for the time series T_{MRS}, there are significant differences between the return levels obtained for each time horizon for every return period. Notably, the values for the P1 time horizon are substantially higher than those for the P2 time horizon. This trend is consistent with the summary statistics presented in Table 2.13. The results obtained for this section are produced by the R code uploaded in the GitHub repository [157].

In the next chapter, the two time series at each time horizon will be analyzed in the multivariate domain. The first approach will assume a particular dependence framework which is constructed on the basis that the marginal distributions are in the form of a GEV distribution. Although, the extremes of the time series T_{MRS} in the P1 time horizon are best modeled by the GP distribution, the GEV distribution is still a reasonable model to consider. Thus, the second approach will still consider the GEV distribution form for the marginals but will explore other dependence structures.

Chapter 4

Multivariate Analysis

The previous chapter focused entirely on modelling extreme univariate observations using techniques based on the GEV, GEV_k, and GP distributions. However, since many physical processes are composed of multiple single processes, relying solely on univariate models may not provide a comprehensive assessment of extreme events in the data, which could lead to inaccurate risk estimation. Additionally, Chapter 2 demonstrated that there is dependence between the observations of the time series T_{MRD} and T_{MRS}, yet the univariate models discussed in Chapter 3 did not account for this dependence framework. Multivariate distributions are more capable of representing the extreme joint structure of multiple components. One of the main aims of Multivariate Extreme Value Theory (MEVT) and related methods is to identify methods which provide insight on the tail structure of distributions that model multivariate extreme observations.

Having had the theoretical foundations for univariate extreme value theory (UEVT), the theory behind multivariate extreme analysis developed quite rapidly, due to its practical applicability. However, the transition from analysing a univariate structure to a multivariate structure created some fundamental theoretical complications. For instance, in a multivariate context, the traditional definition of an extreme observation loses its significance since there is no obvious way to determine order in high dimensions. This was cardinal in univariate extreme analysis. Given this absence of order, the target was to identify ways to determine a certain region, referred to as the failure

region, in which the elements of this region are deemed to be extreme, due to having a low probability of occurrence but, if they do occur, can lead to serious consequences. The objective of this chapter is to introduce two commonly used approaches in MEVT. These methods, namely the Component-wise Block Maxima (CWBM) method and the General Copula-based (GCB) method, are described in Sections 4.2.1 and 4.2, respectively. Both methodologies utilize copula functions as the primary tool for modeling the dependence framework, as will be shown in the Düsseldorf case study. In fact, the CWBM method focuses on Extreme Value copulas, while the GCB method focuses on Archimedean copulas. Like the univariate models, the two approaches discussed in this chapter assume that the observations are independent and identically distributed (IID).

4.1 Component-wise Block Maxima Method

The main objective of this section is to extend to the multivariate dimension the main theoretical result of Section 3.1.1, which was a crucial component of the introduction of the BM approach. Consequently, the CWBM approach and its accompanying techniques are introduced in Sections 4.1.1, 4.1.2, and 4.1.3. The information covered in these sections will then be applied to the Düsseldorf data as a case study, which is presented in Section 4.1.4.

4.1.1 Modelling Component-wise Maxima

In view of the fact that there is no apparent order for the multivariate observations to determine the extremes at either end of the scale, Barnett [10] determined that the marginal ordering is the most fruitful relation for multivariate extreme analysis, where the arithmetic vector operations and order relations are applied component-wise. Consider an arbitrary sample of independent and identically distributed (IID) random vectors in $\mathbb{R}^d$, denoted by $\{\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,d})' \mid 1 \leq i \leq n\}$. These vectors all share the same underlying distribution D, which is different from the one used in Section 3.1. Then, given the nature of the case study, the approach considered in this section is based on the sample component-wise maxima that is defined as follows

$$\boldsymbol{M}_n = (M_{n,1}, \ldots, M_{n,d})' = \left(\max_{1 \leq i \leq n} (X_{i,1}), \ldots, \max_{1 \leq i \leq n} (X_{i,d}) \right)'. \tag{4.1}$$

It is important to note that in this context, the index i in (4.1) might be different for each component, and so the sample component-wise maxima obtained may not be a vector of observations corresponding to an element from the sample considered. Consequently, it can be noted that the shift from univariate to multivariate is not entirely straightforward. However, it is important to note that the sample component-wise maxima is a potential extreme observation. Hence, assuming that a maximum domain of attraction exists in the multivariate setting, it is still reasonable to establish the limiting distribution of the normalized sample component-wise maxima. Thus, let us assume that there are appropriate normalizing sequences of constants $\boldsymbol{a}_n = \{a_{n,1}, \ldots, a_{n,d}\} \in \mathbb{R}_+^d$ and $\boldsymbol{b}_n = \{b_{n,1}, \ldots, b_{n,d}\} \in \mathbb{R}^d$ such that, as $n \to \infty$ we have

$$\begin{aligned} \mathbb{P}\left(\frac{M_{n,1} - b_{n,1}}{a_{n,1}} \leq z_1, \ldots, \frac{M_{n,d} - b_{n,d}}{a_{n,d}} \leq z_d \right) &= (D(a_{n,1} z_1 + b_{n,1}, \ldots, a_{n,d} z_d + b_{n,d}))^n \\ &\xrightarrow{\mathbf{d}} G(z_1, \ldots, z_d), \end{aligned} \tag{4.2}$$

for all continuity points $\boldsymbol{z} = (z_1, \ldots, z_d)' \in \mathbb{R}^d$ of a d-dimensional non-degenerate distribution function G with non-degenerate marginals. If (4.2) holds then G is referred to as a multivariate extreme value (MEV) distribution. For each $j \in \{1, \ldots, d\}$ let D_j and G_j represent the j^{th} univariate marginal distribution of D and G respectively. As it is assumed that the marginal distributions of G are non-degenerate and recall that a sequence of random vectors converges in distribution if and only if the corresponding marginal sequences converge in distribution, then, by setting all z_j but one to ∞ it can be concluded that

$$\lim_{n \to \infty} D_j^n (a_{n,j} z_j + b_{n,j}) = G_j(z_j). \tag{4.3}$$

Consequently, the form of the asymptotic marginal distributions (4.3) is of the same type as the GEV distribution by virtue of Theorem 1 with corresponding normalizing sequences of constants. Let $U_j(n) = D_j^{\leftarrow}(1 - 1/n)$ (see Definition A.2 in Appendix A.2) where $n > 1$. Then, by (4.3) and the extended regular variation result (see Theorem A.8), there are an auxiliary positive functions $a_j(n)$ such that for $z > 0$:

$$\lim_{n \to \infty} \frac{U_j(nz) - U_j(n)}{a_j(n)} = \frac{z^{\xi_j} - 1}{\xi_j}. \tag{4.4}$$

Hence, if the sequence $\left(z_{n,j} = \frac{U_j(nz_j) - U_j(n)}{a_j(n)}\right)$ for all $j \in \{1, \ldots, d\}$ is considered, it will imply that

$$D^n(U_1(nz_1), \ldots, U_d(nz_d)) = D^n(a_1(n)z_{n,1} + U_1(n), \ldots, a_d(n)z_{n,d} + U_d(n)).$$

Considering (4.2), it can be observed that the suitable choice of normalizing constants is $a_{n,j} = a_j(n)$ and $b_{n,j} = U_j(n)$. Since the margins of G are continuous, it is shown in [80] that G is also continuous. Given that D is monotone, it can be deduced that

$$\lim_{n\to\infty} D^n(U_1(nz_1), \ldots, U_d(nz_d)) = G\left(\frac{z_1^{\xi_1} - 1}{\xi_1}, \ldots, \frac{z_d^{\xi_d} - 1}{\xi_d}\right). \tag{4.5}$$

However, from (4.5) it can be inferred that the form of the limiting distribution is influenced by the behavior of the marginal distributions. To remove this influence, the marginal variables need to be standardized to follow a specific distribution. A mathematically convenient approach is to transform the univariate margins to follow the unit Fréchet distribution ($G_{1,1,1}(z) = e^{-1/z}$ for $z > 0$) with $a_{n,j} = n$ and $b_{n,j} = 0$ for all $j \in \{1, \ldots, d\}$. As a result of the transformation and Proposition A.1, (4.2) can be reformulated as follows

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le i\le n}\left(\frac{-1}{\ln(D_1(X_{i,1}))}\right) \le nz_1, \ldots, \max_{1\le i\le n}\left(\frac{-1}{\ln(D_d(X_{i,d}))}\right) \le nz_d\right) \\
&\qquad = D_*^n(nz_1, \ldots, nz_d) \\
&\qquad = D^n\left(\left(\frac{-1}{\ln(D_1)}\right)^{\leftarrow}(nz_1), \ldots, \left(\frac{-1}{\ln(D_d)}\right)^{\leftarrow}(nz_d)\right) \\
&\qquad \xrightarrow{\mathbf{d}} G(G_1^{\leftarrow}(e^{-1/z_1}), \ldots, G_d^{\leftarrow}(e^{-1/z_d})) = G_*(\boldsymbol{z}),
\end{aligned} \tag{4.6}$$

for $z_1 \ge 0, \ldots, z_d \ge 0$, where D_* represents the underlying distribution of the transformed random vectors. As a result, the marginal distribution becomes $G_{\star j}(z) = G_{1,1,1}(z)$ and G_* is referred to as a simple multivariate extreme value distribution. According to Proposition A.1 there is no loss of generality in assuming specific margins as the extreme value properties are still preserved [97]. Consequently, through the following theorem, the class of all possible multivariate extreme value (MEV) distribution functions with prescribed Fréchet marginals is derived.

Theorem 6. *Let $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be a sequence of IID random vectors with a common underlying distribution function D_* with unit Fréchet margins. If G_* is a multivariate extreme value distribution with unit Fréchet marginals then there is a finite measure S on $\Xi = \left\{\boldsymbol{w} \in \mathbb{R}_+^d \mid \sum_{j=1}^d w_j = 1\right\}$ such that G_* has the following form*

$$G_*(z_1, \ldots, z_d) = \exp\left\{-\int_\Xi \max_{1\leq j\leq d}\left(\frac{w_j}{z_j}\right) S(d\boldsymbol{w})\right\}, \tag{4.7}$$

with the mean constraint

$$\int_\Xi w_j \, S(d\boldsymbol{w}) = 1. \tag{4.8}$$

Proof. The proof of this theorem is included in Appendix A.1. □

A generalized representation of Theorem 6 is derived from the fact that as a result of equation (4.8) we have the following

$$S(\Xi) = \int_\Xi 1 \, d\Xi = \int_\Xi \sum_{j\leq d} w_j \, S(d\boldsymbol{w}) = \sum_{j\leq d} \int_\Xi w_j \, S(d\boldsymbol{w}) = d,$$

and as $S(d\boldsymbol{w})/S(\Xi)$ is a probability measure on $[0, 1]$ then

$$S(\Xi) \int_\Xi \max_{1\leq j\leq d}\left(\frac{w_j}{z_j}\right) S(d\boldsymbol{w})/S(\Xi) = d \int_{[0,1]^d} \max_{1\leq j\leq d}\left(\frac{w_j}{z_j}\right) H(d\boldsymbol{w}).$$

Subsequently, as the case study will only consider two-dimensional data, then, considering the above, equation (4.7) will reduce to

$$G_*(z_1, z_2) = \exp\left\{-2\int_0^1 \max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right) H(dw)\right\}, \tag{4.9}$$

with the corresponding mean constraint

$$2\int_0^1 wH(dw) = 1. \tag{4.10}$$

In essence, this representation implies that the class of bivariate extreme value distribution functions with unit Fréchet margins is in one-to-one correspondence with the set of distribution functions that satisfy (4.10). Given that (4.10) is the only constraint on the distribution function H, then it can be concluded that H is not unique and as a

result, (4.9) cannot be represented by any finite dimensional parametric family ($d > 1$). In essence, the probability measure H characterizes the dependence structure among the standardized extreme observations. When the mass of H is concentrated within the interior of $[0,1]^d$, it assigns a strong dependence structure, while concentration at the boundary of $[0,1]^d$ assigns a weak dependence structure. For simplicity in modeling and statistical analysis of extreme observations, copula theory is employed. This approach offers practical and convenient techniques that investigate the non-linear dependence structure among the extreme random observations, which will be discussed in more detail in Section 4.1.2.

Definition 4 (Copula Function)**.** *Let $I = [0,1]$. A d-Copula is a d-dimensional function $C : I^d \mapsto I$ that satisfies the following properties for all $u_i \in I$:*

(*i*) *(Grounded) Whenever $u_i = 0$ for at least one component $i \in \{1, \ldots, d\}$ we have*

$$C(u_1, \ldots, u_d) = 0,$$

(*ii*) (*Uniform margins*) *If for some $i \in \{1, \ldots, d\}$ $u_j = 1$ for all $j \neq i$, then we have*

$$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i, \tag{4.11}$$

(*iii*) (*Non-decreasing*) *For every hyperrectangle $B = \prod_{i=1}^{d} [x_i, y_i] \subset [0,1]^d$ the volume is non-negative*

$$\int_B dC(u) = \sum_{\boldsymbol{z} \in \prod_{i=1}^{d} \{x_i,\, y_i\}} (-1)^{N(\boldsymbol{z})} C(\boldsymbol{z}) \geq 0, \tag{4.12}$$

where $N(\boldsymbol{z}) = \#\{k \mid z_k = z\}$.

Thus, a d-dimensional copula is a joint distribution function on I^d with uniform margins. The relation between the d-dimensional distribution G_* and the corresponding copula function C_{G_*} can be established through the functional equation stated in the following theorem that was developed by Sklar [146].

Theorem 7 (Sklar's Theorem)**.** *Let F be a d-dimensional distribution function with margins F_i for $i \in \{1, \ldots, d\}$. Then there exists a d-dimensional copula C such that for all $(z_1, \ldots, z_d) \in \mathbb{R}^d$ it holds that*

$$F(z_1, \ldots, z_d) = C(F_1(z_1), \ldots, F_d(z_d)). \tag{4.13}$$

If F_i for all $i \in \{1, \ldots, d\}$ are continuous, then C is unique. Conversely, if C is a d-dimensional copula and F_i are univariate distribution functions for each $i \in \{1, \ldots, d\}$, then the function F defined via (4.13) *is a d-dimensional distribution function.*

Proof. The proof of this theorem is found in [104] and references therein. □

Sklar's theorem is an important building block for this dissertation, as the representation (4.13) can be used effectively to construct the limiting joint distribution G_* in simpler form. This is achieved by combining a copula function with the unit Fréchet distributions. In fact, if G_* is a simple MEV distribution function, then by Theorem 7 it can be deduced that there exists a copula function C_{G_*} such that

$$G_*(\boldsymbol{z}) = C_{G_*}(G_{1,1,1}(z_1), \ldots, G_{1,1,1}(z_d)). \tag{4.14}$$

In light of this, the focus has now shifted towards the class of extreme value copulas (EVC) that corresponds to the class of asymptotic distributions G_* through the relation (4.14). When we recall the proof of Theorem 6, we can infer that the simple MEV distribution displays the homogeneity characteristic. In fact, this property manifests itself as an analytic property of the EVC, as stated in the following definition.

Definition 5 (Extreme Value Property)**.** *A copula $C : I^d \mapsto I$ is called an extreme value copula if it satisfies the following*

$$C(u_1^t, \ldots, u_d^t) = C(u_1, \ldots, u_d)^t, \tag{4.15}$$

for all $t > 0$ and $u_i \in I$ for all $i \in \{1, \ldots, d\}$.

Proposition A.1 states that if G_* is an MEV distribution, so is G. Thus, our modelling approach will directly involve C_G, avoiding the interim step of using C_{G_*} and then transitioning back to C_G to extract the required information. According to (4.3), the

univariate marginals of G are GEV distributed, which is a continuous function. Then, by Sklar's Theorem, it can be deduced that the representation (4.14) converted to G is unique. However, there are several standard families that satisfy (4.15) and so there is no unique parameterization for the class of EVC. In practical applications, a plausible approach is to consider a parametric extreme value sub-family of the class of EVC that aligns with the data.

Consequently, for the Düsseldorf case study, the Gumbel-Hougaard copula, also known as the logistic copula, was selected due to its widespread use in climate and hydrological studies, owing to its simple and flexible functional form and its ability to represent a broad range of dependence structures. Let $\alpha \geq 1$ be known as the dependence parameter, and $u_1, \ldots, u_d \in [0,1]$, then the Gumbel copula is defined as follows

$$C_\alpha(u_1, \ldots, u_d) = \exp\left\{-\left[(-\ln u_1)^\alpha + \cdots + (-\ln u_d)^\alpha\right]^{\frac{1}{\alpha}}\right\}. \tag{4.16}$$

This family is particular as it happens to be the only one that is both an EVC and an Archimedean copula (see Definition $A.11$) [56]. The Archimedean generator associated with this family is $\psi(z) = \exp\left(-z^{1/\alpha}\right)$. The proposed dependence structure ranges from independence ($\alpha = 1$) to complete dependence ($\alpha = \infty$). However, some limitations of this family are that the interchangeability of the variables forces certain symmetric properties and this family has asymptotically independent lower tails (4.19).

The copula selected is important for the component-wise block maxima (CWBM) method. The structure of the CWBM approach is similar to that of the BM method. The initial step would be to partition the nm vectors of observations obtained for a variable of interest into m blocks, where each block is treated as if it were a sample of size n. Then, from each block, the component-wise maxima are obtained to form $\boldsymbol{M} = \{\boldsymbol{M}_i = (M_{i,1}, \ldots, M_{i,d})' \mid 1 \leq i \leq m\}$. As a result of Theorem 6 and the arguments that followed, the components of the component-wise maxima in $\boldsymbol{M}$ are considered separately. This results in constructing a univariate sequence of independent block maxima ($\{M_{i,j}\}_{i=1}^m$) for each dimension $j \in \{1, \ldots, d\}$, which are modeled by the three-parameter GEV distribution (3.7) using the approach described in Section 3.1.2.1.

Once the vector of estimates $(\hat{\mu}_j, \hat{\sigma}_j, \hat{\xi}_j)'$ is obtained for each dimension j using univariate maximum likelihood estimation, the component-wise maxima are transformed to form $\left\{\tilde{\boldsymbol{M}}_i = (G_{\hat{\mu}_1,\hat{\sigma}_1,\hat{\xi}_1}(M_{i,1}), \ldots, G_{\hat{\mu}_d,\hat{\sigma}_d,\hat{\xi}_d}(M_{i,d}))' \mid 1 \leq i \leq m\right\}$. Finally, the Gumbel copula is fitted to the new sample using the approach described in Section 4.1.3.1.

Considering that the implementation of the multivariate approach is parallel to the univariate BM method, it follows that the CWBM method heavily relies on the same assumption that the block size is appropriate. Specifically, implementing the CWBM method requires the estimation of the parameters of the GEV distribution for each dimension. As a result, it is crucial to find a balance between estimation bias and variance in the estimates. Considering that if the block size is relatively large, it will result in fewer maxima, leading to a larger variance. Meanwhile, if the block size is relatively small, this will result in an insufficient number of observations in each block. Thus, the approximation assumption of the GEV distribution would no longer be valid, leading to bias in the parameter estimates. The block size is often determined based on the length of the available data, and a typical choice, as mentioned earlier for the BM method, is a length representing one year.

4.1.2 Measures of Copula Dependence

As noted previously, the interrelationship between random variables is a crucial topic in modeling multivariate data. In practice, it is essential to quantitatively measure the strength of the dependence between variables. Several dependence measures have been introduced to quantify the dependence structure in a model, such as Pearson's correlation coefficient, Spearman's rho coefficient, and Kendall's tau coefficient, among others. The copula function provides a general structure that conveniently connects together the univariate distributions, according to a dependence framework. However, directly extracting information on the dependence structure is still complicated. Therefore, it is reasonable to consider the measures of dependence in the context of the copula. The aim of this section is to introduce both linear and non-linear measures, as well as tail dependence measures. These measures will all be used in the Düsseldorf case study.

4.1.2.1 Linear Measures

The Pearson's correlation coefficient is the most popular dependence measure used in both theoretical and applied contexts. The Pearson correlation coefficient, which ranges from -1 to 1, is used to assess the strength and direction of the linear association. If the Pearson's coefficient value is -1, the variables have a perfect negative linear relationship, that is, if one variable increases, then the other variable decreases. Conversely, if the value of the coefficient is 1 it implies that the variables have a perfect positive linear relationship. If the random variables are normally distributed, a coefficient value of 0 implies independence; otherwise, it indicates a non-linear relationship.

4.1.2.2 Non-linear Measures

An apparent shortcoming of the Pearson correlation coefficient is that the variables under consideration might have a heavy-tailed distribution, leading to the possibility that the second order moment does not exist. As a result, Pearson's correlation coefficient may not always exist. According to Lemma A.5, the Pearson correlation coefficient aggregates information from marginal behaviour and the dependence structure given by the joint distribution, making it inadequate to capture non-linear changes in the marginals. To address these limitations, concordance measures were introduced to capture the difference between the probability of concordance and discordance (see Definition A.15), making them capable of capturing non-linear relationships between variables. Thus, the following section describes the two most prominent concordance measures: Kendall's tau coefficient and Spearman's rho coefficient.

4.1.2.2.1 Kendall's tau coefficient

Let us consider two IID bivariate random vectors $(X_1, Y_1)'$ and $(X_2, Y_2)'$. Then, the population version of the Kendall's tau coefficient is obtained as follows

$$\tau = \mathbb{P}(\underbrace{(X_1 - X_2)(Y_1 - Y_2) > 0}_{\text{concordance}}) - \mathbb{P}(\underbrace{(X_1 - X_2)(Y_1 - Y_2) < 0}_{\text{discordance}}). \tag{4.17}$$

Analogous to the Pearson correlation coefficient , the range of the Kendall's tau coefficient is $[-1, 1]$. By Theorem A.9, the Kendall's tau coefficient can be presented in

terms of the copula function:

$$\tau = 4 \int_{[0,1]^2} C(u_1, u_2)\, dC(u_1, u_2) - 1.$$

When examining the formulation of (4.17) it can be concluded that the coefficient is simply the difference between the probability of concordance and the probability of discordance. Hence, the sample version of the Kendall's tau can be easily deduced from the configuration of (4.17). Let $\{(x_{1_j}, x_{2_j})' \mid j = 1, \ldots, n\}$ represent a sample of n vectors of observations. The total unordered distinct pairs of observations in the sample is $\binom{n}{2}$. Let N_c denote the total number of concordant pairs and N_d represent the total number of discordant pairs in the sample. Consequently, the estimate of Kendall's tau coefficient for the sample of size n is defined as follows:

$$\widehat{\tau}_n = \frac{N_c - N_d}{N_c + N_d} = \frac{N_c - N_d}{\binom{n}{2}} = \frac{\sum_{1 \le i < j \le n} sign[(x_{1_j} - x_{1_i})(x_{2_j} - x_{2_i})]}{\frac{n(n-1)}{2}},$$

where

$$sign[x] = \begin{cases} 0 & x = 0, \\ +1 & x > 0, \\ -1 & x < 0. \end{cases}$$

4.1.2.2.2 Spearman's rho coefficient

An alternative and well established concordance measure is Spearman's rho coefficient. Let us consider two IID random vectors $(X_1, Y_1)'$ and $(X_2, Y_2)'$. The population version of Spearman's rho coefficient is also based on the concept of the difference between concordance and discordance and is calculated as follows:

$$S = 3\big[\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) < 0)\big]. \tag{4.18}$$

Similar to the Kendall's tau coefficient, the range of this coefficient is between -1 and 1. It can also be explicitly written in terms of the copula function. As a result of Theorem A.10, the Spearman's rho coefficient can be calculated as follows

$$S = 12 \int_{[0,1]^2} C(u_1, u_2)\, du_1\, du_2 - 3.$$

From Theorem A.10, it can be deduced that the Spearman's rho coefficient can be viewed as a non-parametric version of Pearson's correlation coefficient. Let $U_1 = F_1(X_1)$ and $U_2 = F_2(Y_1)$ denote the probability transform of the random variables X_1 and Y_1 respectively. Then it can be deduced that

$$S = \frac{\mathbb{E}(U_1U_2) - \frac{1}{4}}{\frac{1}{12}} = \frac{\mathbb{E}(U_1U_2) - \mathbb{E}(U_1)\mathbb{E}(U_2)}{\sqrt{\frac{1}{12}}\sqrt{\frac{1}{12}}}$$
$$= \frac{\mathbb{E}(U_1U_2) - \mathbb{E}(U_1)\mathbb{E}(U_2)}{\sqrt{Var(U_1)}\sqrt{Var(U_2)}} = \frac{Cov(F_1(X_1), F_2(Y_1))}{\sqrt{Var(F_1(X_1)) \cdot Var(F_2(Y_1))}}.$$

Subsequently, the remark above suggest that the estimate of Spearman's rho coefficient for a random sample should be based on the sample estimate of the Pearson's linear coefficient. Applying the Pearson's correlation coefficient estimate to the probability transforms of the random sample corresponds to applying the Pearson's linear coefficient estimate to the random sample of bivariate ranks. Consequently, let r_{i_j} represent the rank of the observation x_{i_j} for $j = \{1, 2\}$ in the sample $\{(x_{i_1}, x_{i_2}); i = 1, \ldots, n\}$ where $r_{i_1} = d$ if x_{i_1} is the d^{th} smallest in $\{x_{1_1}, \ldots, x_{n_1}\}$ and similarly for $r_{i_2} = l$ if x_{i2} is the l^{th} smallest in $\{x_{1_2}, \ldots, x_{n_2}\}$. Then the estimate of Kendall's tau coefficient for the Spearman's rho coefficient for the sample of size n is defined as follows:

$$\widehat{\rho_n} = \frac{\sum_{i=1}^{n}(r_{i_1} - \bar{r}_1)(r_{i_2} - \bar{r}_2)}{\sqrt{\sum_{i=1}^{n}(r_{i_1} - \bar{r}_1)^2}\sqrt{\sum_{i=1}^{n}(r_{i_2} - \bar{r}_2)^2}} = \frac{\sum_{i=1}^{n}\left(r_{i_1} - \frac{n+1}{2}\right)\left(r_{i_2} - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_{i_1} - \frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}\left(r_{i_2} - \frac{n+1}{2}\right)^2}}.$$

4.1.2.3 Tail Dependence Measures

The dependence measures mentioned in Sections 4.1.2.1 and 4.1.2.2 have been designed to capture the overall dependence pattern between random variables. However, the dependence structure in the upper range of the distribution may be different from that in the mid-range or lower range. As a result, the measures outlined earlier in this section, may not accurately reflect the dependence in the extreme regions of the distribution. In light of this, the purpose of this section is to introduce tail dependence measures that can help us quantify the association in the extreme parts of the distributions.

Let us consider two random variables, X_1 and X_2, with corresponding marginal distribution function, F_1 and F_2. If the following limits exist, then the lower and upper tail dependence coefficients of X_1 and X_2 are defined as follows

$$\lambda_l = \lambda_l(X_1, X_2) = \lim_{q \downarrow 0} \mathbb{P}\left(X_2 \leq F_2^{-1}(q) \,|X_1 \leq F_1^{-1}(q)\right) \tag{4.19}$$

$$\lambda_u = \lambda_u(X_1, X_2) = \lim_{q \uparrow 1} \mathbb{P}\left(X_2 > F_2^{-}-1(q) \,|X_1 > F_1^{-1}(q)\right). \tag{4.20}$$

As the definitions of tail dependence measures are derived from conditional probability, it follows that they both range from 0 to 1. The upper tail dependence coefficient quantifies the probability of the event that X_2 exceeds a certain high quantile, assuming that X_1 already exceeded that high quantile. The lower tail dependence coefficient is defined in a similar fashion. An appealing property of these measures is that they are also explicitly defined in terms of the copula function and are independent of the marginal distributions. In fact, if the univariate marginals are continuous, then

$$\begin{aligned}\lambda_l &= \lim_{q \downarrow 0} \mathbb{P}\left(X_2 \leq F_{X_2}^{-1}(q) \,|X_1 \leq F_{X_1}^{-1}(q)\right) \\ &= \lim_{q \downarrow 0} \frac{\mathbb{P}\left(X_2 \leq F_{X_2}^{-1}(q), X_1 \leq F_{X_1}^{-1}(q)\right)}{\mathbb{P}(X_1 \leq F_{X_1}^{-1}(q))} = \lim_{q \downarrow 0} \frac{C(q,q)}{q}.\end{aligned}$$

and if $\widehat{C}$ denote the survival copula then

$$\begin{aligned}\lambda_u &= \lim_{q \uparrow 1} \mathbb{P}\left(X_2 > F_{X_2}^{-1}(q) \,|X_1 > F_{X_1}^{-1}(q)\right) \\ &= \lim_{q \uparrow 1} \frac{\mathbb{P}\left(X_2 > F_{X_2}^{-1}(q), X_1 > F_{X_1}^{-1}(q)\right)}{\mathbb{P}(X_1 > F_{X_1}^{-1}(q))} \\ &= \lim_{q \uparrow 1} \frac{1 - 2q + C(q,q)}{1-q} = \lim_{q \downarrow 0} \frac{\widehat{C}(q,q)}{q}.\end{aligned}$$

If the underlying copula is in closed form, the tail measures are simple to calculate once the parameters have been determined using the techniques covered in Section 4.1.3.1. Considering that in the case study the Gumbel copula will be applied, then the calculation of (4.19) and (4.20) can be further simplified using the generator $\psi(z) = \exp\left(-z^{1/\alpha}\right)$, resulting in

$$\lambda_l = 2 \lim_{q \to \infty} \frac{\psi^{-1'}(2q)}{\psi^{-1'}(q)} = 0$$

$$\lambda_u = 2 - 2 \lim_{q \to 0} \frac{\psi^{-1'}(2q)}{\psi^{-1'}(q)} = 2 - 2^{\frac{1}{\alpha}}.$$

Since $\lambda_l = 0$, the Gumbel copula exhibits lower tail independence. However, we are interested in the upper tail dependence, and as $\lambda_u > 0$, the Gumbel copula caters for upper tail dependence with extreme events tending to occur simultaneously. Other dependence measures may be relatively straightforward to estimate, but empirical estimates of the above tail dependence measures are difficult due to their asymptotic nature and the finite sample size used in practice. There are non-parametric approaches, but these can be complex and unstable, especially with small sample sizes [124, Sec 2.6]. Nevertheless, a rough idea of the tail dependence can be obtained by plotting the scatter plot and evaluating the number of observations that exceed high quantiles.

4.1.3 Inference on the Multivariate Extreme Value Distribution

In principle, the framework of the CWBM method allows one to assess and model the vectors of extreme observations through the copula selected. Subsequently, inferences and predictions based on the fitted copula could be obtained. Hence, in the following sections, a technique to estimate the unknown parameters of the copula function and a test that assess the fit are presented. A statistic based on the copula structure will then be presented with the aim of determining the statistical properties of the next extreme observations. These techniques will also be used in the Düsseldorf case study.

4.1.3.1 Parameter Estimation

This section is devoted to the introduction of the procedure that will be used to estimate the parameters of the derived copula based multivariate model for the case study. Analogous to the univariate framework, there are several methods that provide estimates of the unknown parameters. In principle, these methods can be divided into two groups. The first group contains methods that are rank based as the estimation of the parameters is independent of the marginal distribution functions, while the second group contains some degree of dependence on the marginal distribution functions. The standard method is to consider, in a single step, the full likelihood of the marginal and

dependence parameters, which is known as the direct MLE approach. This method is considered to be statistically efficient as it is invariant to the marginal distribution, but the numerical optimization method that is required to maximize the log-likelihood function with respect to the multi-dimensional parameter, becomes increasingly computationally demanding as the dimension gets higher [26].

In setups where both the marginals and the copula are in parametric form, the inference functions for margins (IFM) method is a natural alternative to the direct MLE approach, which was developed to mitigate its drawbacks [90]. The IFM procedure is to split the direct MLE approach into two steps. When the parametric families of the marginal distributions and the copula are selected, the initial step consists of obtaining the univariate parameter estimates by using the MLE method described in Section 3.1.2.1. Then, in the second step, the parameter vector of the copula function is estimated using the univariate parameter estimates obtained in the previous step. According to Patton [115], the IFM method performs relatively better when compared to the MLE method when applied to a small sample size.

Let us assume that the block size is n. Since the blocks are disjoint, then the members of the d-variate component-wise maxima sample $\boldsymbol{M} = \{\boldsymbol{M}_i \mid 1 \leq i \leq m\}$ are IID of each other. Let $\boldsymbol{\eta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_d', \boldsymbol{\alpha}')'$ denote the full parameter vector of the following copula-based parametric model

$$G_{\boldsymbol{\eta}}(z_1, \ldots, z_d) = C_{\boldsymbol{\eta}}(G_{\boldsymbol{\theta}_1}(z_1), \ldots, G_{\boldsymbol{\theta}_d}(z_d)), \tag{4.21}$$

such that $\boldsymbol{\theta}_j$ represents the parameter vector of the j^{th} marginal distribution for $j \in \{1, \ldots, d\}$ and $\boldsymbol{\alpha}$ denotes the vector of parameters that parameterize the copula. Since, in the case study, the Gumbel copula (4.16) will be considered, then $\boldsymbol{\alpha}$ will reduce to the dependence parameter as specified in (4.16). Let us assume that the copula has a density function $\boldsymbol{c}_\eta$ and the marginal distributions have density functions g_j's, then, the joint probability density function can be decomposed as follows

$$g_{\boldsymbol{\eta}}(z_1, \ldots, z_d) = \frac{\partial^d}{\partial z_1 \cdots \partial z_d} \cdot C_\eta \left(G_{\boldsymbol{\theta}_1}(z_1), \ldots, G_{\boldsymbol{\theta}_d}(z_d)\right)$$

$$= c_{\boldsymbol{\eta}}\left(G_{\boldsymbol{\theta}_1}(z_1), \ldots, G_{\boldsymbol{\theta}_d}(z_d)\right) \cdot \prod_{j=1}^{d} g_{\boldsymbol{\theta}_j}(z_j).$$

Hence, following this decomposition, the full log-likelihood function of $\boldsymbol{M}$ is defined as follows

$$\begin{aligned} l(\boldsymbol{\eta}; \boldsymbol{M}) &= \sum_{i=1}^{m} \ln g_{\boldsymbol{\eta}}(M_{i,1}, \ldots, M_{i,d}) \\ &= \sum_{i=1}^{m} \ln c_{\boldsymbol{\eta}}\left(G_{\boldsymbol{\theta}_1}(M_{i,1}), \ldots, G_{\boldsymbol{\theta}_d}(M_{i,d})\right) + \sum_{j=1}^{d}\sum_{i=1}^{m} \ln g_{\boldsymbol{\theta}_j}(M_{i,j}) \\ &= l_D(\boldsymbol{\eta}, \boldsymbol{M}) + \sum_{j=1}^{d} l_j(\boldsymbol{\theta}_j, \boldsymbol{M}) \\ &= \underbrace{l_D(\boldsymbol{\eta}, \boldsymbol{M})}_{\text{Dependence}} + \underbrace{l_M(\boldsymbol{\theta}, \boldsymbol{M})}_{\text{Marginal}}. \end{aligned} \tag{4.22}$$

The direct MLE approach would estimate $\boldsymbol{\eta}$ by maximizing the above full likelihood over an appropriate parameter space using numerical optimization algorithms such as the L-BFGS-B or the BFGS algorithm, as most of the time there are no closed form estimators [165]. This can be time consuming, and a way to reduce computation time is by considering the partitioning of the log-likelihood function into two parts as in 4.22: the dependence component and the marginal component. Accordingly, the IFM method works as follows:

Step 1 : The log-likelihoods (l_j) of the univariate marginal distributions are separately maximized to acquire the corresponding estimates $(\widehat{\boldsymbol{\theta}}_1', \ldots, \widehat{\boldsymbol{\theta}}_d')'$.

Step 2 : The parameter estimates obtained in step 1 are plugged in the log-likelihood function ($l_D((\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_d, \boldsymbol{\alpha}), \boldsymbol{M})$) and maximized over $\boldsymbol{\alpha}$ to obtain the copula vector of estimates $\widehat{\boldsymbol{\alpha}}$.

Alternatively stated, under standard regularity conditions, the IFM estimation $\widehat{\boldsymbol{\eta}}$ is the solution of

$$\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1}, \cdots, \frac{\partial l_d}{\partial \boldsymbol{\theta}_d}, \frac{\partial l_D}{\partial \boldsymbol{\alpha}}\right)' = \mathbf{0}'.$$

Subject to the standard regularity conditions, both the MLE and IFM methods give consistent estimates that exhibit asymptotic normality. However, in the case of the IFM method, the estimation of the covariance matrix is complex, and it is usually estimated using the Godambe information matrix. Further details are available in [59, 88].

A fundamental issue for the two approaches is that the estimates are not robust against misspecification of the marginal distributions, which will often lead to an overvalue of the dependence parameter [84]. To avoid such an issue, the maximum pseudo-likelihood method is a semi-parametric extension of the IFM approach that uses the empirical marginal distribution functions in (4.22) [54]. Detail on the maximum pseudo-likelihood technique will not be provided in this dissertation since, in light of the theory discussed in Chapter 3, the parametric marginal distributions obtained for the data being studied in Section 3.1.3 are adequate for the task at hand. However, a benefit of the partitioning in the IFM method is that the estimation of the univariate marginal distribution parameters is unaffected by a misspecification of the copula function.

Recall that the copula under consideration for the case study is the one-parameter bivariate Gumbel copula, and from Section 4.1.2 it was deduced that Kendall's tau and Spearman's rho are functions of the copula parameter. Accordingly, through the inverse function of the dependence measures, the Gumbel copula parameter can be estimated by substituting the empirical estimate of Kendall's tau [57]. This is a basic method that will serve as a comparison check to the estimate provided by the IFM method. Having established that the IFM method will be used to estimate the parameters, it is essential to consider the parametric uncertainty. In view of the two step split in the procedure, it is reasonable to obtain the confidence intervals of the parameters of the univariate marginal distributions separately from the confidence intervals of the vector of parameters that parameterize the copula. The profile likelihood method, which was considered in Section 3.1.2.2 can be used in both steps to construct the confidence intervals. Consequently, given (4.22), for the univariate marginal distribution parameters, the log-likelihood functions l_j will be considered, and for the copula parameter, the dependence log-likelihood function will be considered.

4.1.3.2 Goodness-of-Fit Tests

In general, which copula family to use for a specific dataset would be known. The ambiguity lies in which member of the family to use, as in the case of the component-wise maxima method. Thus, another important step in EVT is to test whether the random sample of observations obtained is drawn from a specific distribution. Hence, in this section, a goodness-of-fit test for distribution checking will be presented. In the case study, this test can be used to ensure that the Gumbel copula (4.16) provides an adequate fit to the component-wise maxima sample.

Let us recall the component-wise maxima sample of IID random vectors $\boldsymbol{M} = \{\boldsymbol{M}_i = (M_{i,1}, \ldots, M_{i,d})' \mid 1 \leq i \leq m\}$, with a common underlying distribution G and the corresponding copula C. In principle, the copula is unknown, but it is assumed to be part of a class

$$\mathcal{C}_0 = \{C_{\boldsymbol{\eta}} \mid \boldsymbol{\eta} \in \mathcal{O}\},$$

where $\mathcal{O}$ is an open subset of $\mathbb{R}^p$ for some integer $p \geq 1$. The previous section was mainly about the direct MLE procedure and the IFM procedure. The direct MLE approach estimates the full parameter vector ($\boldsymbol{\eta}$) under the following two assumptions

$$H_0 : C \in \mathcal{C}_0 \qquad \text{and} \qquad H_0^{'} : G_{\boldsymbol{\theta}_1} \in \mathcal{F}_1, \ldots, G_{\boldsymbol{\theta}_d} \in \mathcal{F}_d,$$

where $\mathcal{F}_j = \{G_{\boldsymbol{\theta}_j} : \boldsymbol{\theta}_j \in \Gamma_j\}$ and Γ_j is an open subset of $\mathbb{R}^3$. However, as it is time consuming and complex to estimate $\boldsymbol{\eta}$ at once, it is also complicated to test the composite null hypothesis $H_0 \cap H_0^{'}$. The advantage of applying the IFM method instead of the direct MLE method is that, as the estimation procedure is split into two steps, the composite null hypothesis can also be tested in two steps.

Step 1 : The null hypothesis $H_0^{'}$ is tested separately, by considering the goodness-of-fit test on the univariate marginal distribution as described in Chapter 3. Also, the diagnostic plots will serve as a test of the fit.

Step 2 : The marginal distributions are then considered to be nuisance parameters, and the null hypothesis H_0 is tested separately by the following method.

Recall that the copula function is invariant to strictly increasing and continuous transformations of its marginal components. Subsequent to this and the first step, it is more practical to consider the component-wise probability transform of the component-wise maxima

$$\boldsymbol{U}_i = (G_{\boldsymbol{\theta}_1}(M_{i,1}), \ldots, G_{\boldsymbol{\theta}_d}(M_{i,d}))' = (U_{i,1}, \ldots, U_{i,d})',$$

as elements from a sample from the copula under consideration. Conveniently, the related empirical copula provides a summary of the information that the random vectors $\boldsymbol{U}_i$ contain, which is defined as follows

$$C_m(\boldsymbol{u}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}_{\left\{U_{i,1} \leq u_1, \ldots, U_{i,d} \leq u_d\right\}}, \tag{4.23}$$

where $\boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d$. Considering that the empirical copula is a standard estimate of the underlying copula then it is natural to test the significance of the distance between C_m and the hypothetical copula $C_{\widehat{\eta}}$ with the given estimate of the full parameter vector. Consequently, the goodness-of-fit test is based on the Cramer–von Mises empirical statistic

$$S_m = m \int_{[0,1]^d} \left| C_m(\boldsymbol{u}) - C_{\widehat{\eta}}(\boldsymbol{u}) \right|^2 \, dC_m(\boldsymbol{u}). \tag{4.24}$$

The limiting distribution of the test statistic S_m relies on the asymptotic behaviour of the integrand. Since the integrand is dependent on the family of copula under consideration and the estimate of the full parameter vector, the limiting distribution of S_m cannot be tabulated. In fact, this was expressed in the brief review done in [46]. However, the p-values can be approximated through the following parametric bootstrapping procedure.

(1) Calculate C_m according to (4.23) and estimate the full parameter vector ($\boldsymbol{\eta}$) using the IFM method described in Section 4.1.3.1.

(2) Calculate the value of S_m from the analytical expression $C_{\boldsymbol{\eta}}$. If it does not exist then proceed to use the Monte Carlo approximation. In particular, select $t \geq m$ and execute the following additional steps:

   (a) Consider $C_{\widehat{\boldsymbol{\eta}}}$ and generate a random sample $\{\boldsymbol{U}_1^*, \ldots, \boldsymbol{U}_t^*\}$.

(b) Approximate $C_{\widehat{\boldsymbol{\eta}}}$ by

$$B_t^* = \frac{1}{t}\sum_{i=1}^{t} \mathbb{I}_{\{\boldsymbol{U}_i^* \leq u\}}, \qquad \boldsymbol{u} \in [0,1]^d.$$

(c) Then the approximation of S_m follows by

$$S_m = \sum_{i=1}^{m} \left\{C_m(\boldsymbol{U}_i) - B_t^*(\boldsymbol{U}_i)\right\}^2.$$

(3) Select a large integer M and replicate the following steps for every $k \in \{1, \ldots, M\}$:

(a) Consider $C_{\widehat{\boldsymbol{\eta}}}$ and generate a random sample $\{\boldsymbol{Y}_{1,k}^*, \ldots, \boldsymbol{Y}_{m,k}^*\}$ and compute their associated rank vectors $\{\boldsymbol{R}_{1,k}^*, \ldots, \boldsymbol{R}_{m,k}^*\}$.

(b) For $i \in \{1, \ldots, m\}$ calculate $\boldsymbol{U}_{i,k}^* = \boldsymbol{R}_{i,k}^*/(m+1)$ and the corresponding empirical copula

$$C_{m,k}^*(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{\left\{\boldsymbol{U}_{i,k}^* \leq \boldsymbol{u}\right\}}, \qquad \boldsymbol{u} \in [0,1]^d.$$

Also, $\boldsymbol{\eta}_{m,k}^*$ is the estimate of $\boldsymbol{\eta}$ based on $\boldsymbol{U}_{i,k}^*$'s.

(c) If the analytic expression of $C_{\boldsymbol{\eta}}$ exists then

$$S_{m,k}^* = \sum_{i=1}^{m} \left\{C_{m,k}^*(\boldsymbol{U}_{i,k}^*) - C_{\boldsymbol{\eta}_{m,k}^*}(\boldsymbol{U}_{i,k}^*)\right\}^2.$$

Alternatively, proceed using the following steps

(a) Consider $C_{\boldsymbol{\eta}_{m,k}^*}$ and generate a random sample $\{\boldsymbol{Y}_{1,k}^{**}, \ldots, \boldsymbol{Y}_{t,k}^{**}\}$.

(b) Then $C_{\boldsymbol{\eta}_{m,k}^*}$ can be approximated as follows

$$B_{t,k}^{**} = \frac{1}{t}\sum_{i=1}^{t} \mathbb{I}_{\left\{\boldsymbol{Y}_{i,k}^{**} \leq u\right\}}, \qquad \boldsymbol{u} \in [0,1]^d.$$

and let

$$S_{m,k}^* = \sum_{i=1}^{m} \left\{C_{m,k}^*(\boldsymbol{U}_{i,k}^*) - B_{t,k}^{**}(\boldsymbol{U}_{i,k}^*)\right\}^2.$$

Finally, the p-value for the test can be approximated by the following

$$p = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}_{\{S^*_{m,k} > S_m\}}.$$

The validity of the above procedure was demonstrated in [55], and it was also shown that the test statistic S_m is consistent (if $C \notin \mathcal{C}_0$ then H_0 is rejected with probability 1 as $m \to \infty$).

4.1.3.3 Return Level and Return Period

The return level and return period concepts have already been introduced as criteria for univariate risk analysis (see Definition 2). Hence, it is logical to extend such concepts to the multivariate case. In contrast to the univariate definitions, the multivariate concepts are still under development. In fact, over the last two decades, multiple definitions related to these concepts have been proposed, each highlighting particular aspects of the dynamics of extreme events and different probability concepts. However, in this section, we will focus on the copula approach covered in [132, 135].

The following theoretical setup considers bivariate events U and V, but every argument presented here can be directly extended to the multivariate case. Let us consider a sequence $\Xi = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots\}$ that represents arbitrary IID random vectors. Each event is characterised by the joint behaviour of a pair of random variables $(U, V)' \sim \boldsymbol{C}(F_1, F_2)$ and can be represented by the set of marginal events that is defined below.

Definition 6 (Marginal Events)**.** *The following events are called marginal events*

$$E^{<}_{U,u} = \{U \leq u\}, \qquad E^{>}_{U,u} = \{U > u\},$$
$$E^{<}_{V,v} = \{V \leq v\}, \qquad E^{>}_{V,v} = \{V > v\}.$$

Typically, the event of most concern is of the type $\{\boldsymbol{E} \in \mathcal{D}\}$, where $\mathcal{D}$ contains all the realizations that are deemed hazardous.

Definition 7 (Hazard Scenario)**.** *Let $\boldsymbol{E}$ be a random d-vector describing the events of interest. A Hazard Scenario (HS) of level $\zeta \in (0, 1)$ is any Upper Set (see Definition A.12) $\mathcal{D} \subseteq \mathbb{R}^d$ such that $\mathbb{P}(\boldsymbol{E} \in \mathcal{D}) = \zeta$ holds.*

Based on the definition of HS, it can be inferred that if an event is deemed risky, those events with higher realized values are also included in the HS and considered risky. Let us fix a return period T such that the corresponding univariate return level $\boldsymbol{z}^* = (u^*, v^*)'$ will be the vector of thresholds. Accordingly, using the OR and AND mathematical operators, all possible joint behaviour depending on the combinations of the possible marginal behaviour are displayed in the following table.

Hazard Scenarios	Marginal Events	$E^{<}_{U,u^*}$	$E^{>}_{U,u^*}$
OR ($\vee$)	$E^{<}_{V,v^*}$	$\{U \leq u^*\} \vee \{V \leq v^*\}$	$\{U > u^*\} \vee \{V \leq v^*\}$
	$E^{>}_{V,v^*}$	$\{U \leq u^*\} \vee \{V > v^*\}$	$\{U > u^*\} \vee \{V > v^*\}$
AND ($\wedge$)	$E^{<}_{V,v^*}$	$\{U \leq u^*\} \wedge \{V \leq v^*\}$	$\{U > u^*\} \wedge \{V \leq v^*\}$
	$E^{>}_{V,v^*}$	$\{U \leq u^*\} \wedge \{V > v^*\}$	$\{U > u^*\} \wedge \{V > v^*\}$

Table 4.1: All Possible Combinations of Events and Hazard Scenarios.

Although the theoretical results that follow can be easily reformulated to consider all HS's listed in Table 4.1, considering the characteristics of the Düsseldorf case study tackled in this dissertation, we will focus on the following hazard scenarios:

$$\mathcal{D}^{\vee}_{\boldsymbol{z}^*} = \{U > u^*\} \vee \{V > v^*\} \quad \text{and} \quad \mathcal{D}^{\wedge}_{\boldsymbol{z}^*} = \{U > u^*\} \wedge \{V > v^*\}, \tag{4.25}$$

The first HS will be referred to as the OR HS, while the second will be referred to as the AND HS. Figures 4.1 (a) and 4.1 (b) provides visual representations of the OR and AND hazard scenarios. In practice, the selection of which HS is the most suitable to consider depends on:

- the physical attributes of the event being dealt with,
- the required design standards against specific hazardous events.

For instance, the OR HS is relevant when the failure of flooding occurs due to either the excess river discharge or the heavy precipitation, or both. On the other hand, the AND HS is relevant when the risk measures against flooding are based on when both river discharge and heavy precipitation are larger than a predefined criterion. Similar methodology to the arguments that led to Definition 2 in Chapter 3 are used to obtain

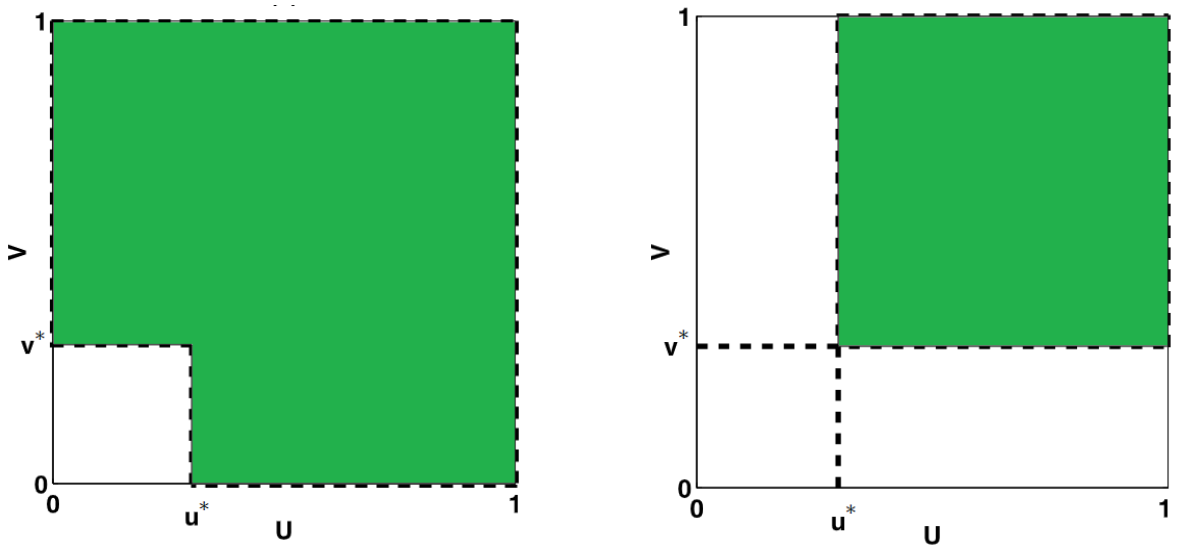


Figure 4.1: The green region in panel (a) corresponds to the OR hazard scenario and the green region in panel (b) corresponds to the AND hazard scenario.

the multivariate return period. Let $\mu > 0$ denote the average inter-arrival time between two successive hazard events. Recall from Section 4.1.1 that it is typical to choose a block size that represents a year. Then, according to [163] and references therein, μ is evaluated in the following manner

$$\mu = \frac{N_{Year}}{N_{HS}}, \tag{4.26}$$

where N_{Year} is the total amount of years under investigation and N_{HS} is the total amount of hazard events that have been recorded during the considered period. As an example, in the case of the component-wise BM method, if each block would represent a year of data, then N_{Year} and N_{HS} would be equal such that $\mu = 1$. Let S represent the first random time such that $\boldsymbol{E}_i \in \mathcal{D}$. Then:

$$S = \mu \cdot \min\{i \mid \boldsymbol{E}_i \in \mathcal{D}\}.$$

Since S/μ can be regarded as a geometric random variable, it follows that the expected value of S is

$$\mathbb{E}(S) = \frac{\mu}{\mathbb{P}(\boldsymbol{E}_i \in \mathcal{D})} = \frac{\mu}{\zeta}. \tag{4.27}$$

The expected value of S gives the multivariate primary return period, which is properly defined in Definition 8.

Definition 8 (Multivariate Primary Return Period)**.** *The multivariate primary Return Period (RP) associated with the event* $\{\boldsymbol{E} \in \mathcal{D}\}$ *is given by* (4.27)*.*

Definition 8 is unspecific with regards to the HS set as it can be adjusted to satisfy broad conditions depending on the application at hand. Subsequently, the return period under an OR HS is given by

$$T^{\vee}_{\boldsymbol{z}^*} = \frac{\mu}{\zeta^{\vee}_{\boldsymbol{z}^*}} = \frac{\mu}{\mathbb{P}(\boldsymbol{E} \in \mathcal{D}^{\vee}_{\boldsymbol{z}^*})} = \frac{\mu}{1 - \boldsymbol{C}(u^*, v^*)}, \tag{4.28}$$

and the return period under the AND HS is given by

$$T^{\wedge}_{\boldsymbol{z}^*} = \frac{\mu}{\zeta^{\wedge}_{\boldsymbol{z}^*}} = \frac{\mu}{\mathbb{P}(\boldsymbol{E} \in \mathcal{D}^{\wedge}_{\boldsymbol{z}^*})} = \frac{\mu}{1 - u^* - v^* + \boldsymbol{C}(u^*, v^*)}. \tag{4.29}$$

Figures 4.1 (a) and 4.1 (b) illustrate the regions in $[0, 1]^2$ that represent the probabilities used in the denominators of (4.28) and (4.29), respectively. As the copula function is a distribution function that is non-negative and non-decreasing, it follows that

$$\zeta^{\vee}_{\boldsymbol{z}^*} \geq \zeta^{\wedge}_{\boldsymbol{z}^*} \qquad \text{and} \qquad T^{\vee}_{\boldsymbol{z}^*} \leq T^{\wedge}_{\boldsymbol{z}^*}. \tag{4.30}$$

This arises due to the return period being shorter for more probable events. Furthermore, for every $t \in [0, 1]$, the diagonal section of a copula function $\delta(t) = (C(t, t))$ satisfies the following inequality

$$\delta(t) \leq t. \tag{4.31}$$

If the copula is from the Archimedean family (as is the logistic copula that will be used in the case study), then the inequality is strict for $t \in (0, 1)$. Given that the univariate RP T is common for both marginals, then $u^* = v^*$. As a result, it follows from (4.30) and (4.31) that

$$T^{\vee}_{\boldsymbol{z}^*} \leq T \leq T^{\wedge}_{\boldsymbol{z}^*}. \tag{4.32}$$

The inequalities are again strict inequalities if the copula used is Archimedean. As a consequence, when determining design standards in practical applications, the inequality (4.32) implies that the events in $\mathcal{D}^{\vee}_{\boldsymbol{z}^*}$ (4.25) occur more frequently than expected.

Accordingly, to have $T = T^{\vee}_{\boldsymbol{z}^*}$, the vector of thresholds have to be increased. In contrast, the events in $\mathcal{D}^{\wedge}_{\boldsymbol{z}^*}$ (4.25) occur less frequently than expected. Therefore, to have $T = T^{\wedge}_{\boldsymbol{z}^*}$, the vector of thresholds has to be decreased.

In applications, it is often desirable to identify a plausible unique event (return level) that corresponds to a given return period, as was done for the univariate case. Hence, it is logical to extend the definition to the multivariate case using the same rationale, but that is based on the HS under consideration. However, the multivariate framework is more complicated. Since there are multiple feasible couples (u, v) lying on the same isoline that yield the same prespecified return period. Typically, the considered isoline is referred to as the critical layer and it is expressed as follows

$$\mathcal{L}_t^C = \left\{ \boldsymbol{u} \in [0,1]^d \mid C(\boldsymbol{u}) = t \right\},$$

for some $t \in (0,1)$ and copula function C. Accordingly, an additional step must be taken into account to select a single design realization that is representative of a HS for the multivariate case. Various distinct approaches have been proposed, but an approach that has been widely used in the past few years is to apply a weight function to the vectors of observations that lie on the critical layer and select the point that maximise this function [135]. Hence, in this context, a design realization is obtained by using the approach defined in Definition 9

Definition 9 (Weight Function Approach)**.** *Let $w : \mathcal{L}_t^C \mapsto [0, \infty)$ be a non-negative weight function. The design realization $\sigma_w \in \mathcal{L}_t^C$ is defined as*

$$\sigma_w(t) = \underset{\boldsymbol{u} \in \mathcal{L}_t^C}{\arg\max} \left\{ w(\boldsymbol{u}) \right\}, \tag{4.33}$$

where $t \in (0,1)$ and provided that the arg max *exists and is finite.*

When adopting the approach defined in Definition 9, a unique solution might not be guaranteed. In that case, the argument of (4.33) must be subjected to further constraints by including additional information attributed to physical conditions or expert advice. Although the points on the critical layer produce the same hazard value, their co-occurrence is not equally likely. For example, a very high precipitation is unlikely

to occur together with a very low discharge. Consequently, the most frequent weight function used is based on the density function of the data restricted over the critical layer, as it gives an indication of the likelihood of the combinations.

Definition 10 (Most-Likely Weight Function)**.** *The Most-Likely design realization σ_{ML} of level t is based on the weight function:*

$$w_{ML}(\boldsymbol{u}) = c(\boldsymbol{u}) \cdot \prod_{i=1}^{d} g_i \left(G_i^{-1}(u_i) \right), \tag{4.34}$$

where $c(\cdot)$ is the density function of C, $g_i(\cdot)$ and $G_i(\cdot)$ denote the marginal density and cumulative distribution function, respectively.

The design realization σ_{ML} might be regarded as the most likely realization to be expected if a hazardous event occurs. In essence, it is expected that the designed flood controls are able to withstand events of that magnitude. Finally, it should be noted that after the above procedure is done, the original values can be retrieved using the inverse of the marginal distribution.

Chapter 3 employed the profile likelihood method to establish the confidence interval for the estimated univariate return period. However, the complexity of the multivariate theory poses a significant challenge for the uncertainty analysis of the multivariate return period [131]. To overcome this limitation, this dissertation implements the approach proposed in [136], which involves utilizing the parametric bootstrap method to identify an appropriate confidence region for the return period under consideration.

4.1.4 Düsseldorf Case Study: CWBM Analysis

In the previous sections, the theoretical framework to model multivariate extreme observations using the Component-wise Block Maxima (CWBM) approach was established. The purpose of this section is to perform the CWBM method on the Düsseldorf data. In other words, the GEV distribution is fitted to the block maxima of each univariate time series data, and then the Gumbel copula (4.16) is fitted to the probability integral transformed observations. After which, the three return period scenarios considered for

the univariate models will be computed using the weight function (4.34). This process is done for the two time horizons (P1 and P2). The results presented in this section were obtained by following the steps outlined in the flowchart shown in Figure 4.2.

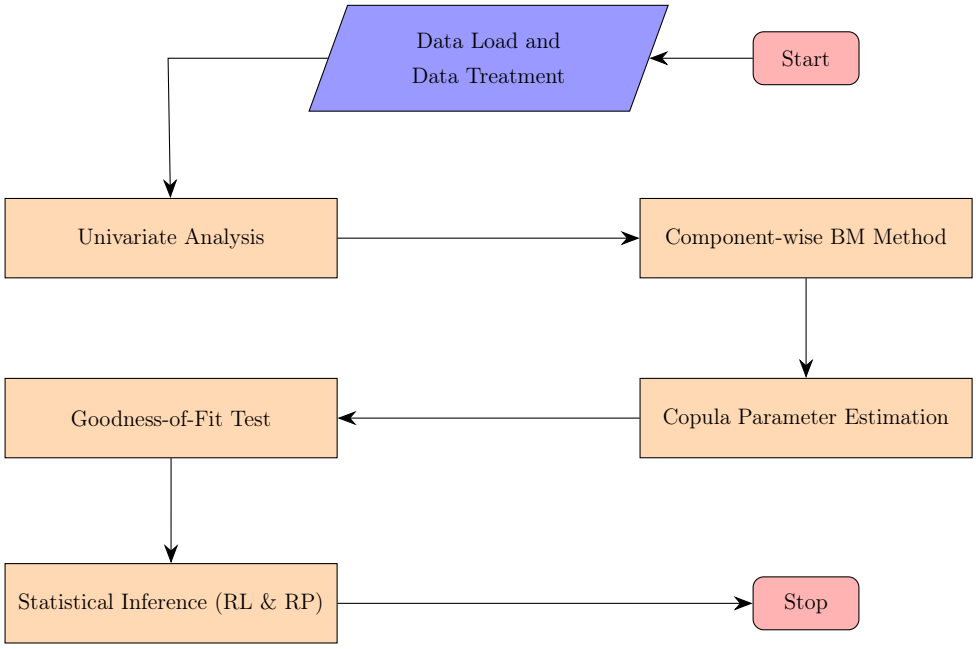


Figure 4.2: Flowchart for the CWBM method process.

This section of the main dissertation text only showcases the results of the extreme value analysis carried out for the two time series during the P1 time horizon. The same analytical approach was taken in the examination during the P2 time horizon, and the results are found in Tables B.10 and B.11. Thus, after the Düsseldorf dataset was loaded and pre-processed, the BM method was applied separately to the horological annual extreme observations of T_{MRD} and T_{MRS} time series during the P1 time horizon. As a result, the ML parameter estimates of the GEV distribution, along with their approximate 95% confidence intervals for both time series, are presented in Table B.1. These findings were discussed in more detail in Section 3.1.3. The maximum values for each hydrological year are shown in Figures 3.3 and B.39. Using equation (4.1), we generated the corresponding component-wise maximum bivariate observations, which are illustrated in Figure 4.3 as the combination of the red and green observations.

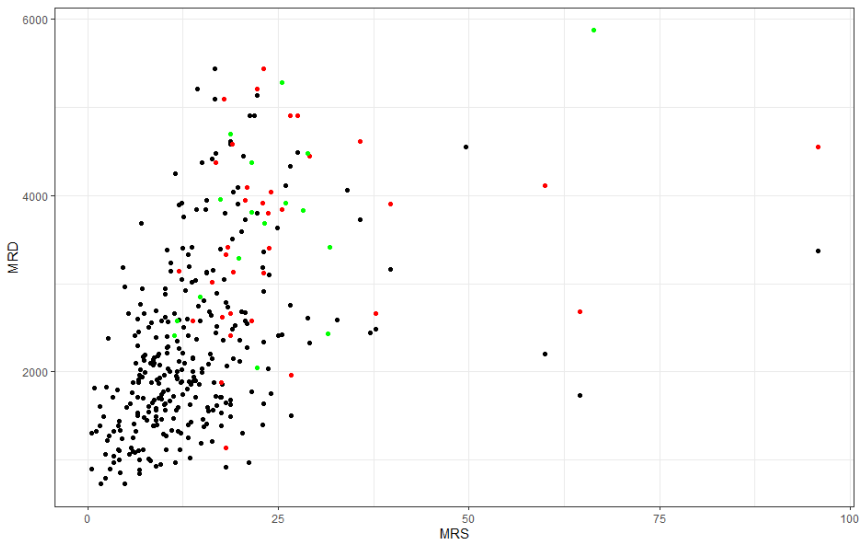


Figure 4.3: The scatter plot shows the relation between the extreme values of the time series T_{MRD} and T_{MRS} in the P1 time horizon. The red and green observations are the component-wise maxima, but the green observations correspond to actual observations.

Based on the scatter plot, 33% of the component-wise observations are actual observations. This outcome was expected since there was only one observation that was common for both panels in Table 2.10. The sampled versions of the linear and non-linear dependence measures were applied to the maxima of the two time series and all resulted in around 0.26 and the upper tail dependence measure was found to be 0.21.

The next step in the CWBM approach is to acquire the following data set $\{(u_i, v_i)' = (G_{\widehat{\boldsymbol{\theta}}_1}(M_{i,1}), G_{\widehat{\boldsymbol{\theta}}_2}(M_{i,2}))' \mid 1 \leq i \leq 52\}$, by applying the marginal GEV distribution $G_{\widehat{\boldsymbol{\theta}}_i}$ to the elements of the component-wise maxima bivariate observations, where $\widehat{\boldsymbol{\theta}}_i$ is the ML parameter estimate. These observations are shown in Figure 4.4. Subsequently, the Gumbel copula (4.16) was fitted to the new data using the IFM procedure mentioned in Section 4.1.3.1. Table 4.2 shows the IFM estimate of the single parameter of the Gumbel copula (4.16), along with their approximate 95% confidence intervals. The confidence intervals of Table 4.2 were obtained using the profile likelihood method and the corresponding profile likelihood function is shown in Figure B.114. In Section 4.1.3.1, we discussed a non-parametric method that utilizes the inverse of Kendall's tau

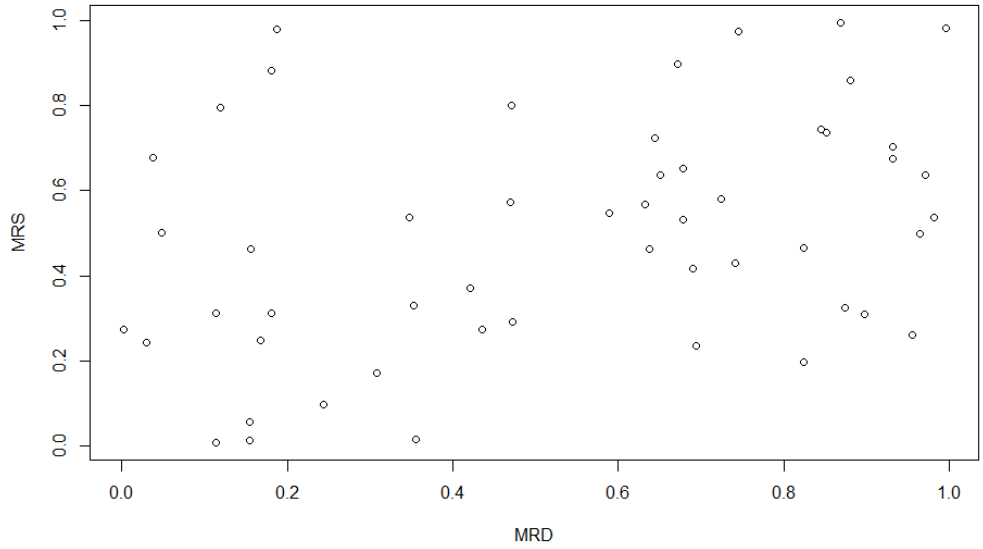


Figure 4.4: The scatter plot of the probability transformed component-wise maxima observations of the time series T_{MRD} and T_{MRS} in the P1 time horizon.

Parameter	95% lower CI	Estimated	95% upper CI
α	1.16	1.19	1.23

Table 4.2: Gumbel copula parameter estimate and corresponding confidence interval.

function to check the IFM estimate. The Kendall's tau coefficient was estimated to be $\hat{\tau} = 0.26$. Then, by using the inverse of the Kendall's tau function of the Gumbel copula, which is defined as follows:

$$\alpha = \frac{1}{1-\tau},$$

the resulted estimated parameter is 1.34, which is close to the IFM parameter estimate. The p-value of the goodness-of-fit test introduced in Section 4.1.3.2 was found to be $0.42 > 0.05$, indicating that that the sample follows the estimated Gumbel copula. Consequently, the points generated by the estimated Gumbel copula are shown in Figure 4.5, with the transformed component-wise maxima marked in red. To complete the CWBM procedure, the final step involves determining the return period that corresponds to the three return level scenarios considered for the univariate models.

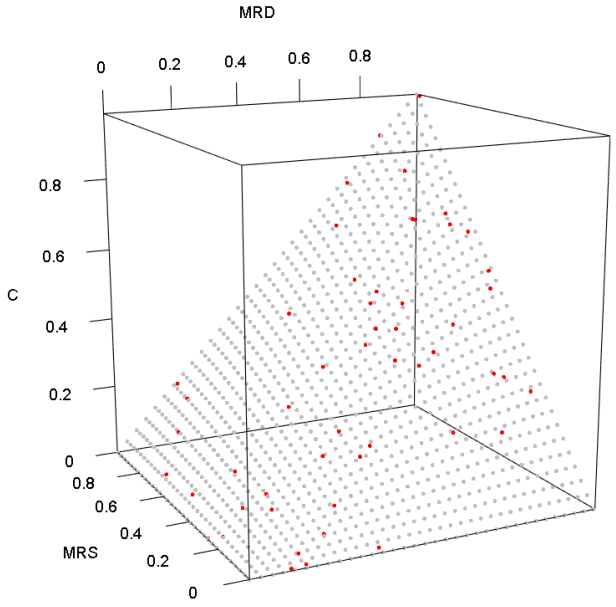


Figure 4.5: The Gumbel copula fit on the transformed component-wise maxima (in red) of the time series T_{MRD} and T_{MRS} in the P1 time horizon.

However, it is important to note that for a given threshold, the return levels produced by the equations may not necessarily be π-year bivariate events as outlined in the relations (4.32). Upon analysis of Table 4.3, it was found that on average, the return period under the OR HS is 44% smaller than π, while the return period under the AND HS is 354% larger than π, assuming that the desired return period is π. Therefore, to achieve a return period of joint events equal to π-year, the quantile $z_{1/\pi}$ must be adjusted accordingly. The last two columns in Table 4.3 provide the required quantiles to ensure that the RP under either the OR HS or the AND HS results in the desired RP when considering equations (4.28) and (4.29). Similar to the univariate case, we are interested in identifying a design realization that corresponds to the RP under consideration. Hence, the weight function defined in (4.34) was considered. The critical layers shown in Figures B.116 and B.117 were computed using the Archimedean representation (A.38). For each quantile z, a sequence from 0 to 1 was produced, with each element x substituted into $\psi(\psi^{-1}(z) - \psi^{-1}(x)) = y$. As a result, the critical layer was constructed from the pairs (x, y). However, in this section, the critical layer levels

π	$z_{1/\pi}$	$T^{\vee}_{(z_{1/\pi},z_{1/\pi})}$	$T^{\wedge}_{(z_{1/\pi},z_{1/\pi})}$	$z^{\vee}_{1/\pi}$	$z^{\wedge}_{1/\pi}$
1 in 20 year event	0.950	11.372	82.884	0.972	0.843
1 in 100 year event	0.990	55.949	470.213	0.994	0.958
1 in 200 year event	0.995	111.676	956.472	0.997	0.977

Table 4.3: The required bivariate quantiles estimates for the Gumbel copula.

considered are the quantiles listed in the final two columns of Table 4.3, such that the return levels under the two HS scenarios are the same as those considered in Chapter 3. To calculate the confidence intervals, we used a parametric bootstrap method with 1000 repetitions applied to the component-wise maxima vector. It should be highlighted that the estimates in Tables 4.3 and 4.4 are produced by the R code uploaded in the GitHub repository [157].

		Return Periods	Lower CI	Estimated	Upper CI
OR ($\vee$)	T_{MRD}	1 in 20 year event	5460.66	5651.26	5691.26
		1 in 100 year event	5899.52	6054.54	6059.48
		1 in 200 year event	6142.36	6186.18	6196.92
	T_{MRS}	1 in 20 year event	20.58	69.66	70.33
		1 in 100 year event	20.44	111.55	111.88
		1 in 200 year event	16.11	135.93	136.64
AND ($\wedge$)	T_{MRD}	1 in 20 year event	3846.27	4034.29	4279.38
		1 in 100 year event	4384.83	4537.42	5028.75
		1 in 200 year event	4588.06	4721.31	5356.89
	T_{MRS}	1 in 20 year event	19.53	22.31	24.71
		1 in 100 year event	20.61	24.52	36.16
		1 in 200 year event	20.89	24.74	47.70

Table 4.4: The Most-Likely estimates of the MRD and MRS design realizations for the three return period scenarios, according to the OR and AND hazard scenarios during the P1 time horizon, based on the Gumbel copula. The corresponding confidence intervals are also included.

As a final note, the return levels extracted from the univariate models and the multivariate models are incomparable. It stems from the fact that the estimates comes from two different classes of return period. Similarly, the return levels estimates based on different HS cannot be compared to each other as they describe different sets of events that correspond to different mechanisms of failure [139]. However, it is still important to consider and evaluate the variations in the design events that result from different modelling choices.

It can be observed that if the Düsseldorf flood protection is based on the highest values listed in Table 2.10, the multivariate scenarios presented in Table B.11 would not cause a flood under the AND HS for both time horizons. However, the multivariate scenarios under the OR HS will lead to flooding, except for the 1 in 20 year event in the two time horizon for both time series, and all the scenarios of T_{MRS} in P2 time horizon. Furthermore, it can also be highlighted that under the OR HS, the return levels are always higher than the AND HS return levels. Additionally, the values in P1 and in P2 are comparable, except for the time series T_{MRS} where the values in P1 are slightly higher than the P2 values. The design realizations for T_{MRD} are comparable to the univariate return levels (see Tables B.2, B.6 and B.8). The design realizations for T_{MRS} are always less than the univariate RLs, except for the values in P1 under the OR HS.

4.2 General Copula-based Method

A fundamental limitation of the component-wise BM method is that, when extending from the univariate to the multivariate dimension, it requires the univariate marginal distributions to belong to the GEV family and the dependence structure to be modeled by an extreme value copula (EVC). Also, according to [116], the EVC is not always able to accurately model dependence structures that exhibit dependence at extreme values and independence at the limit. As a result, in practical applications, the MEVT assumptions may not always be desirable. It is therefore reasonable to consider an alternative approach that takes into account other copula classes that do not impose the limitations of the EVC class and have a dependence framework that adequately

evaluates the tail dependence structure. To this end, this section will focus on presenting the General Copula-based (GCB) model that builds on the work of the preceding section. This model will also be applied to the Düsseldorf data as a case study.

4.2.1 Modelling Extreme Observations

In this section, the general procedure of the GCB method, which exploits the vast list of copula families and the flexibility in the copula framework to examine multivariate extreme events, will be presented. Although the GCB method can be applied with any copula family, in this dissertation, we focus on the Archimedean subclass (see Definition A.11). This class includes a wide variety of families that exhibit different characteristics and dependence structures, yet remain analytically tractable due to their simplistic and closed form (A.38). In fact, the Archimedean copulas have been frequently adopted in recent climate and hydrological publications that are concerned with the modelling of multivariate extreme observations [48, 133, 136]. Thus, in the following, we will present two Archimedean families that will be used in the Düsseldorf case study.

In this section, there is no need to cover new theory, as the necessary tools for analyzing extreme observations using the GCB approach can be obtained directly from Section 4.1. However, the only thing that still needs to be defined are the two well known Archimedean families. Recall from Definition A.11, that the Archimedean copulas have the following representation

$$C(\boldsymbol{u}) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2) + \cdots + \psi^{-1}(u_d)),$$

for some $\boldsymbol{u} \in [0,1]^d$ and for some Archimedean generator ψ (see Definition A.10). Subsequently, the two Archimedean families are defined as follows:

(1) the Clayton copula that is produced by the Archimedean generator $\psi(z) = (1 + \beta z)^{-1/\beta}$, for some $\beta \geq \frac{-1}{d-1}$ and $\beta \neq 0$. Hence, the Clayton copula is defined as follows

$$C_\beta(u_1, \ldots, u_d) = \max\left(\left[\sum_{i=1}^{d} u_i^{-\beta} - (d-1)\right]^{\frac{-1}{\beta}}, 0\right). \tag{4.35}$$

(2) the Frank copula that is produced by the generator $\psi(z) = -\gamma^{-1}\ln(1+e^{-z}(e^{-\gamma}-1))$, for some $\gamma > 0$ (for the bivariate case $\gamma \in \mathbb{R} \setminus \{0\}$). Hence, the Frank copula is defined as follows

$$C_\gamma(u_1, \ldots, u_d) = \frac{-1}{\gamma}\ln\left(1 + \frac{\prod_{i=1}^{d}(e^{-\gamma u_i}-1)}{(e^{-\gamma}-1)^{d-1}}\right). \tag{4.36}$$

The two copulas defined above are comprehensive copulas. This means that their dependence structure ranges from independence to complete dependence. The GCB procedure can now be defined, which has a structure similar to that of the CWBM method. When dealing with a sample of nm vectors of observations, the first step is to consider each time series separately such that the univariate sequence of observations obtained for a variable of interest is partitioned into m blocks, where each block is treated as if it were a sample of size n. Then, from each block, the maximum is obtained. These univariate sequences of maxima are then modeled by the three-parameter GEV distribution (3.7), using the approach described in Section 3.1.2.1. Consequently, after the vector of parameters of the GEV distribution is estimated using the univariate maximum likelihood estimation method, the maxima are probability transformed. Finally, the selected copula is fitted to the probability transformed vectors of maxima using the approach described in Section 4.1.3.1. The procedure stated in Section 4.1.3.2 is then utilized to evaluate the goodness-of-fit of the copula under consideration. Finally, the return level and the return period metrics, as defined in Section 4.1.3.3, are taken into account.

Since the GCB approach relies on the univariate BM method for its implementation, it can be inferred that the GCB method heavily depends on the BM method's assumptions. Specifically, it is crucial to find a balance between estimation bias and variance in the estimates while selecting the block size for the BM method. Similarly, for the univariate models, the block size is often determined based on the length of the available data. A typical choice for the block size, as mentioned earlier for the BM method, is a length representing one year.

4.2.2 Düsseldorf Case Study: GCB Analysis

In Section 4.2.1, the framework to model multivariate extreme observations using the General Copula-based (GCB) approach was established. The purpose of this section is to perform the GCB method on the Düsseldorf data. In other words, the GEV distribution will be fitted to the hydrological annual extremes of the time series T_{MRD} and T_{MRS} separately. Then the copulas (4.35) and (4.36) are fitted to the probability transformed observations. After which, the three return period scenarios will also be computed for each copula, under both hazard scenarios using the weight function (4.34). The results presented in this section were obtained by following the steps outlined in the flowchart shown in Figure 4.2, but the GCB method is used instead of the CWBM method. This section is split into two further sections each presenting the results of the extreme value analysis carried out for the two time series during the P1 time horizon using one of the copula presented in Section 4.2.1. The same analytical approach was taken in the examination during the P2 time horizon, and the results are found in Tables B.10, B.12 and B.13.

4.2.2.1 Clayton Copula Framework

The aim of this section is to apply the GCB method to the Düsseldorf dataset, using the Clayton copula (4.35). After loading and pre-processing the dataset, the initial steps described in Section 4.1.4, involving the fitting of univariate marginal distributions, are replicated in the same manner as in that section. Specifically, the hydrological annual extreme observations of the time series T_{MRD} and T_{MRS} in the P1 time horizon were separately fitted with the GEV distribution. The maximum values for each hydrological year are shown in Figures 3.3 and B.39 and the estimated ML parameters for both distributions, along with their approximate 95% confidence intervals for both time series, are presented in Table B.1, which were further discussed in Sections 3.1.3.

Subsequently, the vectors of component-wise maxima that are illustrated in Figure 4.3 are transformed by applying the univariate marginal GEV distribution with the corresponding estimated vector of parameters in a component-wise manner. As a result,

instead of fitting the Gumbel copula, the Clayton copula (4.35) was fitted to the transformed vectors of component-wise maxima using the IFM procedure, as detailed in Section 4.1.3.1. Table 4.5 shows the IFM estimate of the single parameter of the Clayton copula, along with their approximate 95% confidence intervals that were obtained using the profile likelihood method and the corresponding profile likelihood function is shown in Figure B.120.

Parameter	95% lower CI	Estimated	95% upper CI
β	0.65	0.69	0.73

Table 4.5: Clayton copula parameter estimate and corresponding confidence interval.

The same non-parametric estimation method used in Section 4.1.4 was applied to check the IFM estimate. By substituting the estimated Kendall's tau ($\hat{\tau} = 0.26$) into the inverse of the Kendall's tau function of the Clayton copula, which is defined as follows:

$$\beta = \frac{2\tau}{1-\tau},$$

the resulted estimated parameter is 0.69, which corresponds exactly to the IFM estimate. The p-value of the goodness-of-fit test introduced in Section 4.1.3.2 was found to be $0.66 > 0.05$, indicating that that the sample follows the estimated Clayton copula. To complete the GCB procedure, the final step involves determining the return period that corresponds to the three return level scenarios considered for the univariate models. As outlined in Section 4.1.4, the return levels produced by the equations (4.28) and (4.29) may not necessarily be π-year bivariate events as outlined in the relations (4.32). Upon analysis of Table 4.6, it was found that on average, the return period under the OR HS is 49% smaller than π, while the return period under the AND HS is 4301% larger than π, assuming that the desired return period is π. Thus, to achieve a return period of joint events equal to π, the quantile $z_{1/\pi}$ must be adjusted accordingly. The last two columns in Table 4.6 provide the required quantiles to ensure that the RP under either the OR HS or the AND HS results in the desired RP when considering equations (4.28) and (4.29). Similar to the univariate case, we are interested in identifying a design realization that corresponds to the RP under consideration. Hence, the weight function defined in (4.34) was considered. The procedure to compute the critical

π	$z_{1/\pi}$	$T^{\vee}_{(z_{1/\pi},z_{1/\pi})}$	$T^{\wedge}_{(z_{1/\pi},z_{1/\pi})}$	$z^{\vee}_{1/\pi}$	$z^{\wedge}_{1/\pi}$
1 in 20 year event	0.950	10.425	245.225	0.974	0.817
1 in 100 year event	0.990	50.422	5967.448	0.995	0.921
1 in 200 year event	0.995	100.422	23788.29	0.997	0.945

Table 4.6: The required bivariate quantiles estimates for the Clayton copula.

layers shown in Figures B.122 and B.123 follows the same steps as was done in Section 4.1.4, but using the Archimedean generator mentioned when the Clayton copula (4.35) was defined. In this section, the critical layers of interest are the quantiles listed in the final two columns of Table 4.6, such that the return levels under the two HS scenarios are the same to that were considered in Chapter 3. The parametric bootstrap method with 1000 repetitions was again applied to calculate the confidence intervals of the estimated design realizations. It should be highlighted that the estimates in Tables 4.6 and 4.7 are produced by the R code uploaded in the GitHub repository [157].

		Return Periods	Lower CI	Estimated	Upper CI
OR ($\vee$)	T_{MRD}	1 in 20 year event	5460.66	5682.65	5703.45
		1 in 100 year event	6035.52	6054.54	6124.54
		1 in 200 year event	6156.53	6186.19	6292.36
	T_{MRS}	1 in 20 year event	65.89	75.67	77.98
		1 in 100 year event	111.33	111.48	112.03
		1 in 200 year event	120.71	135.77	276.06
AND ($\wedge$)	T_{MRD}	1 in 20 year event	3453.07	3688.13	4054.59
		1 in 100 year event	3925.63	4176.42	4626.83
		1 in 200 year event	4096.03	4365.90	4822.43
	T_{MRS}	1 in 20 year event	19.81	20.99	22.33
		1 in 100 year event	22.28	23.33	24.04
		1 in 200 year event	23.14	24.09	24.50

Table 4.7: The Most-Likely estimates of the MRD and MRS design realizations for the three return period scenarios, according to the OR and AND hazard scenarios during the P1 time horizon, based on the Clayton copula. The corresponding confidence intervals are also included.

The design realizations obtained by the CWBM approach based on the Gumbel copula (refer to Table B.11), and the GCB approach based on the Clayton copula (refer to Table B.12), exhibit substantial similarity to one another. Therefore, as stated in Section 4.1.4, if the Düsseldorf flood protection is based on the highest values listed in Table 2.10, the multivariate scenarios presented in Table B.12 would not cause a flood under the AND HS for both time horizons. However, the multivariate scenarios under the OR HS will lead to flooding, except for the 1 in 20 year events for both time horizons for both time series, and all the scenarios of T_{MRS} in P2 time horizon. Furthermore, it can be highlighted that under the OR HS, the return levels are always higher than the AND HS return levels. In addition, the values in P1 and P2 are comparable, except for the time series T_{MRS} were the values in P1 are slightly higher than the P2 values under the OR HS. Regarding the design realizations for T_{MRD}, it can be observed that they are comparable to the univariate return levels presented in Tables B.2, B.6, and B.8. On the other hand, the design realizations for T_{MRS} are always less than the return levels of the univariate, except for the values in P1 under the OR HS.

4.2.2.2 Frank Copula Framework

The aim of this section is to apply the GCB method to the Düsseldorf dataset, using the Frank copula (4.36). After loading and pre-processing the dataset, the initial steps described in Sections 4.1.4 and 4.2.2.1, involving the fitting of univariate marginal distributions, are replicated in the same manner as in those sections. Specifically, the hydrological annual extreme observations of the time series T_{MRD} and T_{MRS} in the P1 time horizon were separately fitted with the GEV distribution. The maximum values for each hydrological year are shown in Figures 3.3 and B.39 and the estimated ML parameters for both distributions, along with their approximate 95% confidence intervals for both time series, are presented in Table B.1, which were further discussed in Sections 3.1.3.

Subsequently, the vectors of component-wise maxima that are illustrated in Figure 4.3 are transformed by applying the univariate marginal GEV distribution with the corresponding estimated vector of parameters in a component-wise manner. As a result,

instead of fitting the Gumbel copula or the Clayton copula, the Frank copula (4.36) was fitted to the transformed vectors of component-wise maxima using the IFM procedure, as detailed in Section 4.1.3.1. Table 4.8 shows the IFM estimate of the single parameter of the Frank copula, along with their approximate 95% confidence intervals that were obtained using the profile likelihood method and the corresponding profile likelihood function is shown in Figure B.126.

Parameter	95% lower CI	Estimated	95% upper CI
γ	2.08	2.20	2.37

Table 4.8: Frank copula parameter estimate and corresponding confidence interval.

The same non-parametric estimation method used in Sections 4.1.4 and 4.2.2.1 was used to verify the IFM estimate. The the Kendall's tau function of the Frank copula was utilized, which is defined as follows:

$$\tau = 1 - \frac{4}{\gamma}\left(1 - \frac{1}{\gamma}\int_0^{\gamma} \frac{t}{e^t - 1}\, dt\right).$$

Substituting the estimated Kendall's tau value ($\hat{\tau} = 0.26$) into the estimated inverse of the above function, the resulted parameter estimate is 2.43, which is very close to the IFM estimate. The p-value of the goodness-of-fit test introduced in Section 4.1.3.2 was found to be $0.84 > 0.05$, indicating that that the sample follows the estimated Frank copula. To complete the GCB procedure, the final step involves determining the return period that corresponds to the three return level scenarios considered for the univariate models. As outlined in Section 4.1.4, the return levels produced by the equations may not necessarily be π-year bivariate events as outlined in the relations (4.32). Upon analysis of Table 4.9, it was found that on average, the return period under the OR HS is 49% smaller than π, while the return period under the AND HS is 6263% larger than π, assuming that the desired return period is π. Therefore, to achieve a return period of joint events equal to π, the quantile $z_{1/\pi}$ must be adjusted accordingly. The last two columns in Table 4.9 provide the required quantiles to ensure that the RP under either the OR HS or the AND HS results in the desired RP when considering equations (4.28) and (4.29). Similar to the univariate case, we are interested

π	$z_{1/\pi}$	$T^{\vee}_{(z_{1/\pi},z_{1/\pi})}$	$T^{\wedge}_{(z_{1/\pi},z_{1/\pi})}$	$z^{\vee}_{1/\pi}$	$z^{\wedge}_{1/\pi}$
1 in 20 year event	0.950	10.591	179.248	0.974	0.835
1 in 100 year event	0.990	50.612	4132.027	0.995	0.932
1 in 200 year event	0.995	100.615	16351.140	0.997	0.953

Table 4.9: The required bivariate quantiles estimates for the Clayton copula.

in identifying a design realization that corresponds to the RP under consideration. Hence, the weight function defined in (4.34) was considered. The procedure to compute the critical layers shown in Figures B.128 and B.129 follows the same steps as was done in Section 4.1.4, but using the Archimedean generator mentioned when the Clayton copula (4.35) was defined. In this section, the critical layer levels of interest are the quantiles listed in the final two columns of Table 4.9, such that the return levels under the two HS scenarios are the same to that were considered in Chapter 3. The parametric bootstrap method with 1000 repetitions was again applied to calculate the confidence intervals of the estimated design realizations.

		Return Periods	Lower CI	Estimated	Upper CI
OR ($\vee$)	T_{MRD}	1 in 20 year event	5460.66	5651.26	5682.65
		1 in 100 year event	5899.52	6054.54	6056.58
		1 in 200 year event	6096.23	6186.19	6280.36
	T_{MRS}	1 in 20 year event	68.57	69.21	70.36
		1 in 100 year event	98.82	111.30	120.77
		1 in 200 year event	135.68	135.77	136.22
AND ($\wedge$)	T_{MRD}	1 in 20 year event	3292.10	4087.05	4151.17
		1 in 100 year event	3846.27	4597.58	4616.97
		1 in 200 year event	4090.04	4779.14	4785.12
	T_{MRS}	1 in 20 year event	17.48	23.58	24.95
		1 in 100 year event	18.37	26.23	26.83
		1 in 200 year event	18.46	27.03	27.12

Table 4.10: The Most-Likely estimates of the MRD and MRS design realizations for the three return period scenarios, according to the OR and AND hazard scenarios during the P1 time horizon, based on the Frank copula. The corresponding confidence intervals are also included.

The CWBM approach based on the Gumbel copula (refer to Table B.11), the GCB approach based on the Clayton copula (refer to Table B.12), and the GCB approach based on the Frank copula (refer to Table B.13), all show significant similarities in their design realizations. Therefore, as stated in Section 4.2.2.1, if the Düsseldorf flood protection is based on the highest values listed in Table 2.10, the multivariate scenarios presented in Table B.13 would not cause a flood under the AND HS for both time horizons. However, the multivariate scenarios under the OR HS will lead to flooding, except for the 1 in 20 year events for both time horizons and for both time series and all the scenarios of T_{MRS} in P2 time horizon. Furthermore, it can be highlighted that under the OR HS, the return levels are always higher than the AND HS return levels. In addition, the values in P1 and P2 are comparable, except for the time series T_{MRS} were the values in P1 are slightly higher than the P2 values. Regarding the design realizations for T_{MRD}, it can be observed that they are comparable to the univariate return levels presented in Tables B.2, B.6, and B.8. On the other hand, the design realizations for T_{MRS} are always less than the return levels of the univariate.

The AIC and BIC values presented in Table B.10 indicate that the Frank copula provided the best fit during the P1 time horizon. However, for the P2 time horizon, it was the Clayton copula that provided the best fit. Figures B.132 to B.135 display a graphical representation of the comparison between the flood scenarios listed in Tables B.11, B.12 and B.13 for the three multivariate models and for each HS. As a result, it can be deduced that for the time series T_{MRD} under the OR HS, the design realizations by the different methods are all on the same level for each flood scenario for both time horizons. The same can be said when the flood scenarios were run under the AND HS, but in this case, the design realizations during the P1 time horizon are significantly higher than the values during the P2 time horizon. Conversely, the design realizations from the three methods for the time series T_{MRS} under the AND HS have the same level for each flood scenario for both time horizons. However, there are some significant variations between P1 and P2 values when considering the OR HS. In some instances, the choice of the model did not make a difference, but in some instances it did. In practical situations, the choice of model and the HS selected need to be carefully considered.

Chapter 5

Conclusion

The main objective of the dissertation was to review some of the most commonly used techniques in Extreme Value Theory (EVT) and apply them to the meteorological observations extracted from two stations in Düsseldorf. This chapter will commence with Section 5.1, where the work accomplished in the previous chapters will be highlighted, accompanied by a summary of the main findings. Subsequently, in Section 5.2, the limitations encountered during the compilation of this dissertation will be emphasized, and recommendations for future research will be presented.

5.1 Dissertation Summary

Extreme value analysis is a widely used technique in the fields of meteorology and climatology to model rare and extreme events that could have significant impacts on society's safety. Through these models, risk scenarios can be constructed to provide valuable insights into whether current risk mitigation measures are appropriate. In fact, the aim of this dissertation was to provide insights on the flood risk that is currently present in Düsseldorf. This dissertation begins with an in-depth analysis of the procedure used to obtain the two final target time series from the available daily observations that span over 52 years. The first target time series is composed of monthly mean river discharge observations that were extended by applying the rating curve to the water level observations. The second target time series is composed of the monthly mean of

the combined impact of snow melt and precipitation excess, after a rolling sum over a 7-day period was applied. In order to reduce the seasonality of the target time series, the yearly time horizon was split into semi-yearly time horizons (P1 and P2). Thus, the analysis was then applied to the two target time series for each time period.

Models based on EVT are more flexible and better at modeling extreme events, as they focus on only extreme observations. In Chapter 3, three univariate models that are based on EVT asymptotic results were constructed. The first model was the Block Maxima (BM) approach, for which the time series is split into disjoint blocks such that the GEV distribution is fitted to the maximum of each block. The second approach is an extension of the BM method, for which the K Largest Order Statistics (KLOS) method splits the time series in the same manner, but the GEV_k distribution is fitted to the top k extreme order statistics from each block. The Peak over Threshold (POT) method is the final approach considered for the univariate case. This method takes a different approach by fitting the excess of extreme observations over a high threshold with the GP distribution.

The final step for each model is that of deriving the return level and return period metrics, as they are used to determine design standards. Each of the three univariate models was applied to the two target time series for each time period. The MLE approach was used to estimate the model parameters. Consequently, for each method, the extreme scenarios with $20, 100, 200$ year return period were considered and compared to each other. It was found that in most cases the BM method has the most adequate fit (as measured by AIC and BIC). In Chapter 4, two multivariate models were considered that take into account the inter-relationship and the co-occurrence of multiple events. However, prior to establishing the multivariate models, two hazard scenarios had to be established in order to identify a way to define a multivariate extreme observation. Subsequent to this, the first method was extended from the BM method to produce the Component-wise Block Maxima (CWBM) method. The selection of the extreme vector of observations is done separately for each time series in each block, as was done for the BM method, but as opposed to the univariate model, there is no definite distribution

that could be fitted to the multivariate block maxima. Consequently, copula functions were considered and the Gumbel copula was fitted to the multivariate block maxima of the two target time series for each period using the IFM approach. In the context of multivariate models, determining the return period and return level metrics is not as straightforward as it is for univariate models. As a result, the concept of a weight function had to be established in order to identify a way to define a design realization.

Then the General Copula-based (GCB) method was introduced as a generalization of the CWBM method by considering other dependence structures. In fact, the Clayton copula and the Frank copula that forms part of the Archimedean copula family, were considered. As a result, through the AIC and BIC measures, it was determined that the Frank copula provided the best fit during the P1 time horizon and the Clayton copula provided the best fit for the P2 time horizon. Consequently, for each method, as for the univariate case, the extreme scenarios with a $20, 100, 200$ year return period were considered and compared to each other. It was found that the multivariate results for both models were close to the univariate results. This was attributed to the fact that the dependence structure between the extremes of the time series is weak.

5.2 Limitations and Recommendations

One may take a number of different routes to build on the work this dissertation has covered, but in the following section, we will outline the most popular research directions. The branch of EVT carries a number of challenges and the models presented in the previous chapters are the starting point for extreme value analysis. In fact, there are prospective adjustments to the main IID condition that could provide better modeling of the underlying data.

The most natural generalization of the univariate models is to relax the independence condition and assume that the underlying sequence of observations is stationary, which is a realistic condition when considering physical processes [26]. To provide a general characterization of this model, it is necessary to specify a constraint either on the

long-range dependence or the minimum separation interval. For example, this can be achieved by selecting extreme events from different clusters to reduce temporal dependence. Given that the data considered in this dissertation was meteorological in nature, the long term anthropogenic effects and climate variations violate the stationarity assumption. Hence, the next step in generalization is to consider models that cater for non-stationary observations. For these types of instances, the models are based on the standard extreme models, but variations through time can be captured by substituting the parameters of the distributions with time-varying parameters [37]. This concept is also introduced in the multivariate case, where the copulas considered can be taken to be in terms of time-varying parameters [121].

Another innovative approach is to introduce a flexible dependence structure between multiple variables by considering the Pair Copula Constructions (PCC) [104]. The PCC allows for pairs of random variables to have different dependence frameworks, by enabling the multivariate distribution to be decomposed into a cascade of independent bivariate copulas. Other crucial areas of advancement include methods that translate modelling uncertainties into the confidence intervals of the multivariate design realizations, different hazard scenarios (such as the Kendall and Survival Kendall hazard scenarios), and weight functions that account for physical properties when selecting a design realization. It is hoped that further developments in R packages will also help in implementing these improvements.

Appendix A

Extra Material: Theoretical Results

A.1 Main Theoretical Proofs

Proof of Theorem 1. Let us consider (3.5) and apply Lemma A.1 to acquire an equivalent and more flexible condition. By comparing (3.5) to $(A.25)$ and if $u_n = a_n z + b_n$ and $G(z) = e^{-\tau}$, it follows that

$$\lim_{n\to\infty} n(1 - D(a_n z + b_n)) = -\ln(G(z)), \tag{A.1}$$

or else

$$\lim_{n\to\infty} \frac{1}{n(1 - D(a_n z + b_n))} = \frac{1}{-\ln(G(z))}. \tag{A.2}$$

Let U be the left-continuous inverse (see Definition A.2) of $1/(1 - D)$. In other words, let $V(x) = 1/(1 - D(x))$, then by definition

$$U(y) = V^{\leftarrow}(y) = \inf\{w : V(w) \geq y\}. \tag{A.3}$$

It is evident that in this case the equation $(A.3)$ coincides with the quantile function $Q(y)$ such that the left inverse is equivalent to $U(y) = D^{\leftarrow}\left(1 - \frac{1}{y}\right)$. Since, at the continuity points

$$U(y) = Q\left(1 - \frac{1}{y}\right) = z \;\; \Leftrightarrow \;\; D(z) = 1 - \frac{1}{y},$$

as y increases, $U(y)$ tends to z^*. Hence, for any $n \in \mathbb{N}$,

$$U(nz) = \inf\left\{y : \frac{1}{n(1 - D(y))} \geq z\right\}.$$

Hence, linear normalizing the left hand side will result into the following,

$$\begin{aligned}\frac{U(nz)-b_n}{a_n} &= \inf\left\{\frac{y-b_n}{a_n} : \frac{1}{n(1-D(y))} \geq z\right\} \\ &= \inf\left\{x : \frac{1}{n(1-D(a_n x+b_n))} \geq z\right\}\end{aligned}$$

Considering Lemma $A.2$ and equation $(A.2)$, for positive x it follows that

$$\begin{aligned}\lim_{n\to\infty}\frac{U(nz)-b_n}{a_n} &= \lim_{n\to\infty}\inf\left\{x : \frac{1}{n(1-D(a_n x+b_n))} \geq z\right\} \\ &= \inf\left\{x : \frac{1}{-\ln(G(x))} \geq z\right\} \\ &= \inf\left\{x : \ln(G(x)) \geq \frac{-1}{z}\right\} \\ &= \inf\left\{x : G(x) \geq e^{\frac{-1}{z}}\right\} \\ &= G^{\leftarrow}\left(e^{\frac{-1}{z}}\right) := T(z). \end{aligned} \tag{A.4}$$

Suppose that $z=1$ is a continuity point of the function T. Hence, for any continuity point $z>0$,

$$\begin{aligned}E(z) = T(z) - T(1) &= \lim_{t\to\infty}\frac{U(tz)-b(t)}{a(t)} - \frac{U(t)-b(t)}{a(t)} \\ &= \lim_{t\to\infty}\frac{U(tz)-U(t)}{a(t)}. \end{aligned} \tag{A.5}$$

Suppose that $\xi > 0$, then

$$\frac{U(tzy)-U(t)}{a(t)} = \frac{U(tzy)-U(ty)}{a(ty)}\frac{a(ty)}{a(t)} + \frac{U(ty)-U(t)}{a(t)}. \tag{A.6}$$

Let us claim that the limits of

$$\lim_{t\to\infty}\frac{U(ty)-U(t)}{a(t)} \tag{A.7}$$

and

$$\lim_{t\to\infty}\frac{a(ty)}{a(t)} \tag{A.8}$$

exist. Furthermore, suppose that the claim is not true, then there are more than one limit point. Hence, let A_i and B_i be the distinct limit points of $(A.7)$ and $(A.8)$

respectively for $i \in \{1, 2\}$. Ergo, for $i \in \{1, 2\}$ the following equation is induced from the relation $(A.5)$ and $(A.6)$:

$$E(zy) = E(z)A_i + B_i, \tag{A.9}$$

for all continuity points of $E(\cdot)$ and $E(\cdot y)$. Since, the function E is left-continuous, take a sequence of continuity points z_n such that it converges to an arbitrary point z $(z_n \uparrow z)$ then by definition $E(z_n) \to E(z)$ and $E(z_n y) \to E(zy)$. Therefore, it can be deduced that the relation $(A.9)$ holds for all positive values of x and y. Furthermore, from $(A.9)$ the following is acquired

$$E(z)A_1 + B_1 = E(zy) = E(z)A_2 + B_2,$$

and so

$$E(z)\left(A_1 - A_2\right) = \left(B_2 - B_1\right).$$

for all $z > 0$. Moreover, since the sequence M_n has been normalized in order to obtain a non-degenerate distribution G, then E must be a non-constant function. Thus, $A_1 = A_2$ and $B_1 = B_2$ and hence, the limits of $(A.7)$ and $(A.8)$ exist. Hence,

$$A(y) = \lim_{t\to\infty} \frac{a(ty)}{a(t)}$$

exists for $y > 0$. On the other hand, for $z, y > 0$,

$$E(zy) = E(z)A(y) + E(y).$$

Let $s = \ln(z)$, $t = \ln(y)$ and $K(z) = E(e^z)$. Then,

$$\begin{aligned} K(t+s) &= E(e^{t+s}) \\ &= E(e^t e^s) \\ &= E(e^s)A(e^t) + E(e^t) \\ &= K(s)A(e^t) + K(t). \end{aligned} \tag{A.10}$$

Since, $K(0) = E(1) = D(1) - D(1) = 0$, then

$$\frac{K(t+s) - K(t)}{s} = \frac{K(s) - K(0)}{s} A(e^t). \tag{A.11}$$

Moreover, by definition, T is monotonically increasing and this implies that K is also monotonically increasing. Thus, as a necessary condition for a monotonic increasing function, K must be differentiable at least at one point t. Considering equation $(A.6)$ it can be deduced that K is differentiable everywhere. In addition, from $(A.6)$ and the definition of differentiability, the following is obtained

$$K'(t) = K'(0)A(e^t). \tag{A.12}$$

Moreover, $K'(0) \neq 0$ thus let $Q(t) = K(t)/K'(0)$. Consequently, $Q(0) = 0$ and $Q'(0) = K'(0)/K'(0) = 1$. Consider, the expressions $(A.10)$ and $(A.12)$, then

$$Q(t+s) - Q(t) = Q(s)A(e^t) = Q(s)\frac{K'(t)}{K'(0)} = Q(s)Q'(t). \tag{A.13}$$

Similarly, the same expression with s and t interchanged can be obtained with the same procedure. Therefore, $Q(s+t) - Q(s) = Q(t)Q'(s)$ and if we subtract this expression by $(A.13)$, the following is obtained

$$Q(s) - Q(t) = Q(s)Q'(t) - Q(t)Q'(s)$$

and hence

$$Q(t)\frac{Q'(s) - 1}{s} = \frac{Q(s)}{s}\left(Q'(t) - 1\right).$$

Furthermore, as $s \to 0$, the following differential equation is obtained

$$Q(t)Q''(0) = Q'(t) - 1, \tag{A.14}$$

with conditions $Q'(0) = 1$ and $Q(0) = 0$ and in order to solve the differential equation $(A.14)$ we differentiate both sides with respect to t. Hence,

$$Q''(t) = Q'(t)Q''(0).$$

Which it is simplified as follows

$$\left(\ln\left(Q'\right)\right)'(t) = \frac{Q''(t)}{Q'(t)} = Q''(0) := \xi \in \mathbb{R},$$

for all t. Since $Q'(0) = 1$, then

$$\ln\left(Q'(t)\right) = \int_0^t \left(\ln\left(Q'\right)\right)'(s)\, ds$$

$$= \int_0^t \xi \, ds$$
$$\Rightarrow Q'(t) = e^{\xi t}$$

and since $Q(0) = 0$ then it follows that

$$Q(t) = \int_0^t Q'(s) \, ds = \int_0^t e^{\xi s} \, ds = \frac{e^{\xi t} - 1}{\xi}$$

Which this leads to

$$K(t) = K'(0)\frac{e^{\xi t} - 1}{\xi}.$$

Subsequently, $H(t) = E(e^t) = T(e^t) - T(1)$ and with a change of variable, the following is obtained

$$T(t) = T(1) + K'(0)\frac{t^{\xi} - 1}{\xi}.$$

Therefore, the following identities are obtained

$$T^{\leftarrow}(z) = \left(1 + \xi\frac{z - T(1)}{K'(0)}\right)^{\frac{1}{\xi}} \tag{A.15}$$

and since $T(z) = G^{\leftarrow}\left(e^{\frac{-1}{z}}\right)$, then

$$T^{\leftarrow}(z) = \frac{1}{-\ln G(z)}. \tag{A.16}$$

Combining $(A.15)$ and $(A.16)$ and letting $a = K'(0)$ and $b = T(1)$ we get:

$$\frac{1}{-\ln\left(G(z)\right)} = \left(1 + \xi\frac{z - T(1)}{K'(0)}\right)^{\frac{1}{\xi}}$$
$$G(z) = \exp\left(1 + \xi\frac{z - T(1)}{K'(0)}\right)^{\frac{-1}{\xi}}$$
$$G(az + b) = \exp\left(1 + \xi z\right)^{\frac{-1}{\xi}}.$$

Finally, it was assumed at the beginning that $z = 1$ is a continuity point, if it is not a continuity point the same proof would be repeated with $U(tz_0)$ instead of $U(t)$ in $(A.5)$ where z_0 is a continuity point of T. □

Proof of Theorem 2. Let us consider the definition of $M_n^{(k)}$ and S_n with level ($u_n = a_n z + b_n$) from Theorem $A.6$ in Appendix A. Then it clearly follows that

$$M_n^{(k)} \leq u_n \qquad \Longleftrightarrow \qquad S_n = \sum_{i=1}^{n} \mathbb{I}_{\{\xi_i > u_n\}} < k,$$

which implies that the events $\{M_n^{(k)} \leq u_n\}$ and $\{S_n < k\}$ are identical and hence $\mathbb{P}\left\{M_n^{(k)} \leq u_n\right\} = \mathbb{P}\{S_n < k\}$. Thus, Theorem $A.6$ can be restated in terms of the k^{th} largest order statistics. Subsequently, if (3.12) is assumed to hold, then

$$\mathbb{P}\{M_n^{(k)} \leq u_n\} \to e^{-\tau} \sum_{s=0}^{k-1} \frac{\tau^s}{s\,!} \tag{A.17}$$

holds with $k = 1$ and $\tau = -ln(G(z))$. However, the converse part of the restated Theorem $A.6$ implies that $(A.17)$ holds for all k. Hence, it can be concluded that (3.13) holds. Conversely, if (3.14) holds for some fixed k and z is given then as the right hand side of $(A.17)$ is continuously decreasing from 1 to 0 as τ increases, there exists $\tau \in [0, \infty]$ such that

$$H(z) = e^{-\tau} \sum_{s=0}^{k-1} \frac{\tau^s}{s\,!}.$$

Consequently, as $(A.17)$ holds for this fixed k, then by employing the converse argument of the restated Theorem $A.6$ it follows that $(A.17)$ holds for all k. Thus, it also holds for $k = 1$ which implies (3.12) with $G(z) = e^{-\tau}$. Hence, it can be concluded that (3.13) follows from the above arguments. □

Proof of Theorem 3. This proof follows the same reasoning as the proof of Theorem 2. Accordingly, let $u_n^{(i)} = a_n z^{(i)} + b_n$ and consider the definition of $M_n^{(i)}$ and $S_n^{(i)}$ from Theorem $A.7$ (see Appendix A.2) then it can be deduced that

$$\mathbb{P}\left\{M_n^{(1)} \leq u_n^{(1)}, \ldots, M_n^{(k)} \leq u_n^{(k)}\right\} = \mathbb{P}\left\{S_n^{(1)} = 0, S_n^{(2)} \leq 1, \cdots, S_n^{(k)} \leq k-1\right\}.$$

Subsequently, if (3.16) is assumed to hold, then by considering Theorem $A.7$ and Lemma $A.1$, it can be concluded that

$$\lim_{n\to\infty} \mathbb{P}\left\{M_n^{(1)} \leq u_n^{(1)}, \ldots, M_n^{(k)} \leq u_n^{(k)}\right\}$$

$$= \lim_{n\to\infty} \mathbb{P}\left\{S_n^{(1)} = 0, S_n^{(2)} \leq 1, \cdots, S_n^{(k)} \leq k-1\right\}$$
$$= \lim_{n\to\infty} \sum_{s_1=0}^{0} \sum_{s_2=0}^{1} \cdots \sum_{s_k=0}^{k-1} \mathbb{P}\left\{S_n^{(1)} = s_1, S_n^{(2)} = s_2, \cdots, S_n^{(k)} = s_k\right\}.$$

Let $y_i = s_i - s_{i-1}$ and $y_1 = s_1$ then it follows

$$\lim_{n\to\infty} \mathbb{P}\left\{M_n^{(1)} \leq u_n^{(1)}, \ldots, M_n^{(k)} \leq u_n^{(k)}\right\}$$
$$= \sum_{y_1=0}^{0} \sum_{y_2=0}^{1} \sum_{y_3=0}^{2-y_2} \cdots \sum_{y_k=0}^{k-1-y_1-\cdots-y_{k-1}}$$
$$\lim_{n\to\infty} \mathbb{P}\left\{S_n^{(1)} = y_1, S_n^{(2)} = y_1 + y_2, \cdots, S_n^{(k)} = y_1 + \cdots + y_k\right\}$$
$$= \sum_{y_2=0}^{1} \sum_{y_3=0}^{2-y_2} \cdots \sum_{y_k=0}^{k-1-y_1-\cdots-y_{k-1}} \frac{(\gamma_2 - \gamma_1)^{y_2}}{y_2\,!} \cdots \frac{(\gamma_k - \gamma_{k-1})^{y_k}}{y_k\,!} \cdot e^{-\gamma_k}.$$

By changing the variables again the result (3.18) is obtained. □

Proof of Theorem 4. In the case of $\xi = 0$ this is already shown by the first statement of Theorem A.4 found in Appendix A. If $D \in \mathcal{D}(G_\xi)$ holds for $\xi > 0$ then by the second statement of Theorem A.4 and Definition A.4 in Appendix A it follows that $\bar{D} \in \mathcal{R}_{-1/\xi}$. Considering Theorem A.3 in Appendix A, it follows that for some $t > 0$,

$$\bar{D}(z) = c(z) \exp\left\{-\int_t^z \frac{1}{a(u)}\, du\right\}, \qquad t < z < \infty,$$

where $c(z) \to c > 0$ and $a(z)/z \to \xi$ as $z \to \infty$ locally uniformly. Subsequently,

$$\lim_{u\to\infty} \frac{\bar{D}(u + z\sigma(u))}{\bar{D}(u)} = (1 + \xi z)^{-\frac{1}{\xi}},$$

which is (3.23). Conversely, if (3.23) holds for $\xi > 0$ then let

$$k_n = (1/\bar{D})^{\leftarrow}(n) = U(n).$$

which implies that

$$1/\bar{D}(k_n) \sim n,$$

and with $u = k_n$ in (3.23) it follows that

$$(1 + \xi z)^{-\frac{1}{\xi}} = \lim_{n\to\infty} \frac{\bar{D}(k_n + z\sigma(k_n))}{\bar{D}(k_n)}$$

$$= \lim_{n\to\infty} n\bar{D}(k_n + z\sigma(k_n)).$$

Since, $-\ln(G_\xi(z)) = (1+\xi z)^{-\frac{1}{\xi}}$, then by the equivalence relation $(A.1)$ it follows that $H \in \mathcal{D}(G_\xi)$. The same procedure could be applied to show the above equivalence when $\xi < 0$. □

Proof of Theorem 5. Let ξ be an independent copy of the sequence then according to Theorem 4 it follows that

$$\begin{aligned}
\lim_{u\uparrow z^*} \bar{D}_u(z\sigma(u)) &= \lim_{u\uparrow z^*} \mathbb{P}\left\{Z - u > z\sigma(u) \mid Z > u\right\} \\
&= \lim_{u\uparrow z^*} \mathbb{P}\left\{\frac{Z-u}{\sigma(u)} > z \,\middle|\, Z > u\right\} \\
&= \lim_{u\uparrow z^*} \frac{\mathbb{P}\left\{Z > u + z\sigma(u)\right\}}{\mathbb{P}\left\{Z > u\right\}} \\
&= \lim_{u\uparrow z^*} \frac{1 - H(u + z\sigma(u))}{1 - H(u)} \\
&= \begin{cases} (1+\xi z)^{\frac{-1}{\xi}} & \xi \neq 0, \\ e^{-z} & \xi = 0. \end{cases} \\
&= \bar{H}_\xi(z)
\end{aligned}$$

Therefore, $D \in \mathcal{D}(G_\xi)$ if and only if

$$\lim_{u\uparrow z^*} \left| D_u(z) - H_\xi\left(\frac{z}{\sigma(u)}\right)\right| = 0,$$

However, the GP distribution is continuous and hence the uniform convergence follows from Lemma $A.3$ and Theorem $A.5$ in Appendix A. □

Proof of Theorem 6. Let us assume that the appropriate normalizing sequences of constants exist such that $D_* \in \mathcal{D}(G_*)$. Then, by Proposition $A.1$ (i), Proposition $A.3$ and Definition $A.9$ it can be inferred that for some $k > 0$

$$G_*^k(k\boldsymbol{z}) = G_*(\boldsymbol{z}). \tag{A.18}$$

It is clear that G_*^k is a distribution function for every $k > 0$ and by Definition $A.8$ it can be deduced that G_* is also max-id. According to Proposition $A.2$ there is an

exponent measure μ_* that characterizes the dependence structure of the extreme transformed random vectors. As the marginal distribution concentrates on $[0,\infty)$ then it is reasonable to take $\boldsymbol{l} = \mathbf{0}$ such that μ_* is concentrated on $E = \mathbb{R}^d_+ \setminus \{\mathbf{0}\}$ such that for some set $P_{\boldsymbol{z}} = \{\boldsymbol{x} \in \mathbb{R}^d \mid 0 \leq x_i \leq z_i, 1 \leq i \leq d\}$ we have

$$V_*(\boldsymbol{z} = (z_1, \ldots, z_d)) = -\ln(G_*(\boldsymbol{z})) = \mu_*\left(E \setminus P_{\boldsymbol{z}}\right),$$

for some $\boldsymbol{z} \in [0,\infty]^d$ and $V_*(\boldsymbol{z}) = \infty$ otherwise. Considering that the margins of G_* are $G_{1,1,1}$ then it follows that

$$V_*(\infty, \ldots, z_j, \ldots, \infty) = \mu_*\left(\left\{\boldsymbol{x} \in [0,\infty]^d \mid x_j > z_j\right\}\right) = z_j^{-1}$$

for all $j = 1, \ldots, d$ and $0 < z_j < \infty$. Furthermore, the exponent measures μ and μ_* of G and G_* respectively are connected as follows. If $\boldsymbol{x} \in [q,\infty]^d$ and $\boldsymbol{z} \in [0,\infty]^d$ are related by $z_j = -1/\ln(G_j(x_j))$ then

$$\begin{aligned}\mu\left([q,\infty]^d \setminus [q,x]^d\right) &= \ln(G(\boldsymbol{x})) \\ &= \ln(G_*(\boldsymbol{z})) = \mu_*([0,\infty]^d \setminus P_{\boldsymbol{z}}).\end{aligned}$$

Despite that the characterizing property of μ_* is the homogeneity property derived from $(A.18)$ and expressed as follows

$$\mu_*(E \setminus P_{\boldsymbol{z}}) = k\mu_*(E \setminus k \cdot P_{\boldsymbol{z}})$$

where $\boldsymbol{z} > \mathbf{0}$ and as

$$E \setminus k \cdot P_{\boldsymbol{z}} = k(E \setminus P_{\boldsymbol{z}})$$

the property can be represented as follows

$$\mu_*(E \setminus P_{\boldsymbol{z}}) = k\mu_*(k(E \setminus P_{\boldsymbol{z}})). \tag{A.19}$$

In fact, using measure-theoretic arguments on $(A.19)$ the relation can be extended to hold for all Borel subsets (B) of E. The homogeneity property of the exponent measure suggests a coordinate transformation to pseudo-polar coordinates. Accordingly, let us consider an arbitrary norm on $\mathbb{R}^d$. For instance the most common choices are the L_d-norm ($\|\boldsymbol{z}\| = (|z_1|^d + \cdots + |z_d|^d)^{1/d}$ for $1 \leq d < \infty$) or the max-norm ($\|\boldsymbol{z}\| =$

$\max(|z_1|,\ldots,|z_d|)$). Let $\mathbb{S} = \{\boldsymbol{w} \in E \mid \|\boldsymbol{w}\| = 1\}$ denoting the unit sphere in E with respect to the selected norm. Since, all norms are equivalent on $\mathbb{R}^d$ then $\mathbb{S}$ is bounded away from $\mathbf{0}$ and thus it is a compact set. Let us define the mapping $T : E \mapsto (0,\infty] \times \mathbb{S}$ as follows

$$T(\boldsymbol{z}) = (r, \boldsymbol{w}), \tag{A.20}$$

where $r = \|\boldsymbol{z}\|$ is the radial part and $\boldsymbol{w} = \boldsymbol{z}/\|\boldsymbol{z}\|$ is the angular part of $\boldsymbol{z}$. As a matter of fact the defined map is a bijective function as $T(\boldsymbol{z}) = (r, \boldsymbol{w})$ if and only if $\boldsymbol{z} = r\boldsymbol{w}/\|\boldsymbol{w}\| = T^{-1}(r, \boldsymbol{w})$. Accordingly, let us consider the following measure that is concentrating on $\mathbb{S}$

$$S(B) = \mu_*\left(\{\boldsymbol{z} \in [0,\infty)^d \mid \|\boldsymbol{z}\| \geq 1, \boldsymbol{z}/\|\boldsymbol{z}\| \in B\}\right) \tag{A.21}$$

where B represents a Borel subsets of $\mathbb{S}$. The measure S is known as the spectral measure as it is determined by the selected norm and the exponent measure. As a result of $(A.18)$ the spectral measure can be extended as follows

$$\begin{aligned}
&\mu_*\left(\{\boldsymbol{z} \in [0,\infty)^d \mid \|\boldsymbol{z}\| \geq r, \boldsymbol{z}/\|\boldsymbol{z}\| \in B\}\right) \\
&\quad = r^{-1}\mu_*\left(\{r^{-1}\boldsymbol{z} \in [0,\infty)^d \mid \|\boldsymbol{z}\| \geq r, \boldsymbol{z}/\|\boldsymbol{z}\| \in B\}\right) \\
&\quad = r^{-1}\mu_*\left(\{r^{-1}\boldsymbol{z} \in [0,\infty)^d \mid \|r^{-1}\boldsymbol{z}\| \geq 1, (r^{-1}\boldsymbol{z})/\|r^{-1}\boldsymbol{z}\| \in B\}\right) \\
&\quad = r^{-1}\mu_*\left(\{\boldsymbol{y} \in [0,\infty)^d \mid \|\boldsymbol{y}\| \geq 1, \boldsymbol{y}/\|\boldsymbol{y}\| \in B\}\right) \\
&\quad = r^{-1}S(B),
\end{aligned}$$

for $0 < r < \infty$. Hence, following the new coordinate system it can be deduced that μ_* is a product measure as it factors into a function of the radial coordinate which is equal to $r^{-2}dr$ and an angular component equal to the spectral measure. Hence, if T is defined as $T(\boldsymbol{z}) = (\|\boldsymbol{z}\|, \boldsymbol{z}/\|\boldsymbol{z}\|)$ then

$$\mu_* \circ T^{-1}(dr, dw) = r^{-2}drS(dw) \tag{A.22}$$

which is referred to as the spectral decomposition of the exponent measure. Subsequently, for $\boldsymbol{z} \in E$ it follows that as

$$\mu_*(E \setminus P_{\boldsymbol{z}}) = \mu_* \circ T^{-1} \circ T(E \setminus P_{\boldsymbol{z}})$$

and

$$\begin{aligned} T(E \setminus P_{\boldsymbol{z}}) &= T(\{\boldsymbol{y} \in E \mid y_j > z_j \text{ for some j} = 1, \ldots, \text{d}\}) \\ &= \{(r, \boldsymbol{w}) \in (0, \infty) \times \mathbb{S} \mid rw_j > z_j \text{ for some j} = 1, \ldots, \text{d}\} \\ &= \left\{(r, \boldsymbol{w}) \mid r > \frac{z_j}{w_j} \text{ for some j} = 1, \ldots, \text{d}\right\} \\ &= \left\{(r, \boldsymbol{w}) \mid r > \min_{1 \le j \le d} \left(\frac{z_j}{w_j}\right)\right\} \end{aligned}$$

then it can be deduced that the measure of this set is

$$\begin{aligned} \mu_*(E \setminus P_{\boldsymbol{z}}) &= \int_{T(E \setminus P_{\boldsymbol{z}})} \mu_*(d\boldsymbol{z}) \\ &= \int_{T(E \setminus P_{\boldsymbol{z}})} \mu_* \circ T^{-1}(dr, dw) \\ &= \int_{\mathbb{S}} \left(\int_{\min_{1 \le j \le d} \left(\frac{z_j}{w_j}\right)}^{\infty} r^{-2}\, dr \right) S(d\boldsymbol{w}) \\ &= \int_{\mathbb{S}} \left(\min_{1 \le j \le d} \left(\frac{z_j}{w_j}\right) \right)^{-1} S(d\boldsymbol{w}) \\ &= \int_{\mathbb{S}} \max_{1 \le j \le d} \left(\frac{w_j}{z_j}\right)\, S(d\boldsymbol{w}). \end{aligned} \tag{A.23}$$

Hence, the final form stated in the theorem is found by comparing (A.23) and $V_*(\boldsymbol{z})$. It follows that the condition that the margins of G_* are standard Fréchet distributed is forced by the requirement that

$$\int_{\mathbb{S}} w_j S(d\boldsymbol{w}) = 1, \qquad \text{for all } j = 1, \ldots, d. \qquad \square$$

A.2 Supporting Theoretical Results

Lemma A.1. *Let $\{\xi_n\}_{n \in \mathbb{N}}$ be an IID sequence with common distribution D. Let $\tau \in [0, \infty]$ and suppose that $\{u_n\}_{n \in \mathbb{N}}$ is a sequence of real numbers. Then the following are equivalent*

$$n(1 - D(u_n)) \to \tau \qquad \text{as } n \to \infty, \tag{A.24}$$

$$\mathbb{P}\{M_n \le u_n\} \to e^{-\tau} \qquad \text{as } n \to \infty. \tag{A.25}$$

Proof. Let us first consider the case when $0 \leq \tau < \infty$. If $(A.24)$ holds, then from (3.2), it follows that

$$\begin{aligned}\mathbb{P}\{M_n \leq u_n\} &= D^n(u_n) \\ &= (1-(1-D(u_n)))^n \\ &= \left(1 - \frac{\tau}{n} + o\left(\frac{1}{n}\right)\right)^n. \end{aligned} \tag{A.26}$$

Since, $e^z = \lim_{n\to\infty}(1+z/n)^n$, then $(A.25)$ follows immediately. Conversely, if $(A.25)$ holds, then $1 - D(u_n) \to 0$. If not, for some subsequence $\{n_k\}$, $1 - D(u_{n_k})$ would be bounded away from 0. Then $(A.26)$ would imply that $\mathbb{P}\{M_{n_k} \leq u_{n_k}\} \to 0$. Applying logarithms on both sides of $(A.25)$, it can easily be deduced that

$$-n\ln(1-(1-D(u_n))) \to \tau.$$

If z is close to zero, it can be inferred that $-\ln(1-z) \approx z$, and hence it follows that

$$n(1 - D(u_n) = \tau + o(1),$$

giving result $(A.24)$. For the second case, when $\tau = \infty$, if $(A.24)$ holds but $(A.25)$ does not ($\mathbb{P}\{M_n \leq u_n\} \not\to 0$). Hence, then there must be a subsequence $\{n_k\}$, such that $\mathbb{P}\{M_{n_k} \leq u_{n_k}\} \to e^{-\tau'}$ as $k \to \infty$ for some $\tau' < \infty$. However, since $\tau' < \infty$ and by the above arguments, it follows that $(A.24)$ implies $(A.25)$. Hence, $n_k(1 - D(u_{n_k})) \to \tau' < \infty$, which contradicts the assumption that $(A.24)$ holds with $\tau = \infty$. Subsequently, in the same manner as above, $(A.25)$ implies $(A.24)$ when $\tau = \infty$. □

Definition A.1 (Equality in type)**.** *Two distribution functions G_1 and G_2 are of the same type, if and only if there exist constants $a > 0$ and $b \in \mathbb{R}$ such that*

$$G_2(x) = G_1(ax + b).$$

In other words, two distributions are of the same type if they are obtained from one another by location and scale transformation.

Definition A.2 (Left-continuous inverse)**.** *Let f be a non-decreasing and right continuous function. Then the left-continuous (generalized) inverse of f is the function*

$$f^{\leftarrow}(x) = \inf\{y : f(y) \geq x\}.$$

with the convention $\inf\{\mathbb{R}\} = -\infty$ *and* $\inf\{\emptyset\} = +\infty$. *The properties of the left-continuous inverse of a function are reviewed in* [70].

Definition A.3. *A map* $S : \mathbb{R}^n \to \mathbb{R}^n$ *is an affine transformation on* $\mathbb{R}^n$ *if*

$$S(z) = Az + b$$

for all $z \in \mathbb{R}^n$, *where* A *is a linear transformation and* b *is the translational part of the transformation.*

Theorem A.1. *Let* $\{x_1, \ldots, x_n\}$ *be independent realizations from a distribution within a parametric family* $\mathcal{F}$, *and let* $\ell(\cdot)$ *and* $\widehat{\theta}_0$ *denote repetitively the log-likelihood function and the maximum likelihood estimator of the* d*-dimensional model parameter* θ_0. *Then, under suitable regularity conditions, for large* n

$$\widehat{\theta}_0 \sim \mathcal{N}_d(\theta_0, n^{-1} I_E(\theta_0)^{-1}),$$

where

$$I_E(\theta_0) = \begin{bmatrix} e_{1,1}(\theta) & \ldots & e_{1,d}(\theta) \\ \vdots & \ddots & \vdots \\ e_{d,1}(\theta) & \ldots & e_{d,d}(\theta) \end{bmatrix}$$

with

$$e_{i,j}(\theta) = \mathbb{E}\left\{ -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ell(\theta) \right\}.$$

Proof. The proof of this Theorem can be found in [26, Theorem 2.2] □

Theorem A.2. *Let* $x_1, \ldots, x_2$ *be independent realizations from a distribution within a parametric family* $\mathcal{F}$, *and let* $\widehat{\theta}_0$ *denote the maximum likelihood estimator of the* d*-dimensional model parameter* $\theta_0 = (\theta^{(1)}, \theta^{(2)})$, *where* $\theta^{(1)}$ *is a* k*-dimensional subset of* θ_0. *Then, under suitable regularity conditions, for large* n

$$D_p(\theta^{(1)}) = 2(l(\widehat{\theta}_0) - l_p(\theta^{(1)})) \sim \chi_{k^2}.$$

Proof. The proof of this Theorem can be found in [26, Theorem 2.6] □

Lemma A.2. *Suppose f_n is a sequence of non-decreasing functions and g is a non-decreasing function. Suppose, that for each x in some open interval (a, b) that is a continuity point of g,*

$$\lim_{n\to\infty} f_n(x) = g(x).$$

Let $f_n^{\leftarrow}$, $g^{\leftarrow}$ be the left-continuous inverses of f_n and g. Then, for each x in the interval $(g(a), g(b))$ that is a continuity point of $g^{\leftarrow}$ we have

$$\lim_{n\to\infty} f_n^{\leftarrow}(x) = g^{\leftarrow}(x).$$

Proof. The proof of this Lemma is rigorously explained in [70, 126]. In addition, additional results on the left-continuous generalized inverse can be found in [39] and [126, Section 0.2]. □

Theorem A.3 (Representation theorem for regularly varying functions)**.** *If $h \in \mathcal{R}_\alpha$ for some $\alpha \in \mathbb{R}$, then*

$$h(z) = c(z) \exp\left\{\int_t^z \frac{\delta(u)}{u}\, du\right\}, \qquad z \geq t, \tag{A.27}$$

for some $t > 0$ where c and δ are measurable functions such that $c(z) \to c_0$ where $c_0 \in (0, \infty)$ and $\delta(z) \to \alpha$ as $z \to \infty$. The converse implication also holds.

Proof. The proof of this Theorem can be found in [39, Theorem A3.3] □

Theorem A.4. *The underlying distribution function D belongs to the maximum domain of attraction of the family:*

Type I: $D \in \mathcal{D}(G_\xi)$ *for* $\xi = 0 \iff$ *there exists some positive function f such that*

$$\lim_{t\uparrow z^*} \frac{1 - D(t + zf(t))}{1 - D(t)} = e^{-z} \tag{A.28}$$

for all $z \in \mathbb{R}$. If (A.28) holds for some f, then it follows that $\int_t^{z^*}(1 - D(s))\, ds < \infty$ for $t < z^*$ and (A.28) holds with

$$f(t) = \frac{\int_t^{z^*}(1 - D(s))\, ds}{1 - D(t)};$$

Type II: $D \in \mathcal{D}(G_\xi)$ *for* $\xi > 0 \iff z^* = \infty$ *and*

$$\lim_{t\to\infty} \frac{1 - D(tz)}{1 - D(t)} = z^{-\frac{1}{\xi}}, \qquad \text{for all } z > 0.$$

Type III: $D \in \mathcal{D}(G_\xi)$ *for* $\xi < 0 \iff z^* < \infty$ *and*

$$\lim_{t\downarrow 0} \frac{1 - D(z^* - tz)}{1 - D(z^* - t)} = z^{-\frac{1}{\xi}}, \qquad \text{for all } z > 0.$$

Proof. The full proof of this theorem is omitted as it is lengthy and in order to prove result it requires a lot of supplementary results. However, it can be found in [58], [70, Theorem 1.2.1] and [67, Theorem 1.6.1]. □

Definition A.4 (Slowly varying and regularly varying functions)**.**

(*i*) *A positive, Lebesgue-measurable function* L *on* $(0, \infty)$ *is slowly varying* ($L \in \mathcal{R}_0$) *at* ∞ *if*

$$\lim_{z\to\infty} \frac{L(tz)}{L(z)} = 1, \qquad t > 0.$$

(*ii*) *A positive, Lebesgue-measurable function* h *on* $(0, \infty)$ *is regularly varying* ($h \in \mathcal{R}_\rho$) *at* ∞ *with index* $\rho \in \mathbb{R}$ *if*

$$\lim_{z\to\infty} \frac{h(tz)}{h(z)} = t^\rho, \qquad t > 0.$$

The variation described in Definition A.4 can be transformed to define the variation at 0 *by replacing* $z \to 0$*. The family of slow varying functions contains functions that for large* z *exhibit a relative slow change when compared to the power functions. On the other hand, the family of regularly varying functions contains functions which can be decomposed into a power function and a slowly varying function* ($h(z) = z^\rho L(z)$)*. More in depth information on regular variation can be found in* [39, Appendix A3] *and* [70, Appendix B].

Definition A.5 (Von Mises function)**.** *A distribution function* G *with right end point* z^* *is called a Von Mises function if it has the following representation: There must exist* $x_0 < z^*$ *such that for* $x_0 < z < z^*$ *and* $c > 0$ *such that*

$$1 - G(z) = c \cdot exp\left\{ - \int_{x_0}^{z} (1/f(u))\, du \right\}, \tag{A.29}$$

where $f(u) > 0$. Also, f is absolutely continuous on (x_0, z^*) with density $f'(u)$ and $\lim_{u\uparrow z^*} f'(u) = 0$.

Lemma A.3. *Suppose that $X_n \xrightarrow{\mathbf{d}} X$ for a random vector X with a continuous distribution function. Then*

$$\sup_x(|\mathbb{P}\{X_n \leq x\} - \mathbb{P}\{X \leq x\}|) \to 0.$$

Proof. The proof of this Lemma can be found in [156, Lemma 2.11] □

Theorem A.5. *A sequence (f_n) of function on the set $S \subseteq \mathbb{R}$ converges uniformly to a function f on X if and only if*

$$\lim_{n\to\infty} \sup_{x\in X} \{|f(x) - f_n(x)|\} = 0.$$

Proof. Suppose that f_n converges uniformly on X and let $\epsilon > 0$. Then there exists N such that for all $n > N$ implies that $|f(x) - f_n(x)| < \epsilon/2$ for all $x \in X$. It follows that $0 \leq \sup_{x\in X}(|f(x) - f_n(x)|) \leq \epsilon/2 < \epsilon$ for all $n > N$. Hence, it can be concluded that $\lim_{n\to\infty}(\sup_{x\in X}(|f(x) - f_n(x)|)) = 0$.

Conversely, if $\lim_{n\to\infty}(\sup_{x\in X}(|f(x) - f_n(x)|)) = 0$ holds then there exists N such that for $\epsilon > 0$ and for all $n > N$ it follows that $|\sup_{x\in X}(f(x) - f_n(x)) - 0| = \sup_{x\in X}(f(x) - f_n(x)) < \epsilon/2$. Then by the definition of supremum it follows that $|f(x) - f_n(x)| \leq \epsilon/2 < \epsilon$ for all $x \in X$ and $n > N$. Hence, $f_n \to f$ uniformly on X. □

Definition A.6 (Order Statistics (OS))**.** *Let $\{X_1, \ldots, X_n\}$ denote a random sample of size n extracted from a distribution function D. Arranging the sample in descending order of magnitude generates a new family of observations represented as follows $X^{(1)} \geq X^{(2)} \geq \cdots \geq X^{(n)}$ and called the order statistics associated with the original sample. In particular, for $i \in \{1, \ldots, n\}$ the random variable $X^{(i)}$ denotes the i^{th} order statistic.*

Theorem A.6. *Let $\tau \in [0, \infty]$ and S_n denoting the number of exceedances of a threshold level (u_n) by an IID sequence $\{\xi_1, \ldots \xi_n\}$. If $\{u_n\}_{n\in\mathbb{N}}$ satisfies (A.24) then for $k = 0, 1, 2, \ldots,$*

$$\mathbb{P}\{S_n \leq k\} \to e^{-\tau} \sum_{s=0}^{k} \frac{\tau^s}{s\,!} \quad \textit{for } \tau \in (0, \infty). \tag{A.30}$$

When $\tau = \infty$ the right-hand side is taken as zero and when $\tau = 0$ the right-hand side is taken as one. Conversely, if (A.30) holds for any fixed k, then (A.24) holds and thus (A.30) holds for all k.

Proof. For a fixed $n \in \mathbb{N}$ the number of exceedances of a threshold level u_n by an IID sequence $\{\xi_1, \ldots \xi_n\}$ can be represented as follows

$$S_n = \sum_{i=1}^{n} \mathbb{I}_{\{\xi_i > u_n\}}.$$

Then it can be deduced that S_n is a sum of IID Bernoulli random variables with success probability:

$$\mathbb{E}\left(\mathbb{I}_{\{\xi_i > u_n\}}\right) = \mathbb{P}\{\xi_i > u_n\} = 1 - D(u_n) = \overline{D}(u_n) = p_n.$$

Hence B_n is a binomial random variable with parameters n and p_n. If it is assumed that $np_n \to \tau$ then (A.30) follows directly. For $\tau \in (0, \infty)$ the limit follows by the application of the Poisson limit theorem:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\{S_n \leq k\} &= \lim_{n\to\infty} \sum_{r=0}^{k} \mathbb{P}\{S_n = r\} \\
&= \lim_{n\to\infty} \sum_{r=0}^{k} \binom{n}{r} (p)^r (1-p)^{n-r} \\
&= \sum_{r=0}^{k} \lim_{n\to\infty} \binom{n}{r} (p)^r (1-p)^{n-r} \\
&= e^{-\tau} \sum_{r=0}^{k} \frac{\tau^r}{r\,!}.
\end{aligned}$$

On the other hand, when $\tau = 0$ it can be inferred that $p_n = o\,(1/n)$. Hence, due to the monotone non-decreasing property

$$\mathbb{P}\{S_n \leq k\} \geq \mathbb{P}\{S_n = 0\} = (1 - p_n)^n = \left(1 - o\left(\frac{1}{n}\right)\right)^n \to 1.$$

Which implies that $\mathbb{P}\{S_n \leq k\} \to 1$. Finally, when $\tau = \infty$, an arbitrary $\theta > 0$ is selected such that $np_n \geq \theta$, for some large n. Consequently, it can be deduced that

$$\mathbb{P}\{S_n \leq k\} \leq \sum_{r=0}^{k} \binom{n}{r} \left(\frac{\theta}{n}\right)^s \left(\frac{1-\theta}{n}\right)^{n-s}.$$

Since, the binomial distribution on the right-hand side is decreasing in θ, then for a fixed k it follows that

$$\limsup_{n\to\infty} \mathbb{P}\{S_n \leq k\} \leq e^{-\theta} \sum_{s=0}^{k} \frac{\theta^s}{s\,!} \to 0 \quad \text{as } \theta \to \infty,$$

which implies that $\lim_{n\to\infty} \mathbb{P}\{S_n \leq k\} = 0$. Conversely, suppose that $(A.30)$ holds for some k but $np_n \not\to \tau$. Then there exists some $\tau' \neq \tau$ in $[0,\infty]$ and a subsequence $\{n_l\}$ such that $n_l p_{n_l} \to \tau'$. Applying the same arguments as above implies that

$$\mathbb{P}\{S_{n_l} \leq k\} \to e^{-\tau'} \sum_{s=0}^{k} \frac{(\tau')^s}{s\,!} \quad \text{as } l \to \infty.$$

However, this contradicts $(A.30)$ as the function $e^{-x}\sum_{s=0}^{k} \frac{x^s}{s\,!}$ is strictly decreasing in $x \in [0,\infty]$. □

Theorem A.7. *Suppose that $\{X_1,\ldots,X_n\}$ be an IID sequence with common underlying distribution function D. Analogous to Lemma A.1, let $\{u_n^{(r)}\}_{r\in\{1,\ldots,k\}}$ be a sequence of levels such that $u_n^{(1)} \geq u_n^{(2)} \geq \cdots \geq u_n^{(k)}$ and*

$$\begin{aligned} n\big(1 - D\big(u_n^{(1)}\big)\big) &\to \tau_1 \\ &\vdots \\ n\big(1 - D\big(u_n^{(k)}\big)\big) &\to \tau_k, \end{aligned} \tag{A.31}$$

where $0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k \leq \infty$ and define $S_n^{(k)}$ to be the number of exceedances of $u_n^{(k)}$ by $\{X_1,\ldots,X_n\}$. Then for $s_1 \geq 0,\ldots,s_k \geq 0$,

$$\begin{aligned} \mathbb{P}\big\{S_n^{(1)} = s_1,\ S_n^{(2)} = s_1 + s_2, \ldots,\ S_{(n)}^{(k)} = s_1 + \cdots + s_k\big\} \\ \to \frac{\tau_1^{s_1}}{s_1\,!} \cdot \frac{(\tau_2 - \tau_1)^{s_2}}{s_2\,!} \cdots \frac{(\tau_k - \tau_{k-1})^{s_k}}{s_k\,!} e^{-\tau_k} \end{aligned} \tag{A.32}$$

as $n \to \infty$. When $\tau_k = \infty$ the right-hand side converges to 0.

Proof. In Theorem $A.6$, it was found that S_n is a binomial random variable and hence it follows that $\boldsymbol{S} = (S_n^{(1)}, \ldots, S_n^{(n)})$ it a multinomial random variable. Consequently, if $p_{n,k} = 1 - D(u_n^{(k)})$ it can be deduced that the left-hand side of $(A.32)$ is

$$\begin{aligned}
&\frac{n\,!}{s_1\,!\,s_2\,!\cdots s_k\,!}\mathbb{P}\{X_1,\ldots,X_{s_1} > u_n^{(1)} \geq X_{s_1+1},\ldots,X_{s_1+s_2} > u_n^{(2)},\ldots \\
&\qquad \ldots, u_n^{(k-1)} \geq X_{s_1+\cdots+s_{k-1}+1},\ldots,X_{s_1+\ldots+s_k} > u_n^{(k)} \geq X_{s_1+\cdots+s_k+1},\ldots,X_n\} \\
&= \frac{n\,!}{s_1\,!\,s_2\,!\cdots s_k\,!}p_{n,\,1}^{s_1}\cdot(p_{n,\,2}-p_{n,\,1})^{s_2}\cdots(p_{n,\,k}-p_{n,\,k-1})^{s_k}\cdot(1-p_{n,\,k})^{n-s_1-\cdots-s_k} \\
&= \binom{n}{s_1}p_{n,\,1}^{s_1}\binom{n-s_1}{s_2}(p_{n,\,2}-p_{n,\,1})^{s_2}\cdots \\
&\quad\cdot\binom{n-s_1-\cdots-s_{k-1}}{s_k}(p_{n,\,k}-p_{n,\,k-1})^{s_k}\cdot(1-p_{n,\,k})^{n-s_1-\cdots-s_k}. \qquad (A.33)
\end{aligned}$$

If $(A.31)$ is considered and $\tau_k < \infty$ then it follows that

$$\begin{aligned}
\binom{n}{s_1}p_{n,\,1}^{s_1} &= \frac{n\cdot\cdots\cdot(n-s_1+1)p_{n,\,1}^{s_1}}{s_1\,!} \to \frac{\tau_1^{s_1}}{s_1\,!}, \\
(1-p_{n,\,k})^{n-s_1-\cdots-s_k} &= \left(1-\frac{np_{n,\,k}}{n}\right)^{n-s_1-\cdots-s_k} \to e^{-\tau_k},
\end{aligned}$$

and for $2 \leq l \leq k$ it can be concluded that

$$\begin{aligned}
&\binom{n-s_1-\cdots-s_{l-1}}{s_l}(p_{n,\,l}-p_{n,\,l-1})^{s_l} \\
&\quad= \frac{(n-s_1-\cdots-s_{l-1})\cdot\cdots\cdot(n-s_1-\cdots-s_l+1)(p_{n,\,l}-p_{n,\,l-1})^{s_l}}{s_l\,!} \\
&\quad\to \frac{(\tau_l-\tau_{l-1})^{s_l}}{s_l\,!}.
\end{aligned}$$

Consequently, from $(A.31)$ and $(A.33)$ and the above, the right-hand side of $(A.32)$ is obtained. On another note, if $\tau_k = \infty$ then the left-hand side of $(A.32)$ is less than $\mathbb{P}\{S_n^{(k)} = s_1 + \cdots + s_k\}$. In accordance with Theorem $A.6$, $\mathbb{P}\{S_n^{(k)} = s_1 + \cdots + s_k\} \to 0$ and hence $(A.32)$ still holds. □

Definition A.7 (Maximum Domain of Attraction (MDA))**.** *The underlying distribution D is said to belong to the Maximum Domain of Attraction of the distribution G ($D \in \mathcal{D}(G)$), if there exist sequences of constants $a_n > 0$ and b_n such that the limit* (3.5) *holds.*

Lemma A.4 (Threshold Stability Property)**.** *Suppose that the conditions in Theorem* 1 *holds then considering Theorem* 5 *it follows that*

$$D_u(z) \approx H_{\xi,\sigma}(z),$$

for some high threshold u*. Then for any higher threshold* $v \geq u$ *it follows that*

$$D_v(z) \approx H_{\xi,\,\sigma+\xi(v-u)}(z). \tag{A.34}$$

Proof. Consider the form of the excess distribution (3.22), then it can be deduced that

$$\begin{aligned}
\bar{D}_v(z) = \frac{1-D(z+v)}{1-D(v)} &= \frac{\bar{D}(u+(z+v-u))}{\bar{D}(u)}\frac{\bar{D}(u)}{\bar{D}(u+(v-u))} \\
&= \frac{\bar{D}_u(z+v-u)}{\bar{D}_u(v-u)} \\
&\approx \frac{\bar{H}_{\xi,\sigma}(z+v-u)}{\bar{H}_{\xi,\sigma}(v-u)} \\
&= \begin{cases} \left[\frac{1+\xi\frac{z+v-u}{\sigma}}{1+\xi\frac{v-u}{\sigma}}\right]^{-1/\xi} & if\ \xi \neq 0, \\ \frac{e^{-\frac{z+v-u}{\sigma}}}{e^{-\frac{v-u}{\sigma}}} & if\ \xi = 0 \end{cases} \\
&= \begin{cases} \left[1+\xi\left(\frac{z}{\sigma+\xi(v-u)}\right)\right]^{-1/\xi} & if\ \xi \neq 0, \\ e^{-\frac{z}{\sigma}} & if\ \xi = 0 \end{cases} \\
&= \bar{H}_{\xi,\,\beta}(z),
\end{aligned}$$

where $\beta = \sigma$ if $\xi = 0$ and $\beta = \sigma + \xi(v-u)$ if $\xi \neq 0$. □

Theorem A.8 (Extended regular variation)**.** *Let* U *represent the left continuous inverse of* $1/(1-D)$ *such that for* $\xi \in \mathbb{R}$ *it follows that*

$$D \in \mathcal{D}(G_\xi) \qquad \textit{if and only if} \qquad \lim_{t\to\infty}\frac{U(tz)-U(t)}{a(t)} = \frac{z^\xi - 1}{\xi},$$

for some positive function $a(\cdot)$ *with* $z > 0$*. When* $\xi = 0$ *the right hand side of the limit is interpreted as* $\ln(z)$*.*

Proof. The proof of this theorem can be found in [70, Theorem 1.16] □

Proposition A.1. *Consider the following:*

(*i*) *Suppose G is a multivariate distribution function with continuous marginals. Define for $j \in \{1, \ldots, d\}$*

$$\psi_j(z) = (1/(-\ln(G_j)))^{\leftarrow}(z), \qquad z > 0$$

and

$$G_*(\boldsymbol{z}) = G(\psi_1(z_1), \ldots, \psi_d(z_d)) \qquad \boldsymbol{z} \geq \mathbf{0}.$$

Then G_* has marginal distribution $G_{*i}(z) = G_{1,1,1}(z)$ and G is a multivariate extreme value distribution if and only if G_* is also.

(*ii*) *Define $U_j = -1/\ln(D_j)$, for $1 \leq j \leq d$ and let D_* be the distribution of $(U_1(Z_1), \ldots, U_d(Z_d))$ so that*

$$D_*(\boldsymbol{z}) = D(U_1^{\leftarrow}(z_1), \ldots, U_d^{\leftarrow}(z_d)).$$

If (4.2) holds, so that $D \in \mathcal{D}(G)$, then $D_* \in \mathcal{D}(G_*)$ and

$$\mathbb{P}\left(\max_{1\leq i\leq n} \frac{U_j(Z_{i,j})}{n} \leq z_j, 1 \leq j \leq d\right) = D_*^n(nz) \to G_*. \tag{A.35}$$

Conversely if (A.35) holds as well as (4.5) and G_* has non-degenerate marginals then $D \in \mathcal{D}(G)$ and (4.2) is true.

Proof. The proof of this proposition can be found in [126, Prop 5.10] □

Definition A.8. *A distribution function D on $\mathbb{R}^d$ is max-infinitely divisible (max-id) if for every n there exists a distribution D_n on $\mathbb{R}^d$ such that*

$$D = (D_n)^n.$$

Hence, it can be deduced that $D^{\frac{1}{n}}$ is a distribution function. In general a random vector with max-id distribution is called max-id.

Proposition A.2. *The following are equivalent:*

(*i*) *D is max-id.*

(*ii*) *For some* $\boldsymbol{l} = (l_1, \ldots, ld) \in [-\infty, \infty)^d$, *there exists an exponent measure* μ *on*

$$E = \left\{ \boldsymbol{x} \in \mathbb{R}^d \mid l_i < x_i \leq \infty,\ 1 \leq i \leq d \right\},$$

such that

$$D(\boldsymbol{z} = (z_1, \ldots z_d)) = \begin{cases} e^{-\mu(E \setminus \left\{ \boldsymbol{x} \in \mathbb{R}^d \mid -\infty \leq x_i \leq z_i,\, 1 \leq i \leq d \right\})} & \boldsymbol{z} \geq \boldsymbol{l}, \\ 0 & \textit{otherwise.} \end{cases} \tag{A.36}$$

Proof. The proof of this proposition can be found in [126, Prop 5.8] □

Definition A.9. *A distribution function* $G(\boldsymbol{z})$ *is max-stable if for* $j \in \{1, \ldots, d\}$ *and every* $k > 0$ *there exist functions* $\alpha_j(k) > 0$ *and* $\beta_j(k)$ *such that*

$$G(z_1, \ldots, z_d) = G^k(\alpha_1(k)z_1 + \beta_1(k), \ldots, \alpha_d(k)z_d + \beta_d(k)). \tag{A.37}$$

Proposition A.3. *The class of multivariate extreme value distributions is precisely the class of max-stable distributions functions with non-degenerate marginals.*

Proof. The proof of this proposition can be found in [126, Prop 5.9] □

Definition A.10. *An Archimedean generator is a decreasing and continuous function* $\psi : [0, \infty) \mapsto [0, 1]$ *that satisfies the following conditions:* $\psi(0) = 1$, $\lim_{t \to \infty} \psi(t) = 0$ *and it is strictly decreasing on* $[0, \inf\{t \mid \psi(t) = 0\})$, *where* $\inf\{\emptyset\} = \infty$.

Definition A.11. *A d-dimensional copula* C *is called Archimedean if it admits the representation*

$$C(\boldsymbol{u}) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2) + \cdots + \psi^{-1}(u_d)), \tag{A.38}$$

for all $\boldsymbol{u} \in [0, 1]^d$ *and for some Archimedean generator* ψ *(Definition A.10).*

Definition A.12 (Upper Set)**.** *A Set* $\mathcal{S} \subseteq \mathbb{R}^n$ *is called an Upper Set if, and only if,* $\boldsymbol{x} \in \mathcal{S}$ *and* $\boldsymbol{y} \geq \boldsymbol{x}$ *(in a component-wise manner) implying that* $\boldsymbol{y} \in \mathcal{S}$.

Definition A.13 (Penman-Monteith equation)**.** *As an alternative to direct evapotranspiration measurement, the rate of evapotranspiration can be estimated from meteorological data with the Penman-Monteith equation:*

$$ET = \frac{1}{\lambda} \frac{\Delta(R_n - G) + \rho c_p (e_s - e_a)/r_a}{\Delta + \gamma(1 + r_s/r_a)},$$

where

R_n *is the net irradiance;*

ρ *is the mean air density;*

G *is the ground heat flux;*

γ *is the psychometric constant;*

r_s *is the bulk surface resistance;*

r_a *is the aerodynamic resistance;*

λ *is the latent heat of vaporization;*

c_p *is the specific heat of air at constant pressure;*

e_s *is the vapour pressure in the saturated region of a water surface;*

e_a *is the vapour pressure in the atmospheric space above the saturated region;*

Δ *is the rate of change of saturation specific humidity with air temperature.*

Moreover, the difference $(R_n - G)$ *represents the net balance of energy available at the surface, and the difference* $(e_s - e_a)$ *represents the vapor pressure deficit. The above equation can be directly used to compute the evapotranspiration for different crops, as the aerodynamic and the surface resistances are crop-specific.*

Definition A.14 (Richards equation)**.** *The water movement in soils is simulated by the following Richards equation:*

$$\frac{\partial \theta}{\partial t} = \nabla \cdot \boldsymbol{K}(h)(\nabla h + \nabla z) - q_e,$$

where

t *is time;*

z *is depth;*

h *is the soil water pressure head;*

θ is the volumetric water content;

$\boldsymbol{K}(h)$ is the hydraulic conductivity;

q_e is the sink term due to the removal of water by plants and/or losses of water by evaporation from soil surface layers.

The soil water movement can be reduced to one-dimensional infiltration movement as follows:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left(\boldsymbol{K}(\theta)\left(\frac{\partial h}{\partial z} + 1\right)\right) - q_e.$$

Lemma A.5. *Let $F_{X,Y}(x, y)$, $F_X(x)$ and $F_Y(y)$ denote the joint and marginal distributions of andom vector (X, Y) where $\mathbb{E}(XY)$, $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ are assumed to be finite. Then we have*

$$Cov(X_1., X_2) = \int_{-\infty}^{\infty} \{F_{X,Y}(x, y) - F_X(x)F_Y(y)\}\, dxdy.$$

Proof. The proof of this lemma can be found in [38, Lemma 1]. □

Definition A.15 (Concordant & Discordant)**.** *Let the arbitrary points (x_i, y_i) and (x_j, y_j) represent two observations from the vector (X, Y) in the usual probability space.*

- *(Concordant): The two observations are considered to be concordant if $x_i < x_j$ and $y_i < y_j$ or if $x_j < x_i$ and $y_j < y_i$. In other words, the two observations are concordant if $(x_i - x_j)(y_i - y_j) > 0$.*
- *(Discordant): The two observations are considered to be discordant if $x_i < x_j$ and $y_j < y_i$ or if $x_j < x_i$ and $y_i < y_j$. In other words, the two observations are concordant if $(x_i - x_j)(y_i - y_j) < 0$.*

Theorem A.9. *Let X_1 and X_2 be a continuous random variables whose copula is C. Then the population version of Kendall's tau for X_1 and X_2 is given by*

$$\begin{aligned}\tau(X_1, X_2) = \tau(C) &= 4 \int\int_{[0,1]^2} C(u_1, u_2)\, dC(u_1, u_2) - 1 \\ &= 4 \cdot \mathbb{E}(C(U_1, U_2)) - 1, \end{aligned} \tag{A.39}$$

where U_1 and U_2 are uniform $[0, 1]$ random variables with copula C and $\mathbb{E}$ denoting the expectation with respect to the probability measure induced by the copula.

Proof. Suppose that (X_{11}, X_{12}) and (X_{21}, X_{22}) are independent copies of (X_1, X_2). Since, it is assumed that the random variables are continuous then it can be deduced that

$$\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) = 0) = 0.$$

This means that the regions of discordance and concordance split the probability space into two non-intersecting regions. Hence, we have that

$$\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) > 0) + \mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) < 0) = 1. \tag{A.40}$$

Substituting equation (A.40) in equation (4.17), it follows easily that

$$\tau(X_1, X_2) = 2 \cdot \mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) > 0) - 1. \tag{A.41}$$

In addition, it can be inferred that the probability of the event $\{(X_{11}-X_{21})(X_{12}-X_{22}) > 0\}$ is equal to the summation of the probabilities of the events $\{X_{11} > X_{21}, X_{12} > X_{22}\}$ and $\{X_{11} < X_{21}, X_{12} < X_{22}\}$. Therefore, it follows that

$$\begin{aligned}&\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) > 0)\\ &\qquad = \mathbb{P}(X_{11} > X_{21}, X_{12} > X_{22}) + \mathbb{P}(X_{11} < X_{21}, X_{12} < X_{22}).\end{aligned}$$

If the first part is considered with the probability transformations $u_1 = F_1(x_1)$ and $u_2 = F_2(x_2)$ it will result in the following

$$\begin{aligned}&\mathbb{P}(X_{21} < X_{11}, X_{22} < X_{12})\\ &\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}(X_{21} < x_1, X_{22} < x_2)\, dH(x_1, x_2)\\ &\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(F_1(x_1), F_2(x_2))\, dC(F_1(x_1), F_2(x_2))\\ &\qquad = \int_{0}^{1}\int_{0}^{1} C(u_1, u_2)\, dC(u_1, u_2).\end{aligned}$$

In similar fashion, by the interchangeability, it can be deduced that

$$\begin{aligned}&\mathbb{P}(X_{11} < X_{21}, X_{12} < X_{22})\\ &\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}(x_1 < X_{21}, x_2 < X_{22})\, dC(F_1(x_1), F_2(x_2))\end{aligned}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1 - \mathbb{P}(x_1 \geq X_{21}, x_2 \geq X_{22})\, dC(F_1(x_1), F_2(x_2))$$
$$= \int_0^1 \int_0^1 1 - u_1 - v_1 + C(u_1, u_2)\, dC(u_1, u_2).$$

Following the fact that C is the distribution function of the random variables $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$, and that $\mathbb{E}(U_i) = 1/2$, then it can be concluded that

$$\mathbb{P}(X_{11} < X_{21}, X_{12} < X_{22}) = 1 - \frac{1}{2} - \frac{1}{2} + \int_0^1 \int_0^1 C(u_1, u_2)\, dC(u_1, u_2)$$
$$= \int_0^1 \int_0^1 C(u_1, u_2)\, dC(u_1, u_2).$$

Finally, substituting everything in equation (A.41), equation (A.39) follows and concludes the first part. In addition, the integral in (A.39) can be reformulated to be interpreted as the expected value of the function $C(U_1, U_2)$ with respect to the measure induced by the copula, where $U_i \sim \mathbb{U}[0, 1]$ with joint distribution function C. Therefore,

$$\tau(X_1, X_2) = \tau(C) = 4 \cdot \mathbb{E}(C(U_1, U_2)) - 1.$$

This result concludes the proof. □

Theorem A.10. *Let X_1 and X_2 be continuous random variables whose copula is C. Then the population version of Spearman's rho for X_1 and X_2 is given by*

$$\rho(X_1, X_2) = \rho(C) = 12 \int\int_{[0,1]^2} u_1 u_2\, dC(u_1, u_2) - 3 \tag{A.42}$$
$$= 12 \int\int_{[0,1]^2} C(u_1, u_2)\, du_1\, du_2 - 3 \tag{A.43}$$
$$= 12 \cdot \mathbb{E}(U_1, U_2) - 3 \tag{A.44}$$

where U_1 and U_2 are uniform $[0, 1]$ random variables with copula C.

Proof. The first equivalence of this Theorem follows the same path as the proof done in Theorem A.9. Hence, assume that (X_{11}, X_{12}), (X_{21}, X_{22}) and (X_{31}, X_{32}) are three independent copies of (X_1, X_2). Then

$$\rho(X_1, X_2) = 3\big[2\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{32}) > 0) - 1\big]$$

$$= 6\,\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{32}) > 0) - 3, \tag{A.45}$$

where

$$\begin{aligned}&\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{32}) > 0)\\ &\quad = \mathbb{P}(X_{11} > X_{21}, X_{12} > X_{32}) + \mathbb{P}(X_{11} < X_{21}, X_{12} < X_{32}).\end{aligned} \tag{A.46}$$

By integrating (A.46) with respect to the distribution of (X_{11}, X_{12}) and considering the fact that X_{21} and X_{32} are independent, it follows that

$$\begin{aligned}&\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{32}) > 0)\\ &\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}(X_{21} < x_1, X_{32} < x_2)\\ &\qquad\qquad + \mathbb{P}(x_1 < X_{21}, x_2 < X_{32})\, dH(x_1, x_2)\\ &\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}(X_{21} < x_1)\mathbb{P}(X_{32} < x_2)\\ &\qquad\qquad + \mathbb{P}(x_1 < X_{21})\mathbb{P}(x_2 < X_{32})\, dC(F_1(x_1), F_2(x_2))\\ &\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_1(x_1)F_2(x_2)\\ &\qquad\qquad + (1 - F_1(x_1))(1 - F_2(x_2))\, dC(F_1(x_1), F_2(x_2))\\ &\qquad = \int_0^1\int_0^1 u_1u_2 + (1-u_1)(1-u_2)\, dC(u_1, u_2)\\ &\qquad = 1 - \frac{1}{2} - \frac{1}{2} + 2\int_0^1\int_0^1 u_1u_2\, dC(u_1, u_2)\\ &\qquad = 2\int_0^1\int_0^1 u_1u_2\, dC(u_1, u_2).\end{aligned}$$

The result follows by substituting in $(A.45)$,

$$\rho(X_1, X_2) = \rho(C) = 12\int\int_{[0,1]^2} u_1u_2\, dC(u_1, u_2) - 3.$$

Since, u_1 and u_2 are realizations of the probability transforms $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$ which are uniform $[0, 1]$ random variables with joint distribution function C, then it can be deduced that the above integral can be interpreted as follows

$$\rho(C) = 12\int\int_{[0,1]^2} u_1u_2\, dC(u_1, u_2) - 3 = 12 \cdot \mathbb{E}(U_1U_2) - 3.$$

On the other hand, if instead $(A.46)$ is integrated with respect to the distribution of (X_{21}, X_{32}) then an alternative form is obtained. Hence,

$$\begin{aligned}
&\mathbb{P}((X_{11} - X_{21})(X_{12} - X_{32}) > 0) \\
&\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}(X_{11} > x_1, X_{12} > x_2) \\
&\qquad\qquad\qquad + \mathbb{P}(x_1 > X_{11}, x_2 > X_{12})\, dF_1(x_1)\, dF_2(x_2) \\
&\qquad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (1 - F_1(x_1) - F_2(x_2) + H(x_1, x_2)) \\
&\qquad\qquad\qquad + H(x_1, x_2)\, dF_1(x_1)\, dF_2(x_2) \\
&\qquad = \int_0^1\int_0^1 1 - u_1 - u_2 + 2C(u_1, u_2)\, du_1\, du_2 \\
&\qquad = 1 - \frac{1}{2} - \frac{1}{2} + 2\int_0^1\int_0^1 C(u_1, u_2)\, du_1\, du_2 \\
&\qquad = 2\int_0^1\int_0^1 C(u_1, u_2)\, du_1\, du_2.
\end{aligned}$$

The result follows by substituting in $(A.41)$,

$$\begin{aligned}
\rho(X_1, X_2) = \rho(C) &= 12\int\int_{[0,1]^2} C(u_1, u_2)\, du_1\, du_2 - 3 \\
&= 12\int\int_{[0,1]^2} C(u_1, u_2) - u_1 u_2\, du_1\, du_2. \qquad \text{(A.47)}
\end{aligned}$$

This result concludes the proof. □

Appendix B

Extra Material: Supplementary Figures and Tables

B.1 Additional Material for Chapter 2

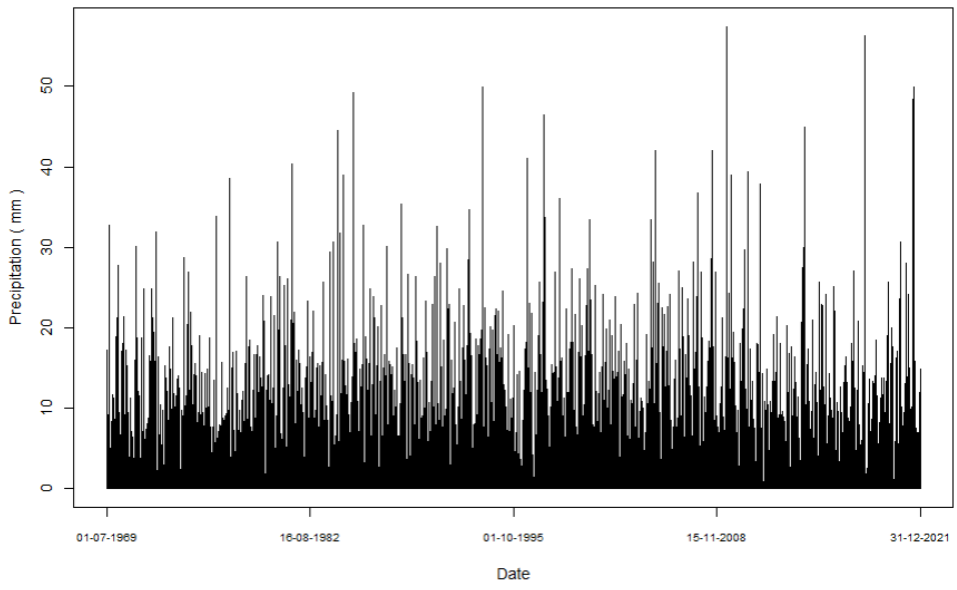


Figure B.1: Düsseldorf daily precipitation time series.

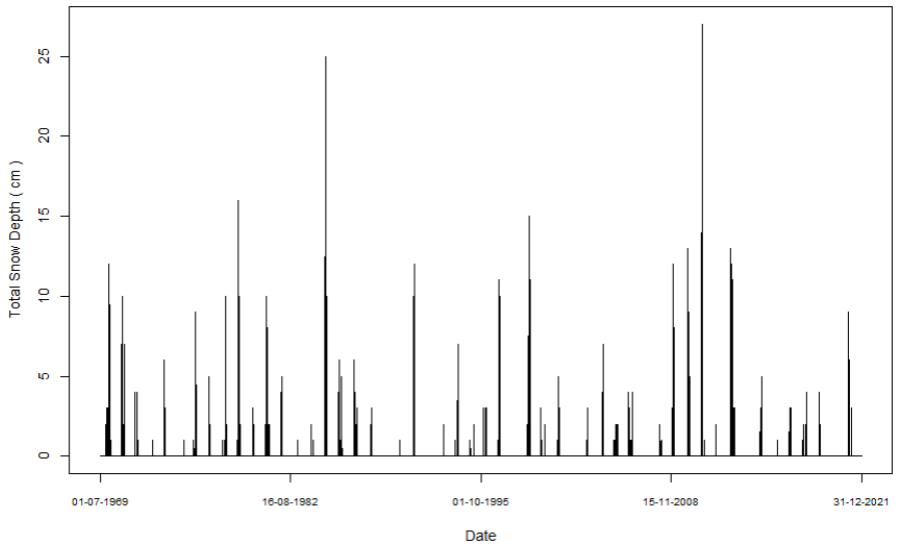


Figure B.2: Düsseldorf daily total snow depth time series.

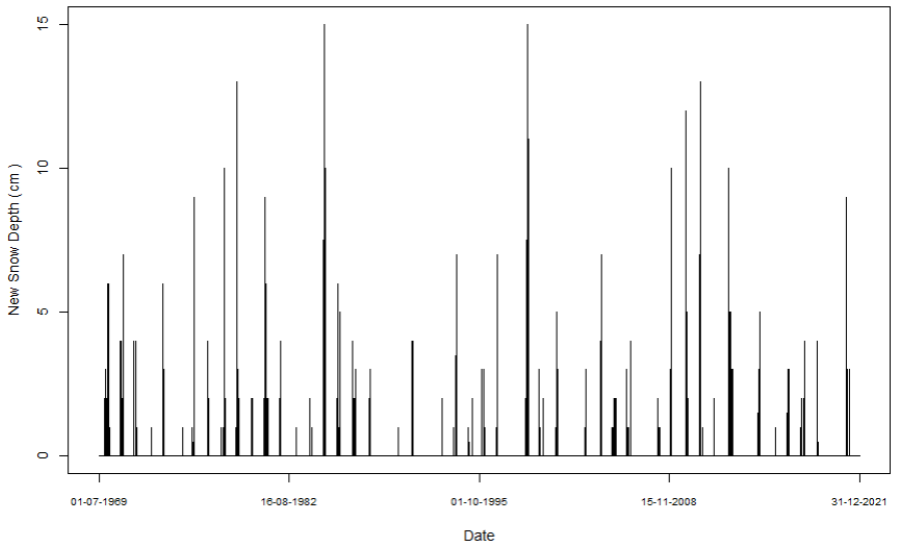


Figure B.3: Düsseldorf daily new snow depth time series.

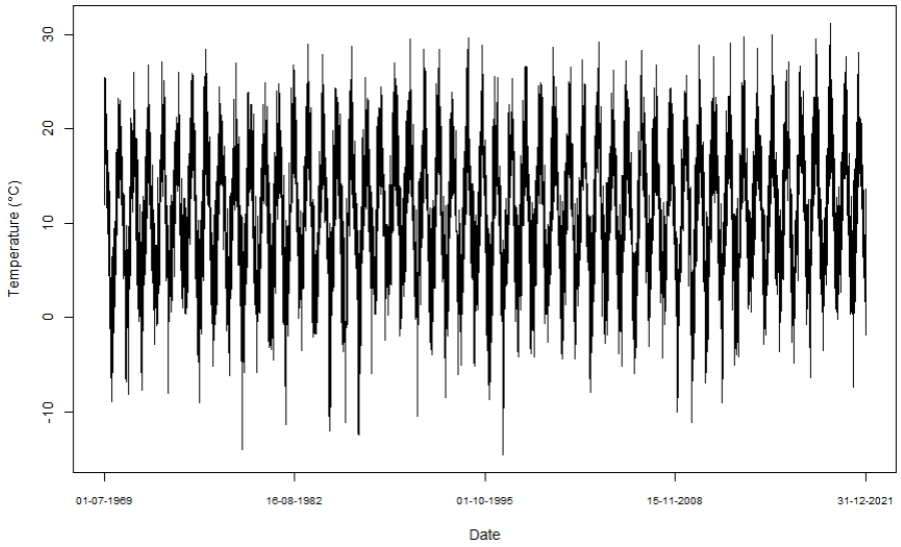


Figure B.4: Düsseldorf daily temperature time series.

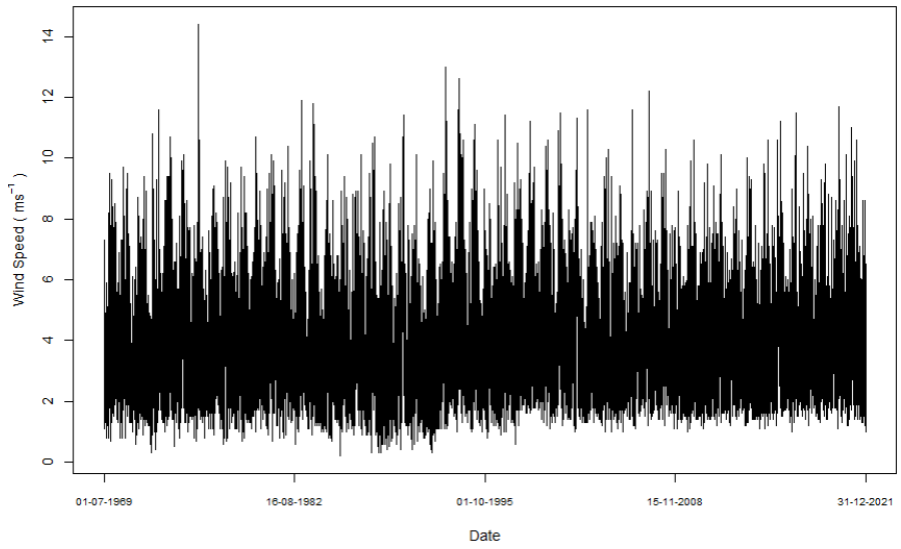


Figure B.5: Düsseldorf daily wind speed time series.

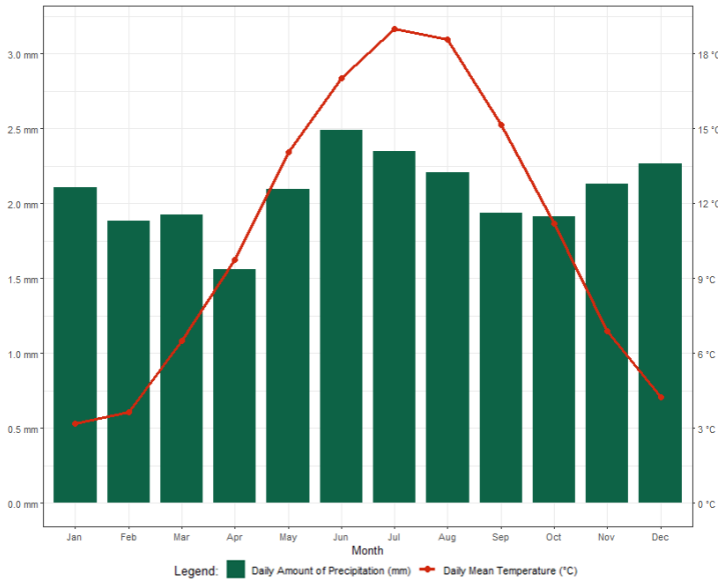


Figure B.6: Düsseldorf climate chart of the expected amount of daily precipitation and mean temperature per month from 1970 – 2021.

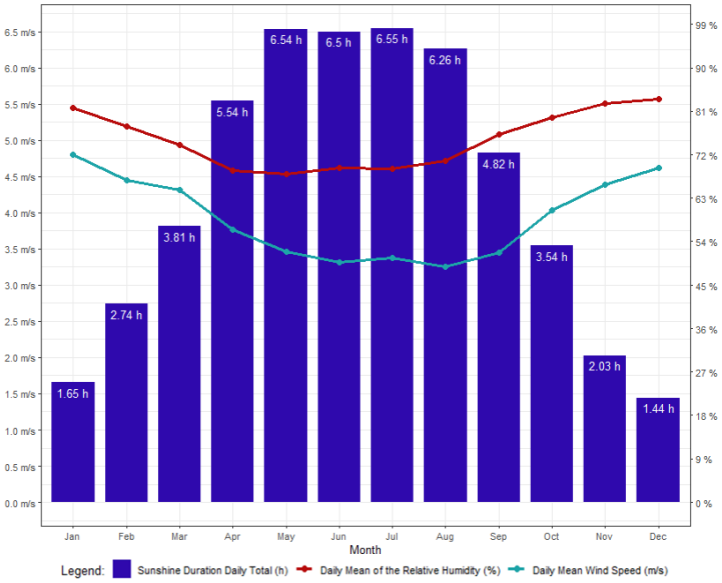


Figure B.7: Düsseldorf climate chart of the expected daily sunshine duration, mean humidity and mean wind speed per month from 1970 – 2021.

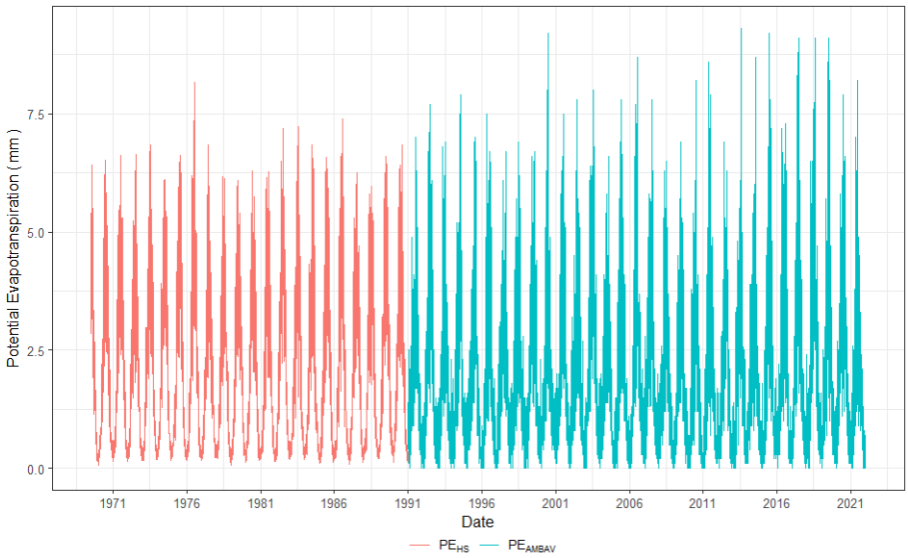


Figure B.8: Düsseldorf daily potential evapotranspiration time series.

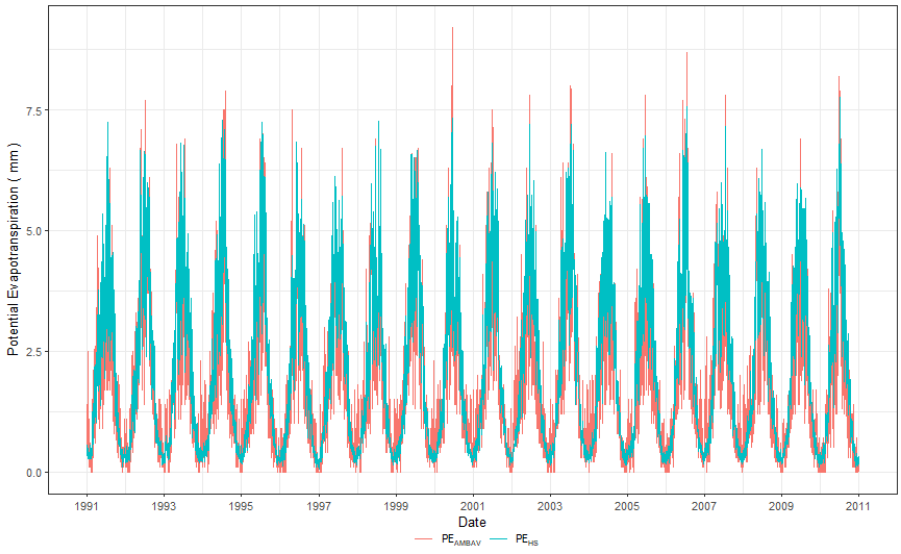


Figure B.9: Comparison of the potential evapotranspiration estimates.

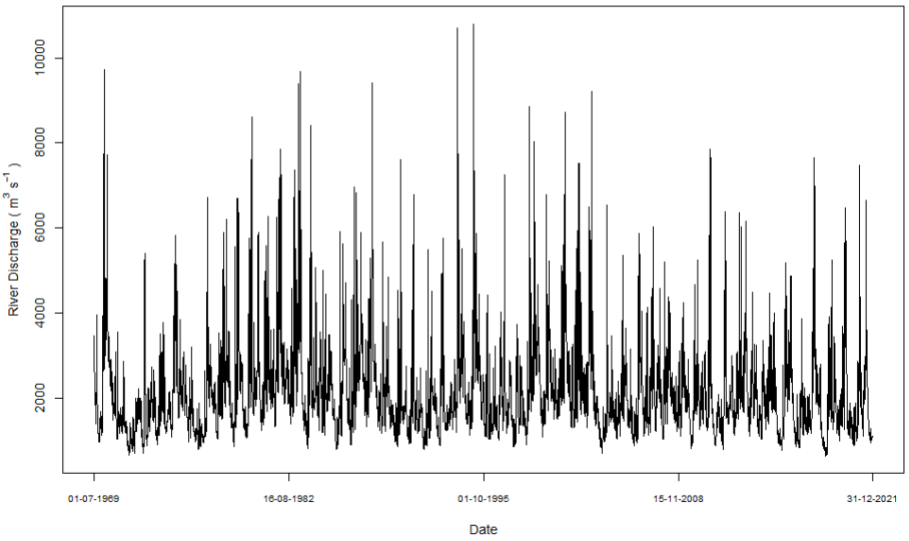


Figure B.10: Düsseldorf daily river discharge time series.

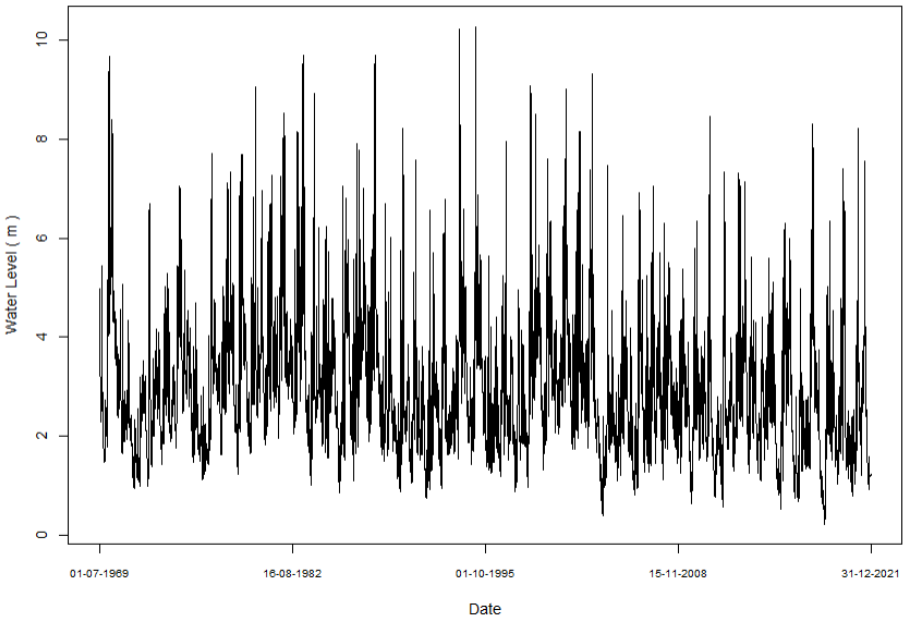


Figure B.11: Düsseldorf daily water level time series.

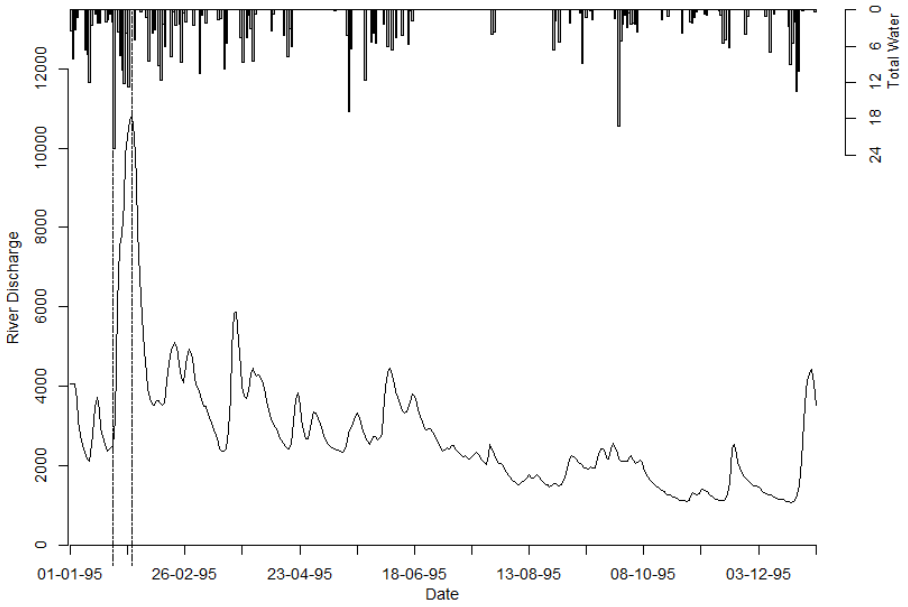


Figure B.12: The lag time in the 1995 Düsseldorf hydrograph.

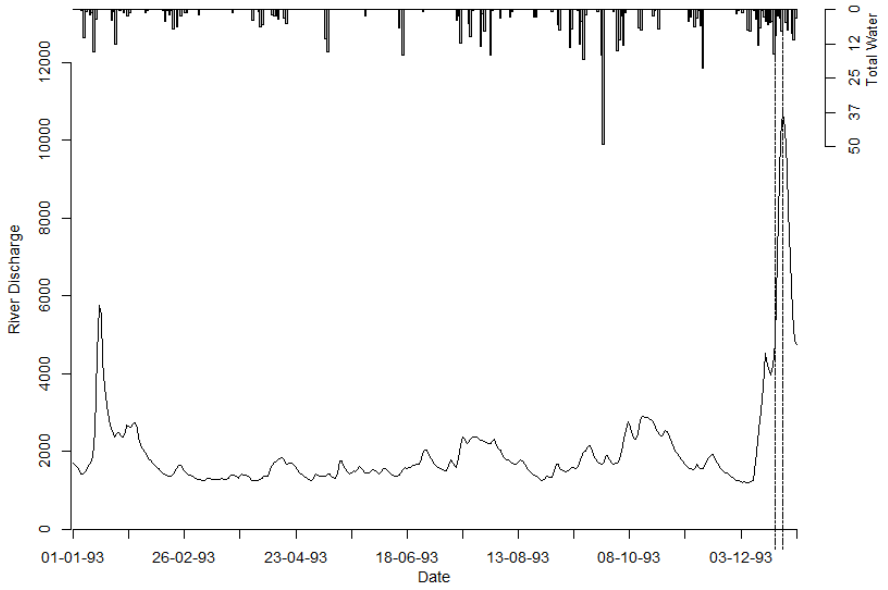


Figure B.13: The lag time in the 1993 Düsseldorf hydrograph.

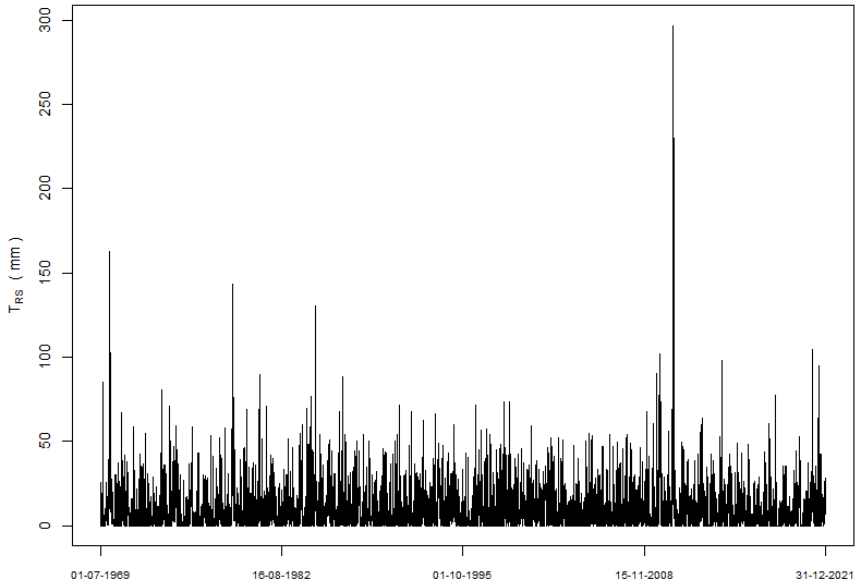
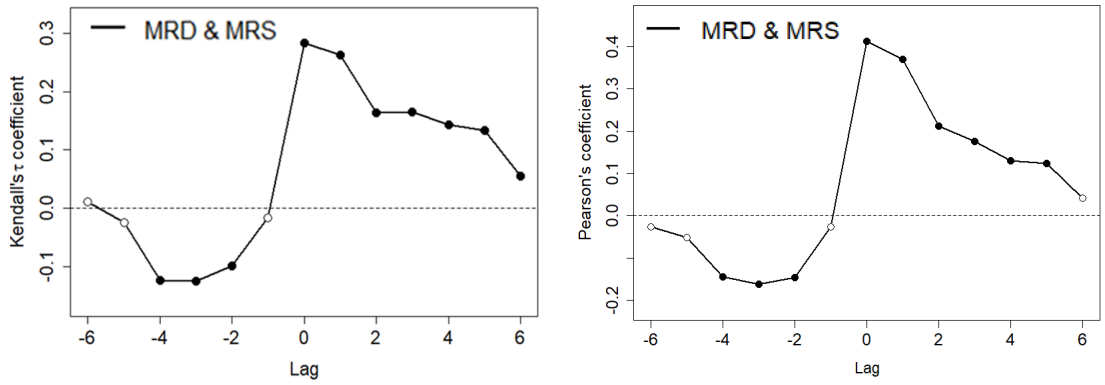


Figure B.14: Düsseldorf 7-Day rolling sum of the collective impact time series.



(a) (b)

Figure B.15: The graph in panel (a) corresponds to the Kendall's tau coefficient lag between the time series T_{MRD} and T_{MRS} and the graph in panel (b) corresponds to the Pearson's correlation coefficient lag between the the time series T_{MRD} and T_{MRS}.

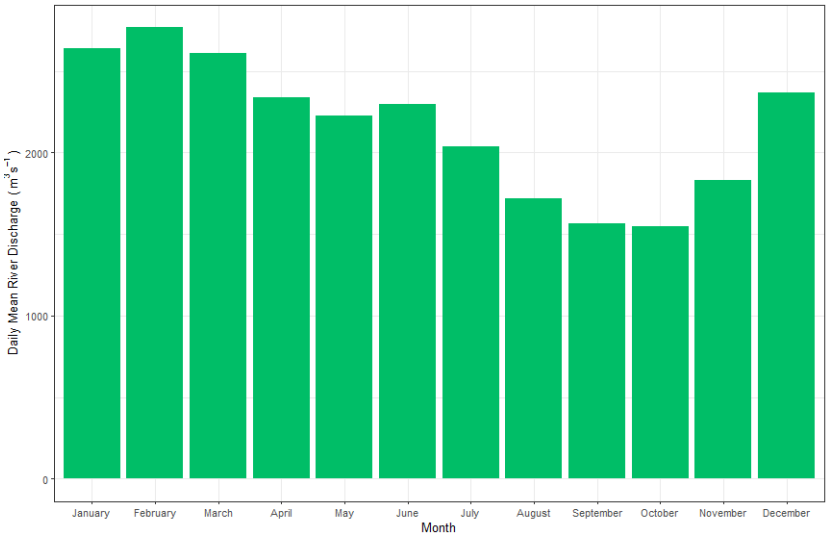


Figure B.16: Düsseldorf expected daily river discharge per month.

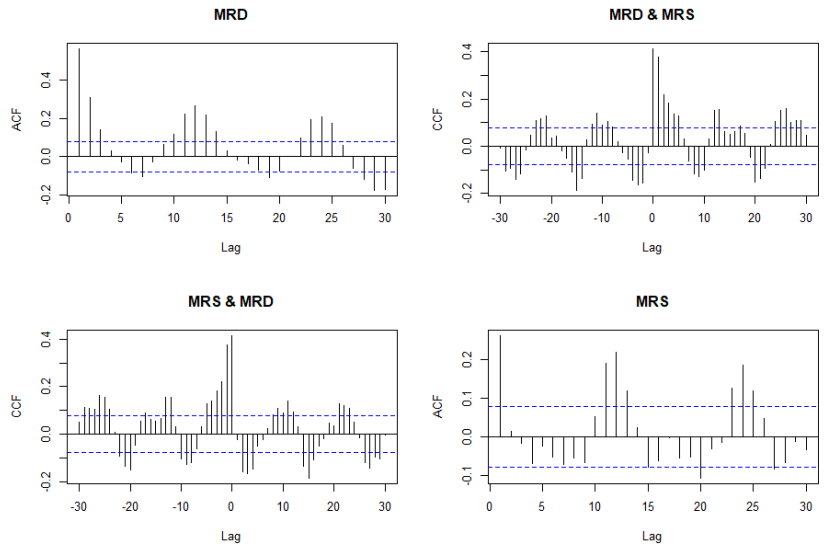


Figure B.17: The ACF and CCF plots for the T_{MRD} and T_{MRS} time series.

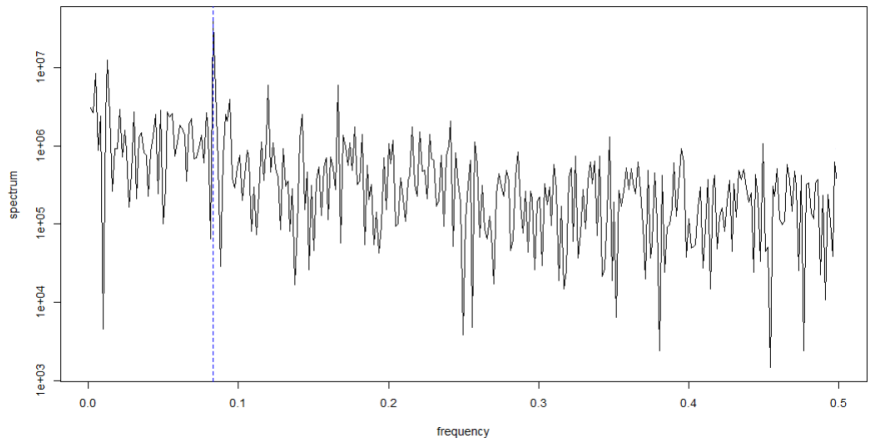


Figure B.18: Periodogram plot of the full time series T_{MRD}.

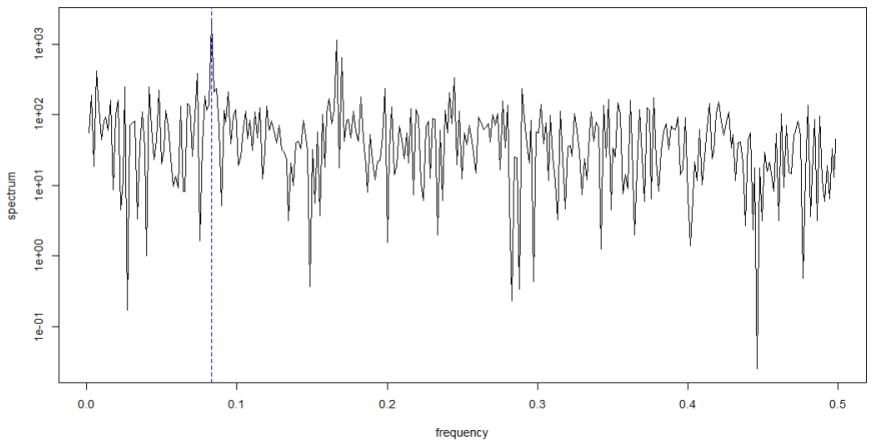


Figure B.19: Periodogram plot of the full time series T_{MRS}.

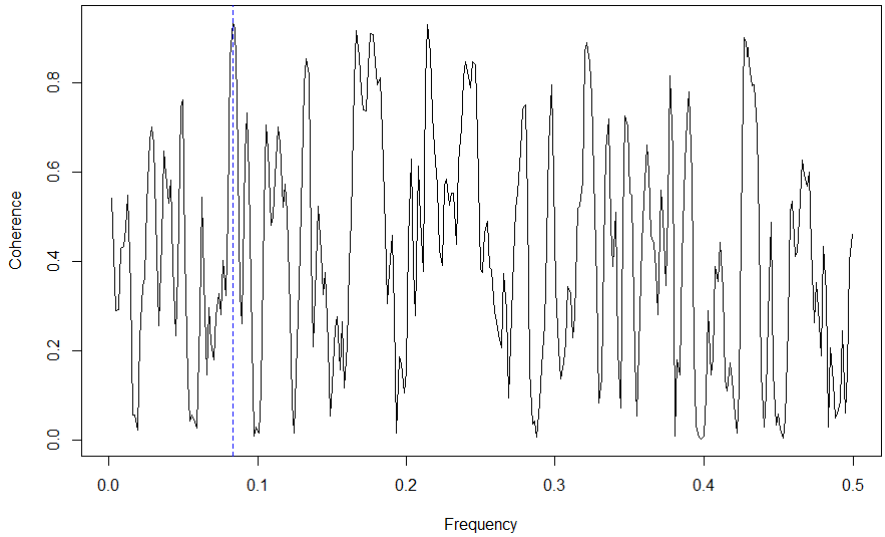


Figure B.20: The coherence spectrum plot for the time series T_{MRD} and T_{MRS}.

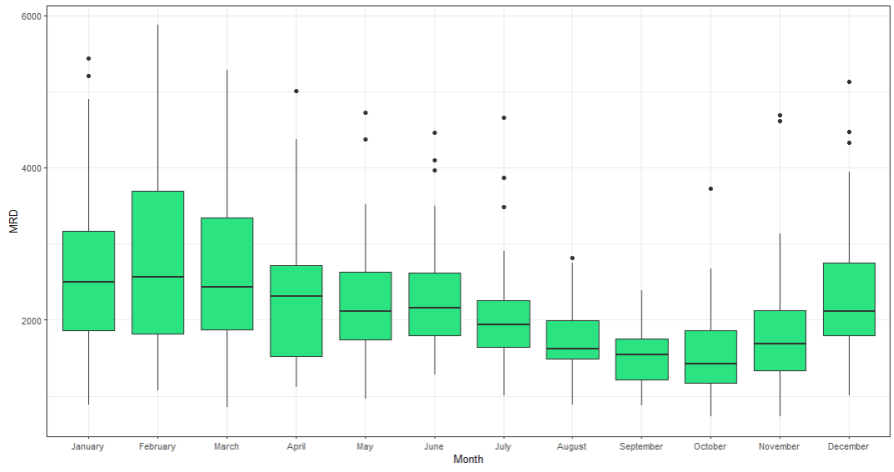


Figure B.21: Statistical summary for each month of the time series T_{MRD} over the time period $1969 - 2020$.

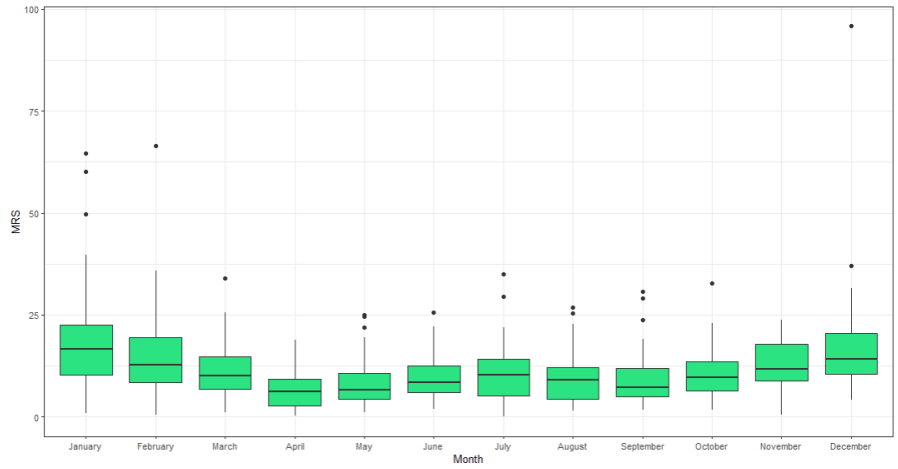


Figure B.22: Statistical summary for each month of the time series T_{MRS} over the time period $1969 - 2020$.

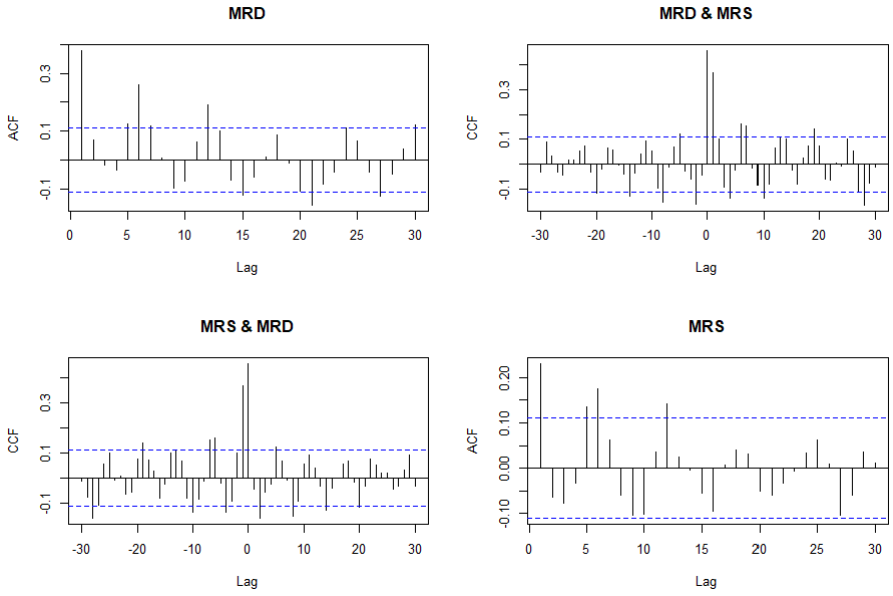


Figure B.23: The ACF and CCF plots for the T_{MRD} and T_{MRS} time series when considering the P1 time horizon.

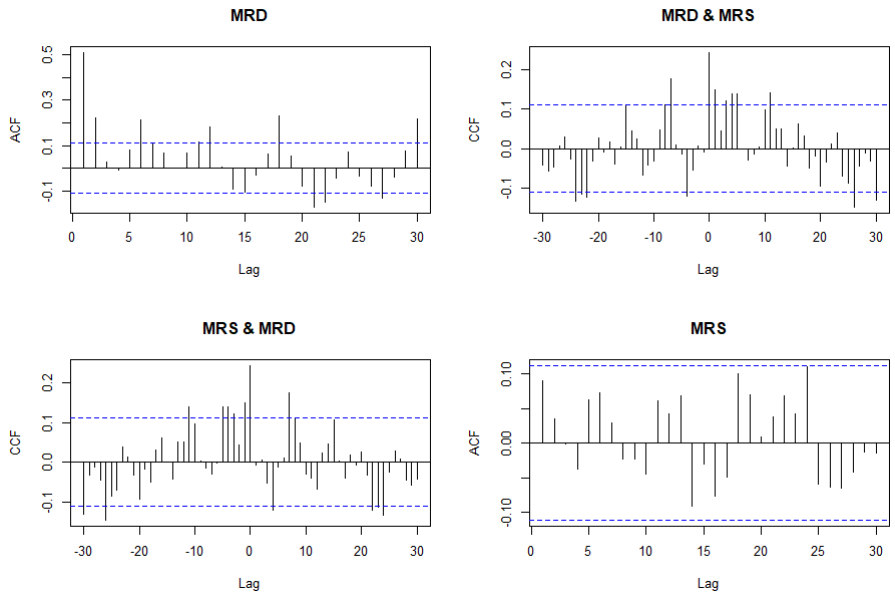


Figure B.24: The ACF and CCF plots for the T_{MRD} and T_{MRS} time series when considering the P2 time horizon.

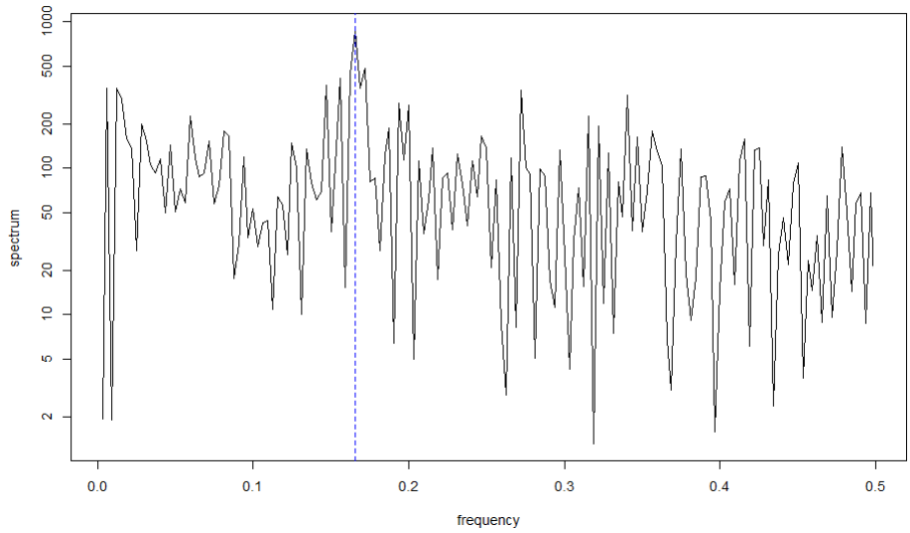


Figure B.25: The periodogram plot of the time series T_{MRS} covering the P1 time horizon.

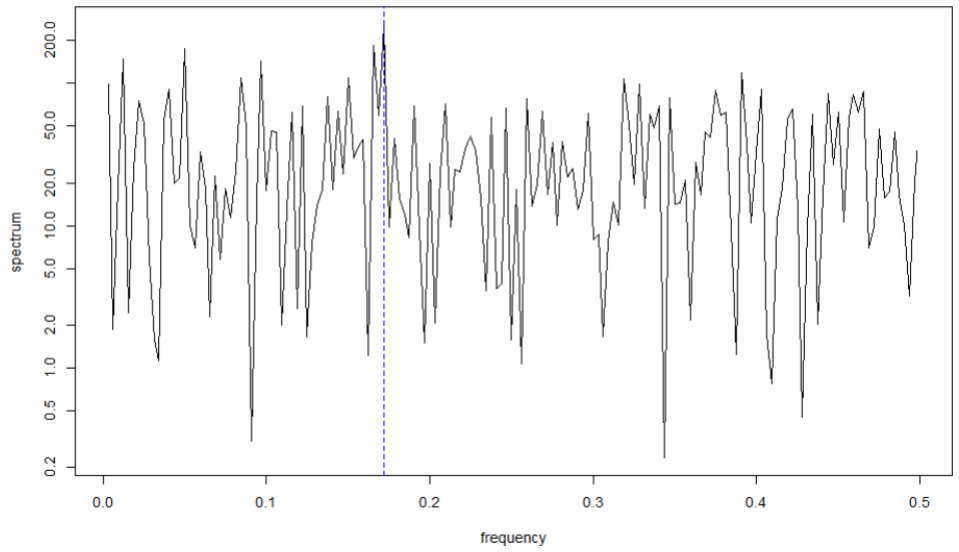


Figure B.26: The periodogram plot of the time series T_{MRS} covering the P2 time horizon.

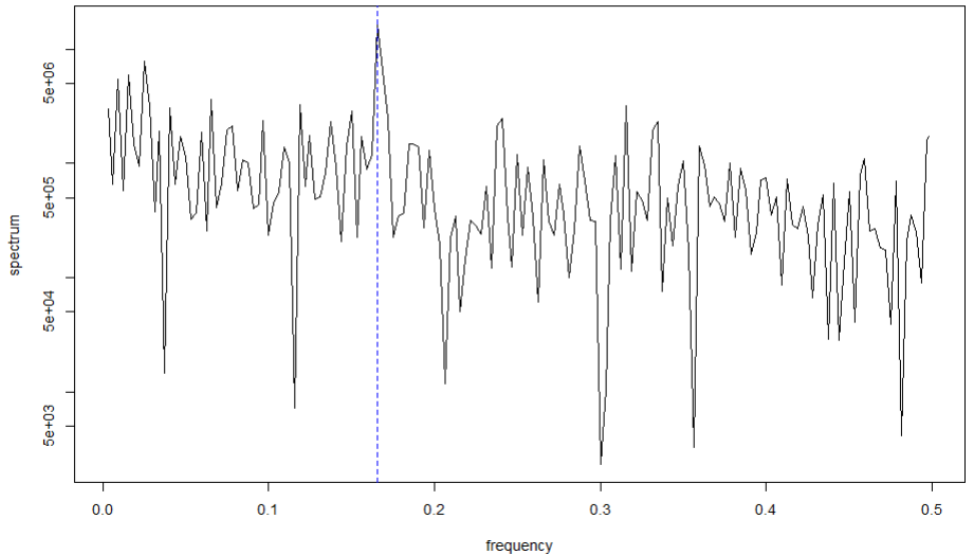


Figure B.27: The periodogram plot of the time series T_{MRD} covering the P1 time horizon.

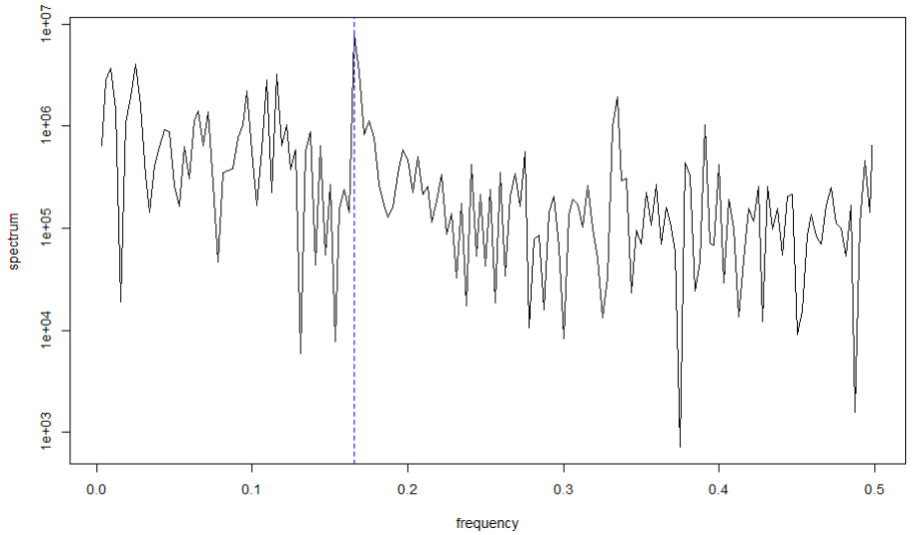


Figure B.28: The periodogram plot of the time series T_{MRD} covering the P2 time horizon.

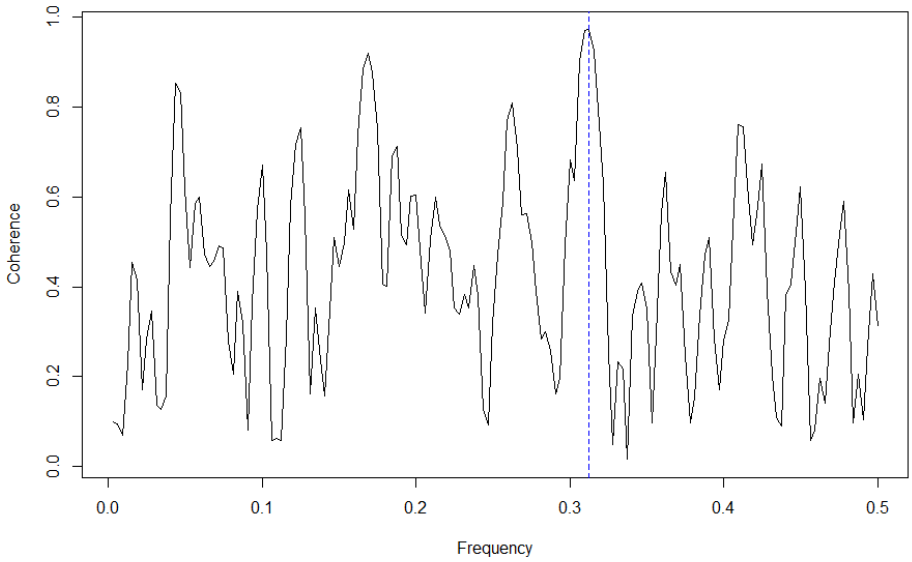


Figure B.29: The coherence spectrum plot of the time series T_{MRD} and T_{MRS} in the P2 time horizon.

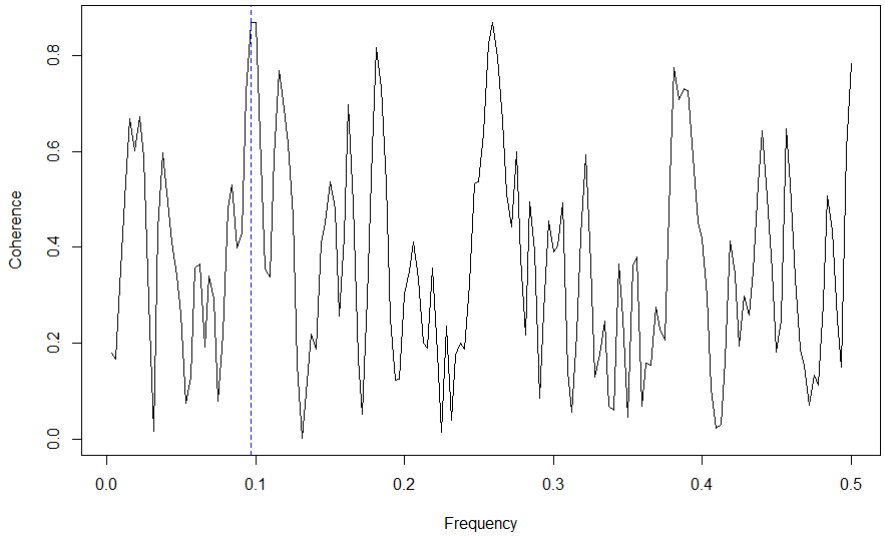


Figure B.30: The coherence spectrum plot of the time series T_{MRD} and T_{MRS} in the P1 time horizon.

B.2 Additional Material for Chapter 3

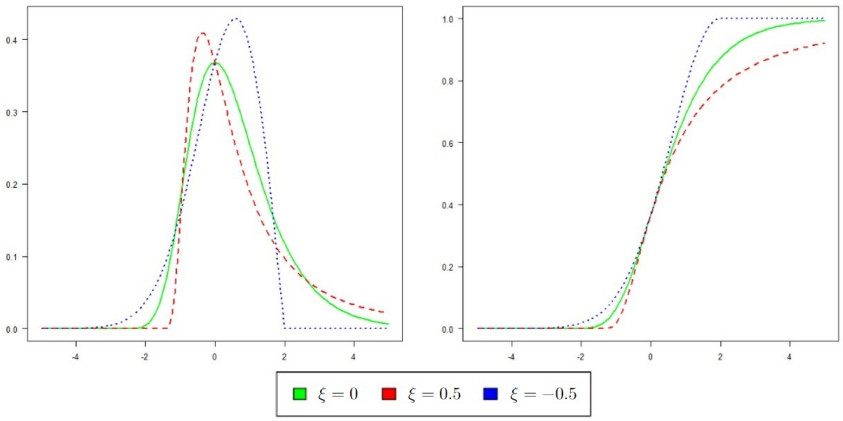


Figure B.31: The GEV density and cumulative distributions functions.

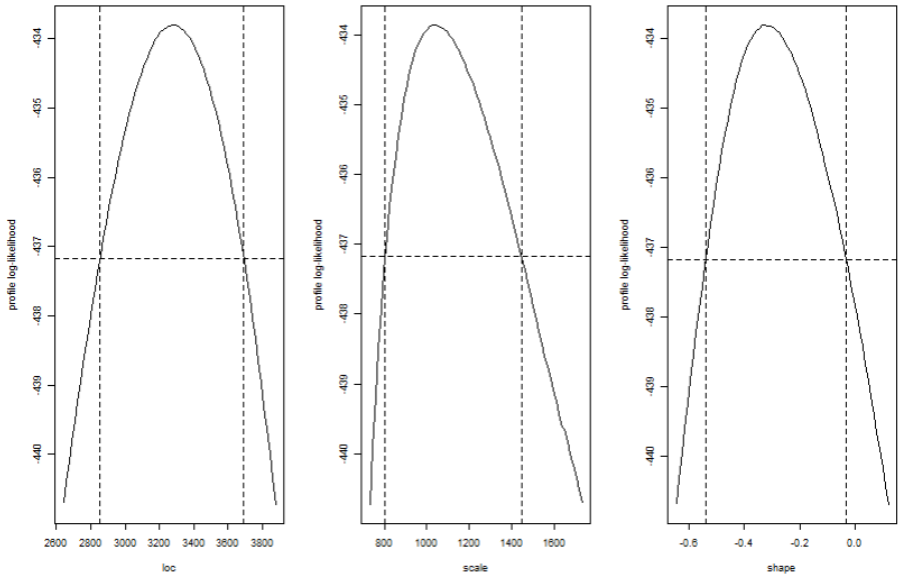


Figure B.32: The BM profile log-likelihood plots for each parameter estimate of the GEV distribution fit on the hydrological annual extremes of the time series T_{MRD} in the P1 time horizon.

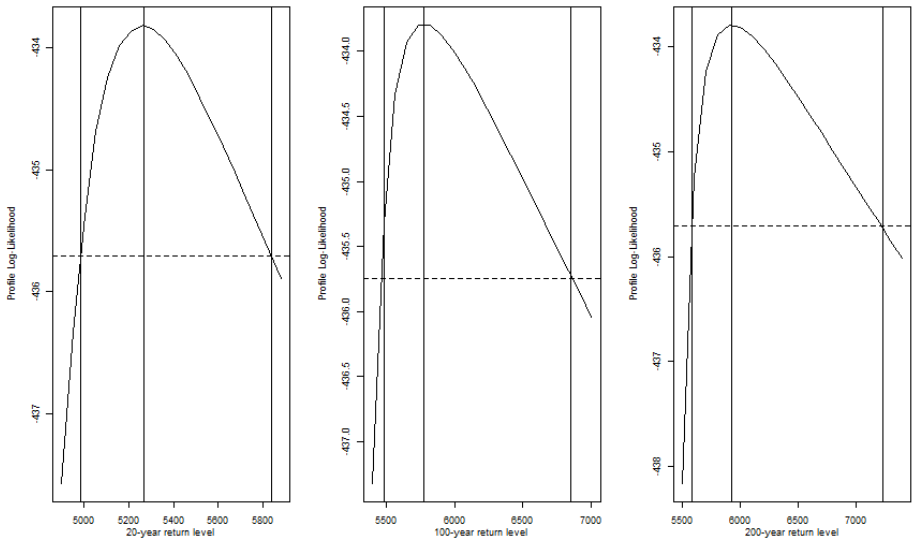


Figure B.33: The BM profile log-likelihood plots for the three return periods considered for the hydrological annual extremes of the time series T_{MRD} in the P1 time horizon.

		Parameters	95% lower CI	Estimate	95% upper CI
T_{MRD}	P1	ξ	-0.54	-0.33	-0.03
		σ	803.84	1055.62	1444.59
		μ	2853.56	3258.75	3691.14
	P2	ξ	0	0	0
		σ	589.48	723.30	908.7
		μ	2179.55	2384.82	2602.16
T_{MRS}	P1	ξ	0.10	0.27	0.51
		σ	5.15	6.52	8.47
		μ	17.85	19.73	21.84
	P2	ξ	0	0	0
		σ	4.11	5.04	6.32
		μ	13.30	14.74	16.25

Table B.1: Summary of the GEV parameter estimates and their corresponding confidence intervals, evaluated on the hydrological annual extremes of the time series T_{MRD} and T_{MRS} for both time horizons.

		Return Period	95% lower CI	Return Levels	95% upper CI
T_{MRD}	P1	1 in 20 year event	4984.24	5269.34	5838.01
		1 in 100 year event	5484.21	5777.78	6847.66
		1 in 200 year event	5600.00	5925.04	7228.14
	P2	1 in 20 year event	4043.22	4533.15	5185.77
		1 in 100 year event	4904.75	5712.09	6519.44
		1 in 200 year event	5302.11	6215.26	7128.42
T_{MRS}	P1	1 in 20 year event	39.99	49.53	72.04
		1 in 100 year event	55.89	79.53	155.60
		1 in 200 year event	73.97	96.98	169.36
	P2	1 in 20 year event	26.31	29.72	34.26
		1 in 100 year event	32.32	37.94	43.55
		1 in 200 year event	35.10	41.45	47.79

Table B.2: Summary of the BM return levels and the corresponding confidence intervals for the three return periods considered for the hydrological annual extremes of the time series T_{MRD} and T_{MRS} for both time horizons.

		AIC	BIC
T_{MRD}	P1	873.61	879.47
	P2	853.79	857.69
T_{MRS}	P1	381.53	387.39
	P2	336.26	340.16

Table B.3: Summary of the adequacy of the BM model fit for each time series T_{MRD} and T_{MRS} for both time horizons.

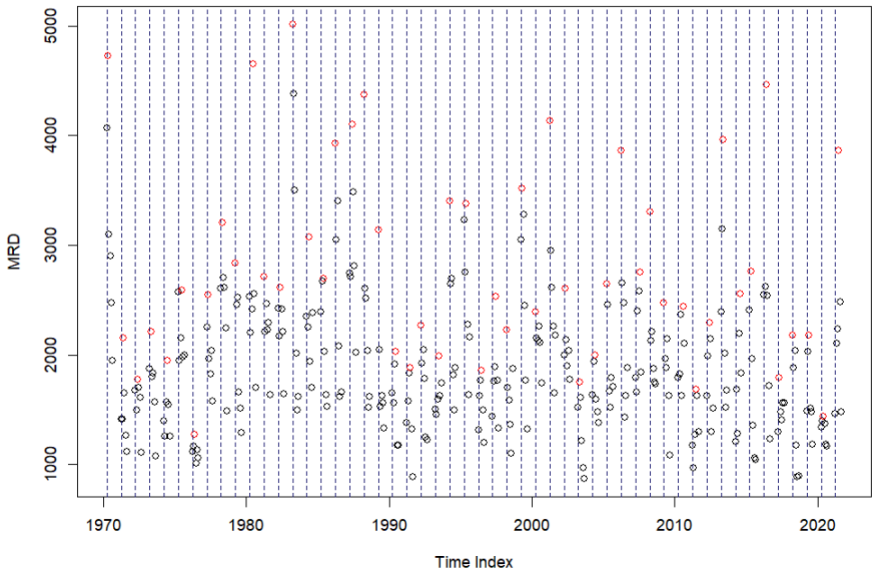


Figure B.34: The MRD hydrological annual maxima observations are shown in red, while the hydrological annual blocks are shown in blue for the P2 time horizon..

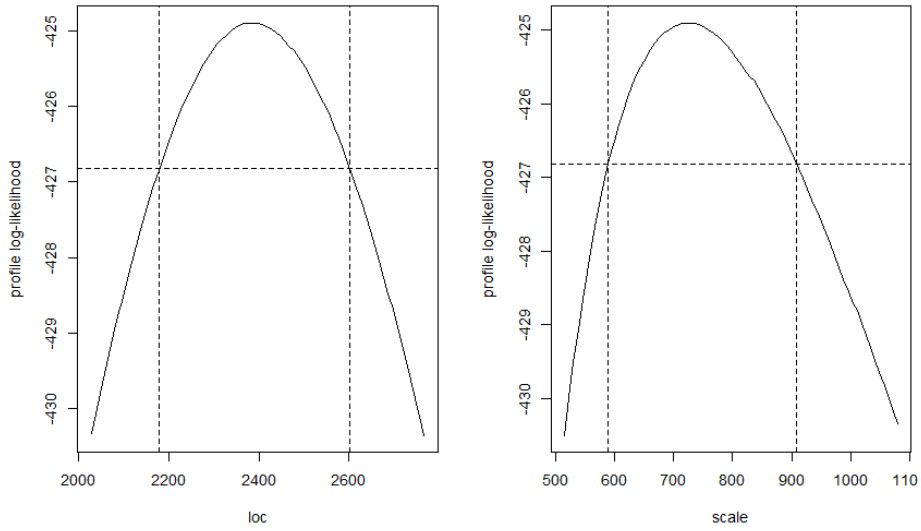


Figure B.35: The BM profile log-likelihood plots for each parameter estimate of the GEV distribution fit on the hydrological annual extremes of the time series T_{MRD} in the P2 time horizon.

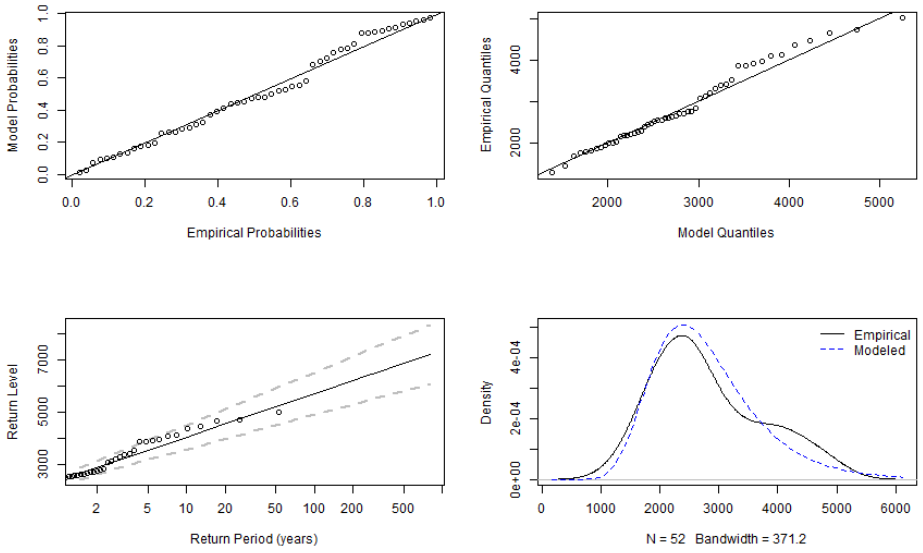


Figure B.36: The diagnostic plots of the GEV distribution fit on the hydrological annual maxima observations of the time series T_{MRD} in the P2 time horizon.

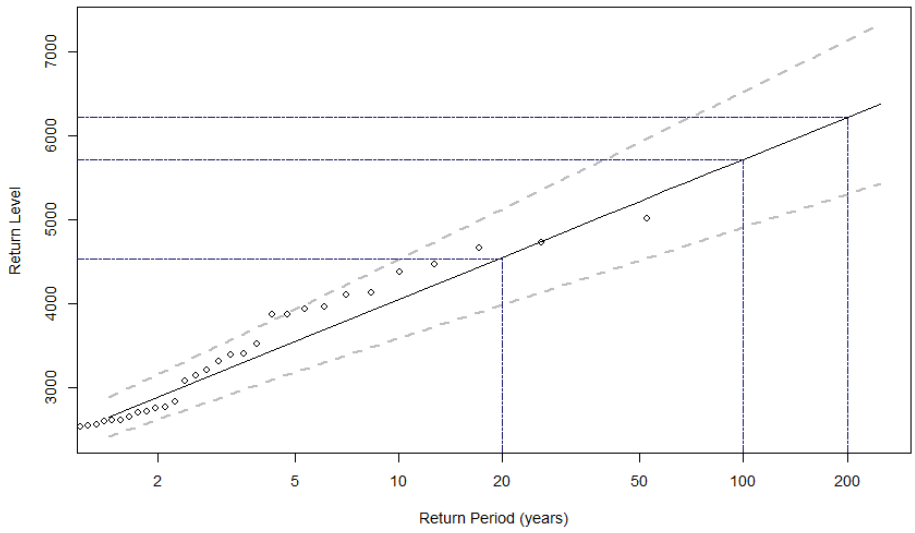


Figure B.37: The BM return level curve, with the 95% profile likelihood confidence intervals represented by grey dotted lines, and the empirical return levels denoted by black circles, is shown for the time series T_{MRD} in the P2 time horizon. The estimate flood scenarios shown in Table B.2 are displayed by the blue dotted lines.

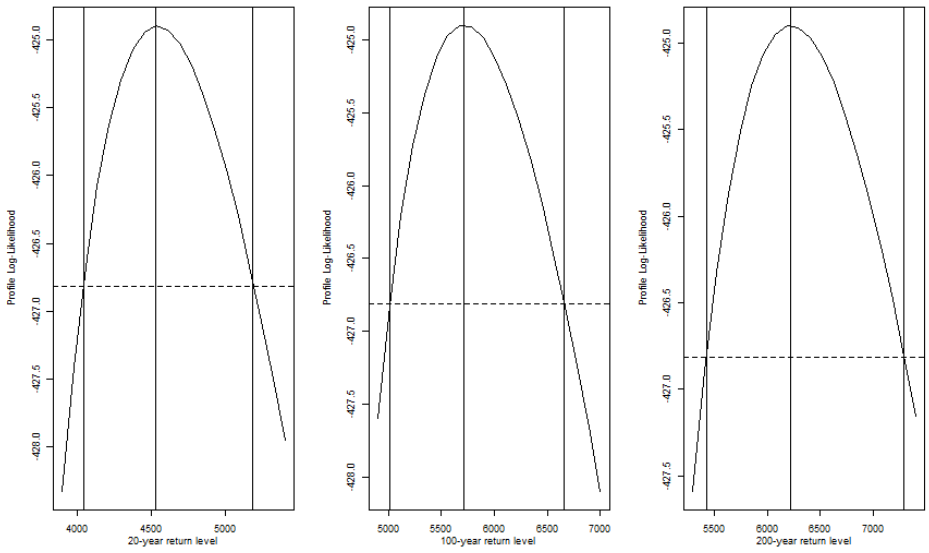


Figure B.38: The BM profile log-likelihood plots for each parameter estimate of the GEV distribution fit on the hydrological annual extremes of the time series T_{MRD} in the P2 time horizon.

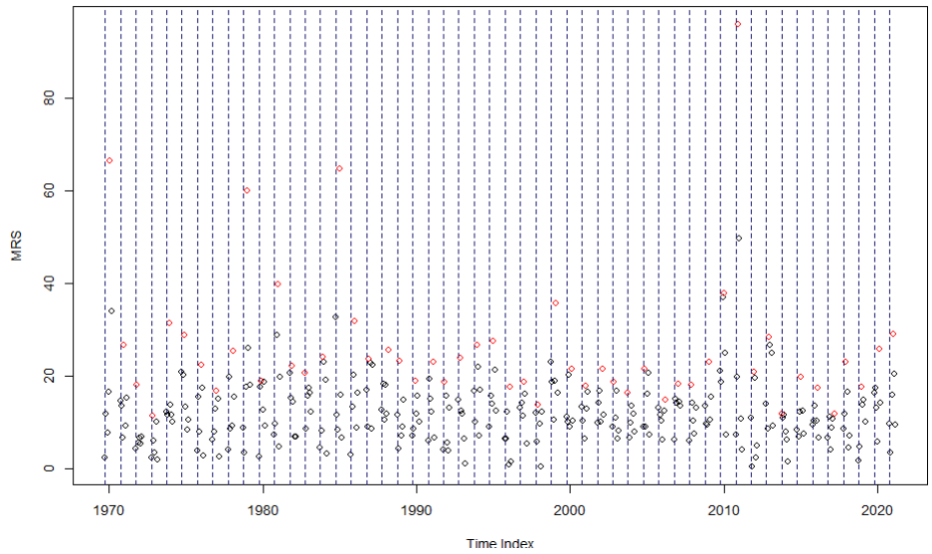


Figure B.39: The MRS hydrological annual maxima observations are shown in red, while the hydrological annual blocks are shown in blue for the P1 time horizon.

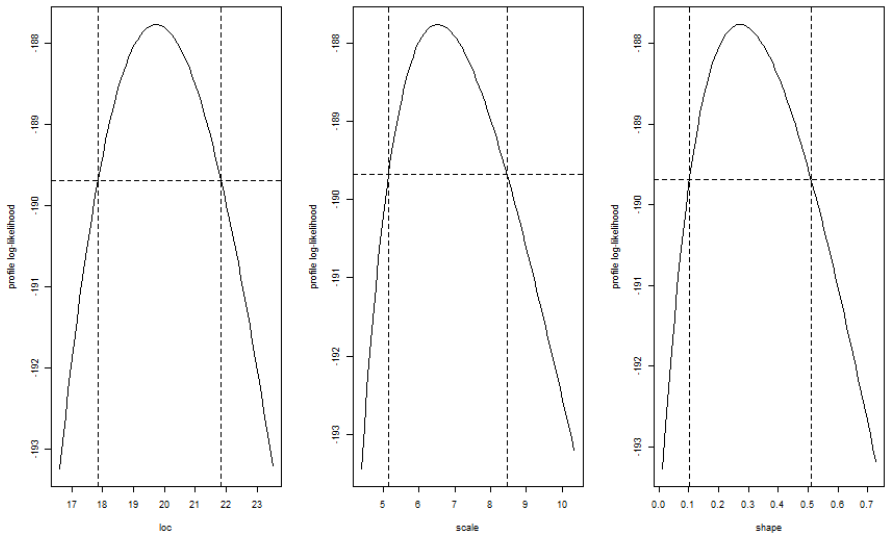


Figure B.40: The BM profile log-likelihood plots for each parameter estimate of the GEV distribution fit on the hydrological annual extremes of the time series T_{MRS} in the P1 time horizon.

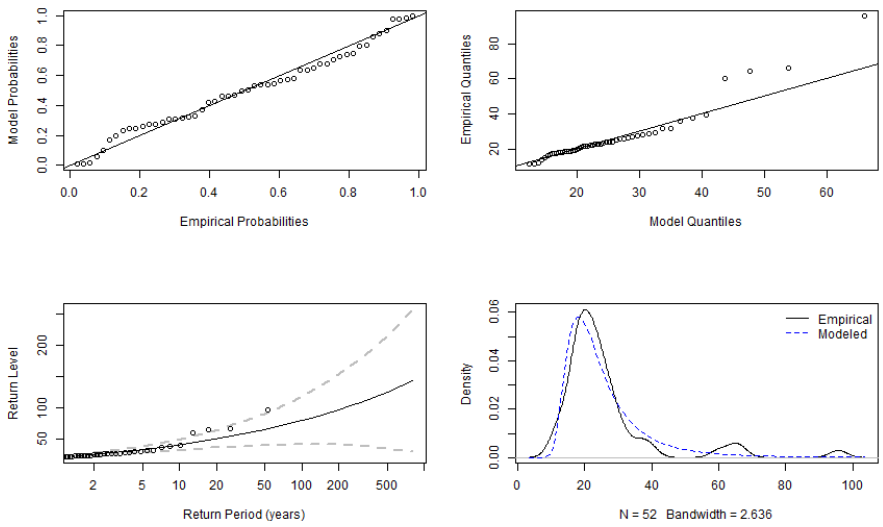


Figure B.41: The diagnostic plots of the GEV distribution fit on the hydrological annual maxima observations of the time series T_{MRS} in the P1 time horizon.

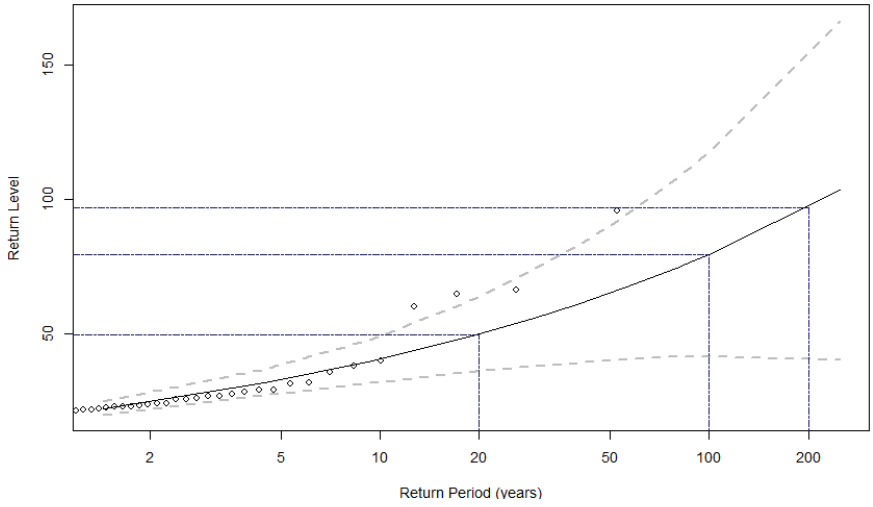


Figure B.42: The BM return level curve, with the 95% profile likelihood confidence intervals represented by grey dotted lines, and the empirical return levels denoted by black circles, is shown for the time series T_{MRS} in the P1 time horizon. The estimate flood scenarios shown in Table B.2 are displayed by the blue dotted lines.

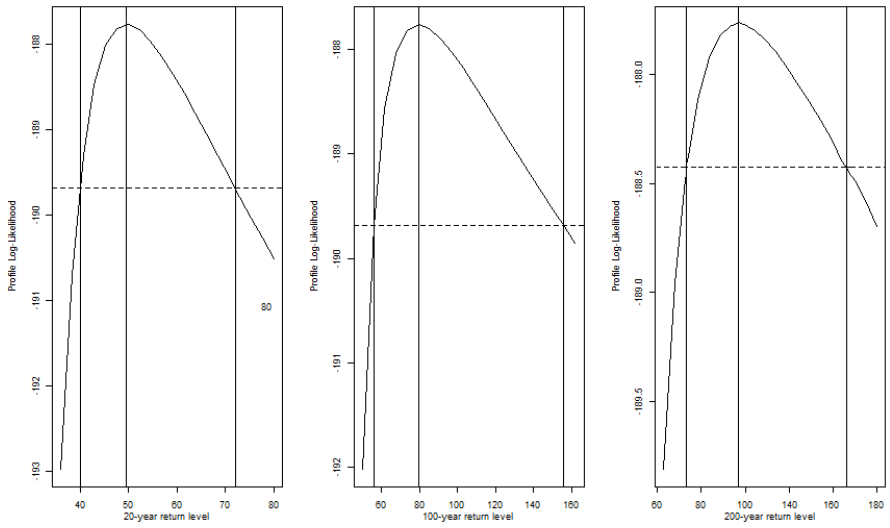


Figure B.43: The BM profile log-likelihood plots for the three return periods considered for the hydrological annual extremes of the time series T_{MRS} in the P1 time horizon.

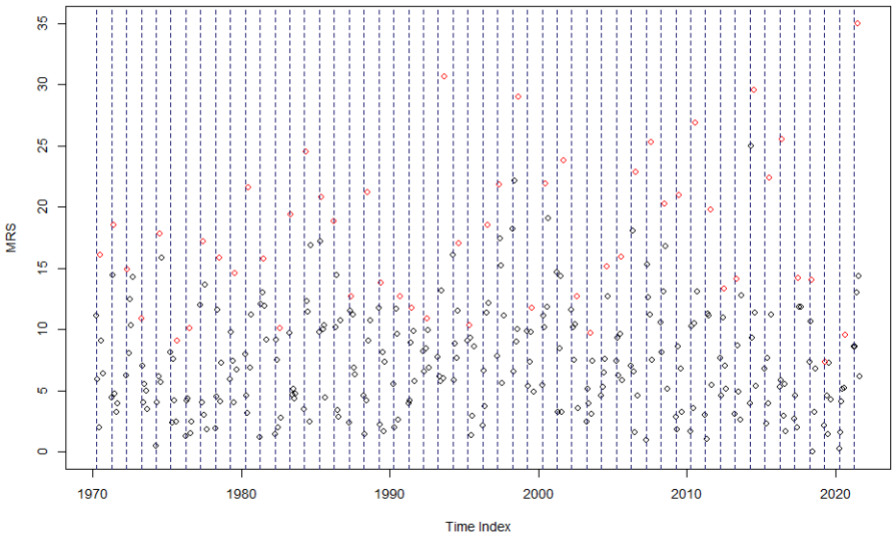


Figure B.44: The MRS hydrological annual maxima observations are shown in red, while the hydrological annual blocks are shown in blue for the P2 time horizon.

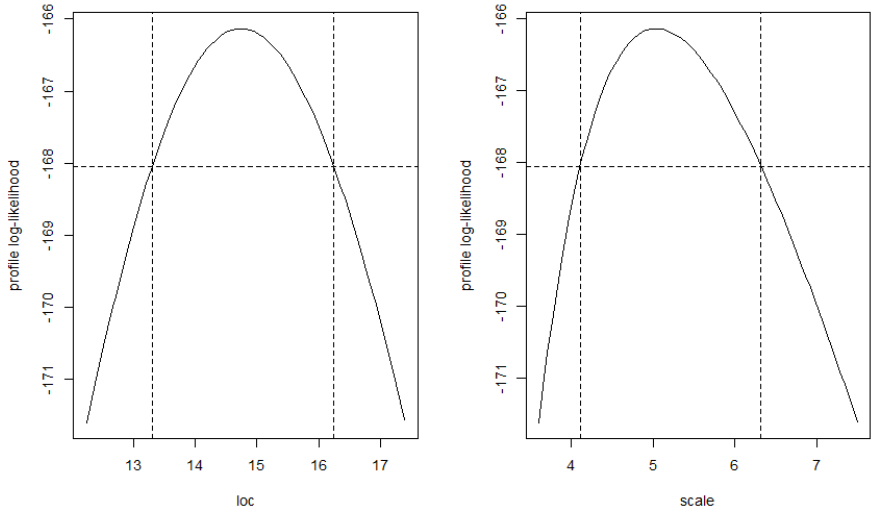


Figure B.45: The BM profile log-likelihood plots for each parameter estimate of the GEV distribution fit on the hydrological annual extremes of the time series T_{MRS} in the P2 time horizon.

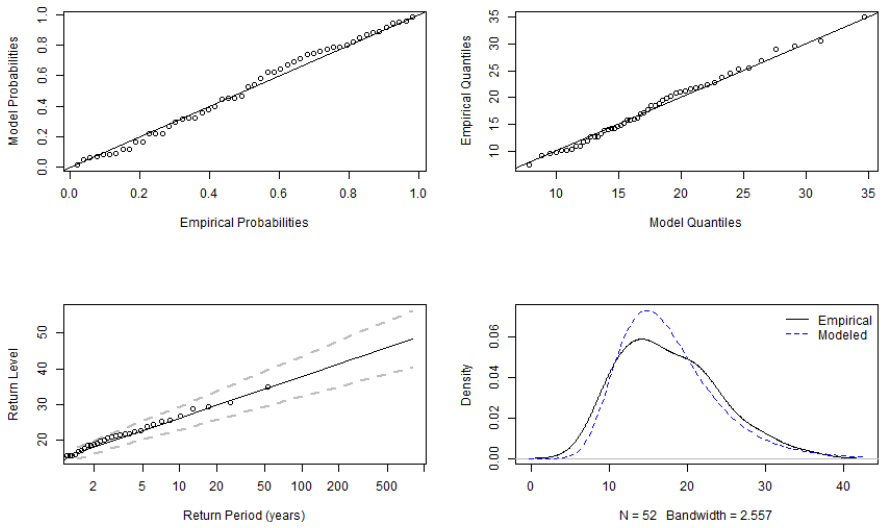


Figure B.46: The diagnostic plots of the GEV distribution fit on the hydrological annual maxima observations of the T_{MRS} time series in the P2 time horizon.

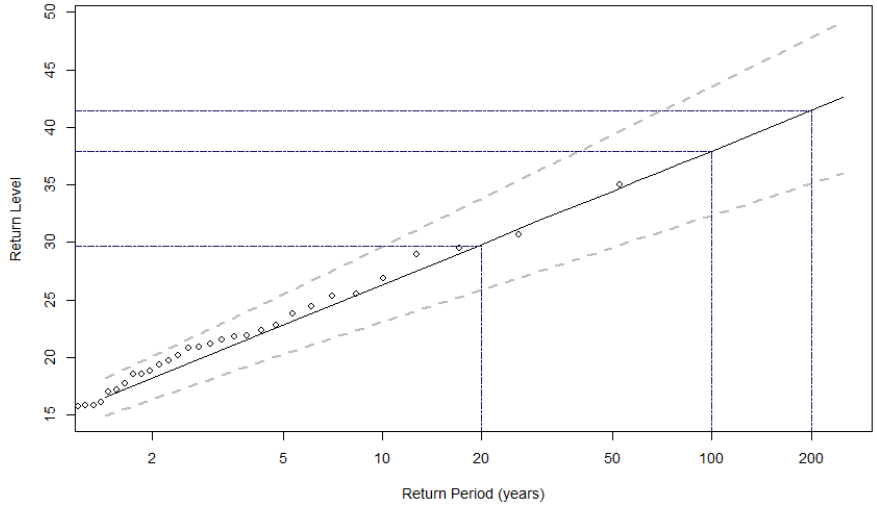


Figure B.47: The BM return level curve, with the 95% profile likelihood confidence intervals represented by grey dotted lines, and the empirical return levels denoted by black circles, is shown for the time series T_{MRS} in the P2 time horizon. The estimate flood scenarios shown in Table 3.2 are displayed by the blue dotted lines.

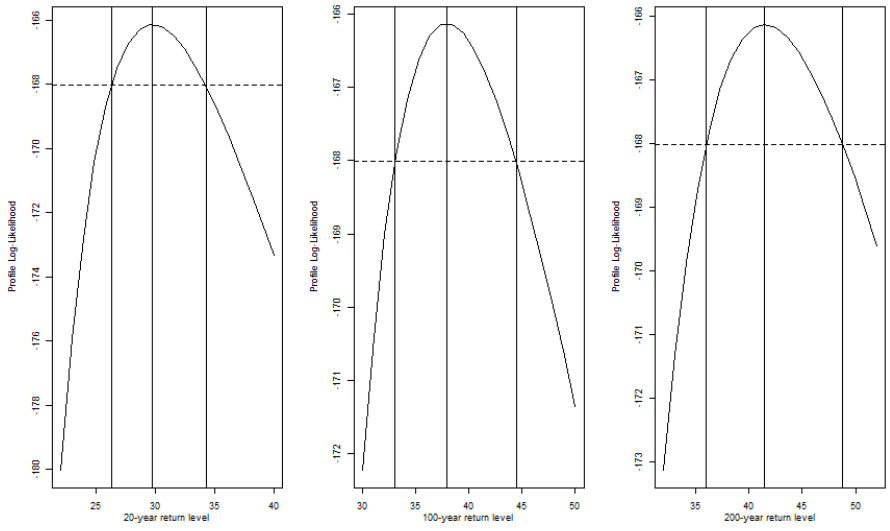


Figure B.48: The BM profile log-likelihood plots for the three return periods considered for the hydrological annual extremes of the time series T_{MRS} in the P2 time horizon.

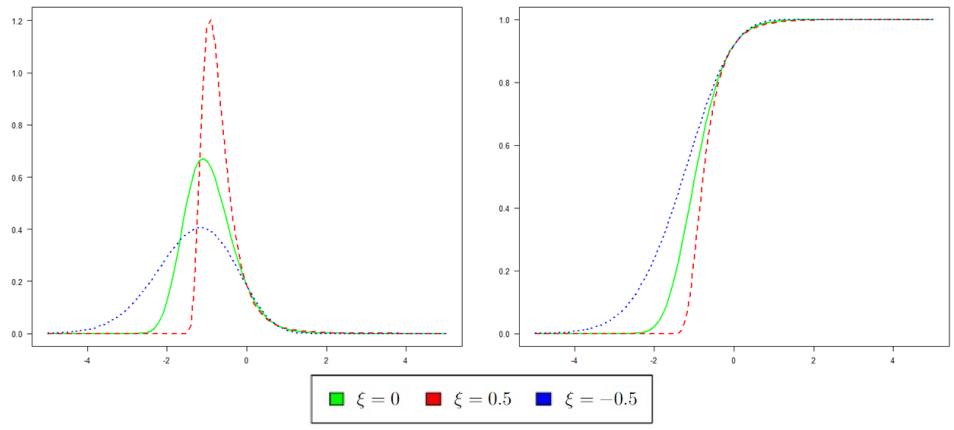


Figure B.49: The family of the 3rd order statistic probability density functions and cumulative distribution functions.

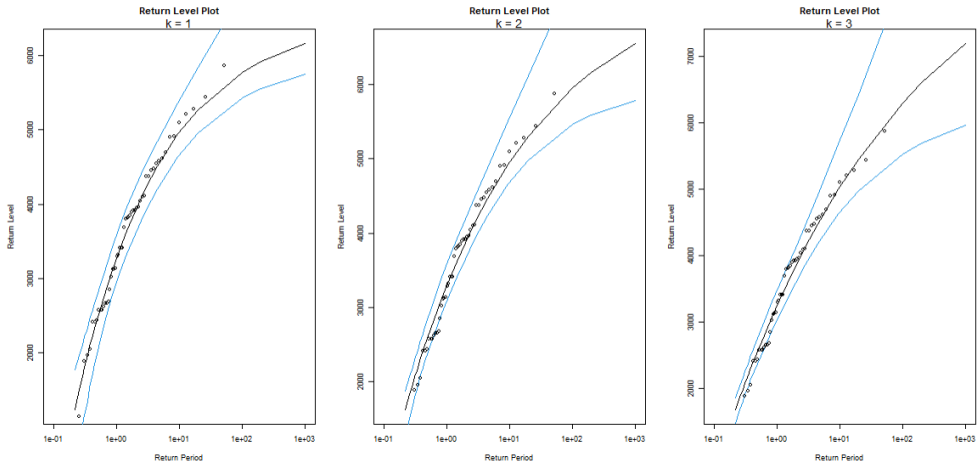


Figure B.50: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{1, 2, 3\}$ largest order statistics of the time series T_{MRD} in the P1 time horizon.

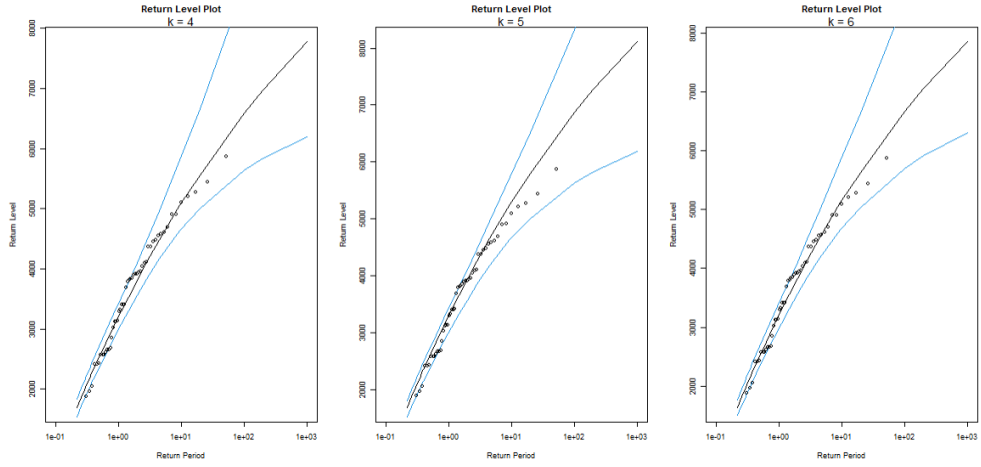


Figure B.51: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{4, 5, 6\}$ largest order statistics of the time series T_{MRD} in the P1 time horizon.

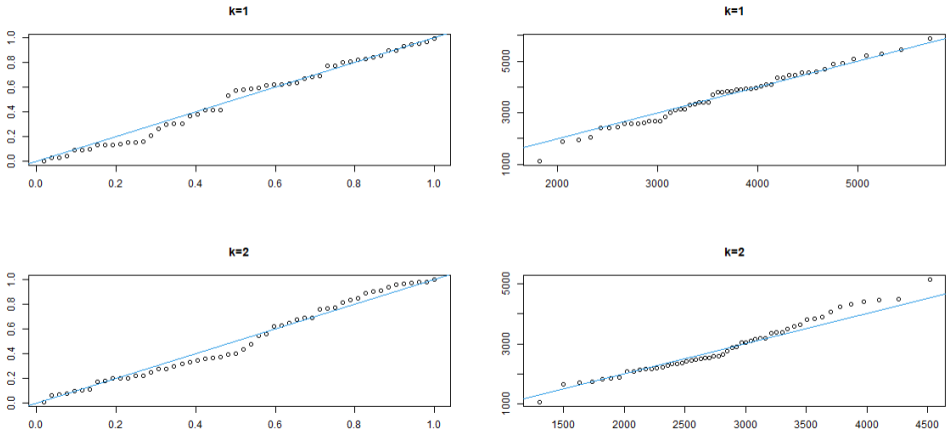


Figure B.52: The model diagnostic plots for the hydrological annual $k \in \{1, 2\}$ largest order statistics observations of the time series T_{MRD} in the P1 time horizon. The left column contains the probability plots, and the right column contains the quantile plots.

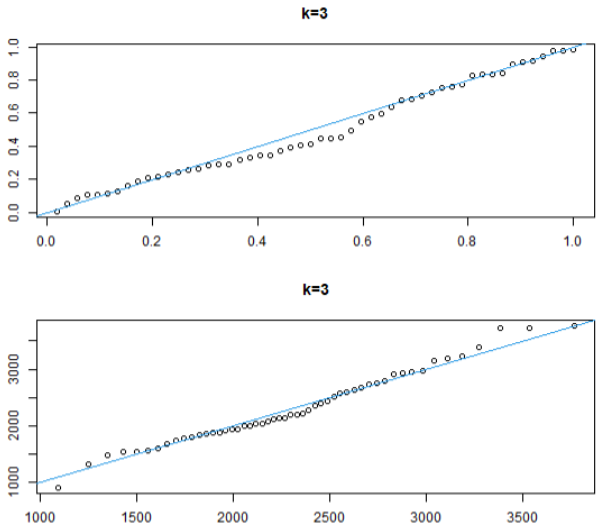


Figure B.53: The model diagnostic plots for the 3^{rd} ordered maxima of the time series T_{MRD} in the P1 time horizon. The top row contains the probability plots, and the bottom row contains the quantile plots.

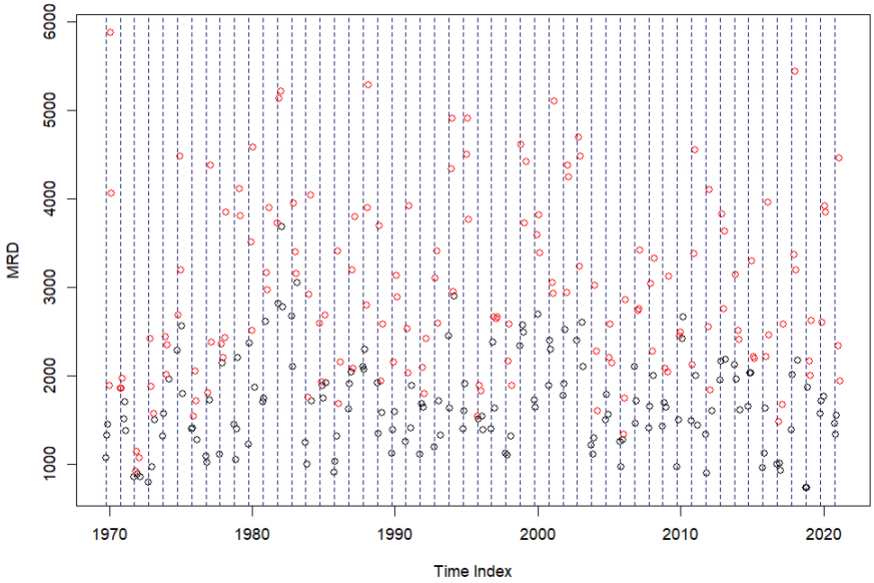


Figure B.54: The MRD hydrological annual top 3 extreme observations are shown in red, while the hydrological annual blocks are shown in blue for the P1 time horizon.

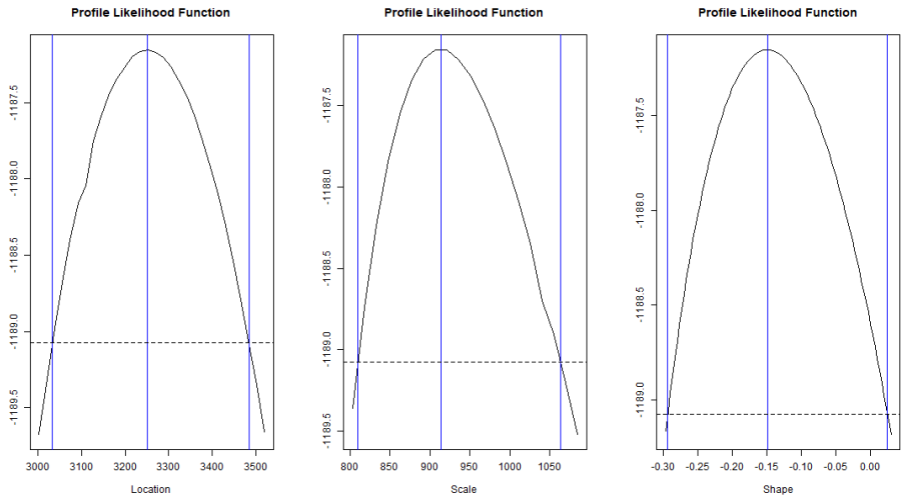


Figure B.55: The KLOS profile log-likelihood plots for each parameter estimate of the GEV_3 distribution fit on the hydrological annual top 3 extreme observations of the time series T_{MRD} in the P1 time horizon.

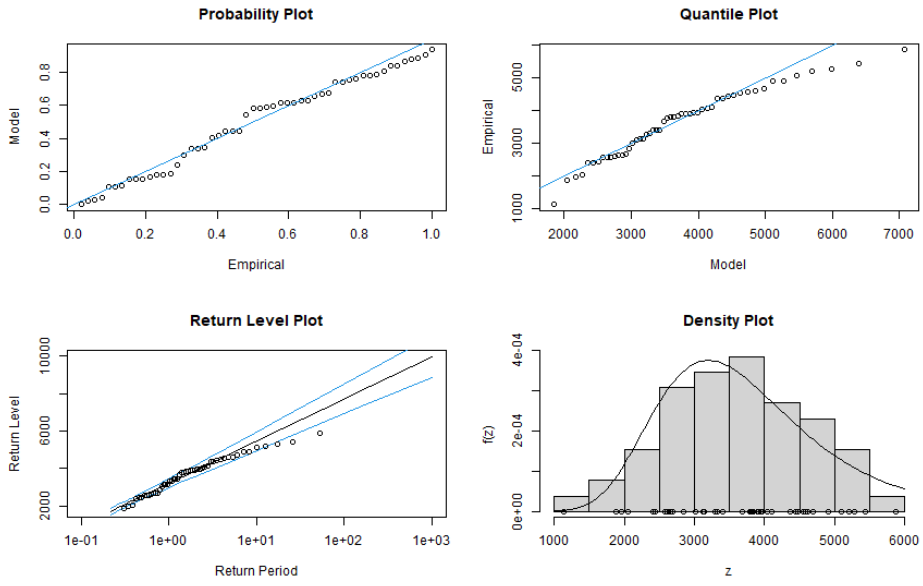


Figure B.56: The diagnostic plots of the GEV_3 distribution fit on the hydrological annual top 3 extreme observations of the time series T_{MRD} in the P1 time horizon.

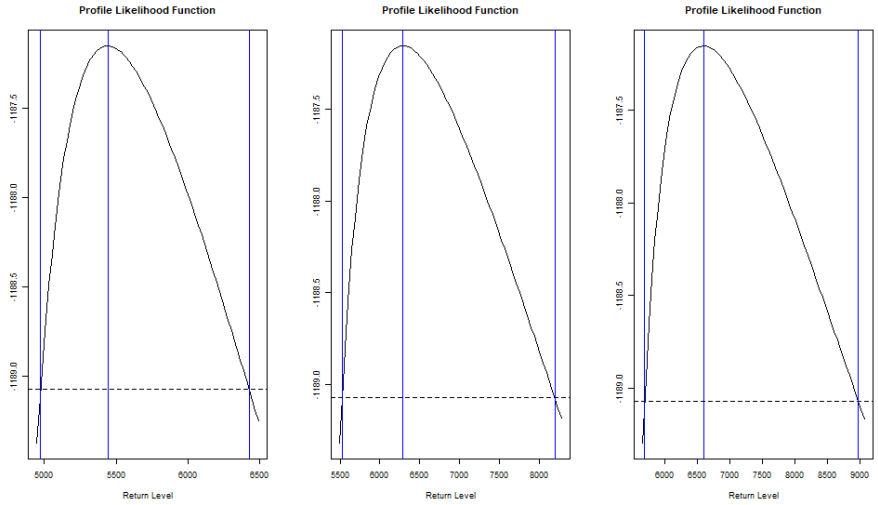


Figure B.57: The KLOS profile log-likelihood plots for the three return periods considered for the hydrological annual top 3 extreme observations of the time series T_{MRD} in the P1 time horizon.

		k	AIC	BIC
T_{MRD}	P1	3	2380.31	2380.05
	P2	2	1568.44	1568.19
T_{MRS}	P1	3	853.06	852.81
	P2	3	763.65	763.39

Table B.4: Summary of the adequacy of the KLOS model fit for each time series T_{MRD} and T_{MRS} for both time horizons.

		Parameters	95% lower CI	Estimate	95% upper CI
T_{MRD}	P1	ξ	-0.29	-0.15	0.03
		σ	809.53	913.22	1063.38
		μ	3032.40	3250.37	3428.52
	P2	ξ	-0.12	0.03	0.21
		σ	566.75	663.34	794.89
		μ	2388.70	2551.64	2733.54
T_{MRS}	P1	ξ	0.04	0.14	0.26
		σ	6.46	7.58	9.16
		μ	19.46	21.10	23.04
	P2	ξ	-0.15	-0.02	0.14
		σ	4.54	5.21	6.17
		μ	13.15	14.35	15.68

Table B.5: Summary of the GEV_k parameter estimates and their corresponding confidence intervals, evaluated on the hydrological annual top k extreme observations of the time series T_{MRD} and T_{MRS} for both time horizons.

		Return Period	95% lower CI	Return Levels	95% upper CI
T_{MRD}	P1	1 in 20 year event	4973.91	5442.17	6425.93
		1 in 100 year event	5525.87	6290.67	8200.14
		1 in 200 year event	5689.13	6594.52	8983.38
	P2	1 in 20 year event	4087.78	4614.68	5679.79
		1 in 100 year event	4816.80	5829.51	8293.20
		1 in 200 year event	5082.81	6366.82	9706.56
T_{MRS}	P1	1 in 20 year event	41.77	49.00	61.98
		1 in 100 year event	55.54	70.02	100.61
		1 in 200 year event	61.71	80.56	115.39
	P2	1 in 20 year event	25.77	29.34	36.55
		1 in 100 year event	30.75	37.15	52.74
		1 in 200 year event	32.51	40.41	60.92

Table B.6: Summary of the KLOS return levels and their corresponding confidence intervals for the three return periods considered for the hydrological annual top k extreme observations of the time series T_{MRD} and T_{MRS} for both time horizons.

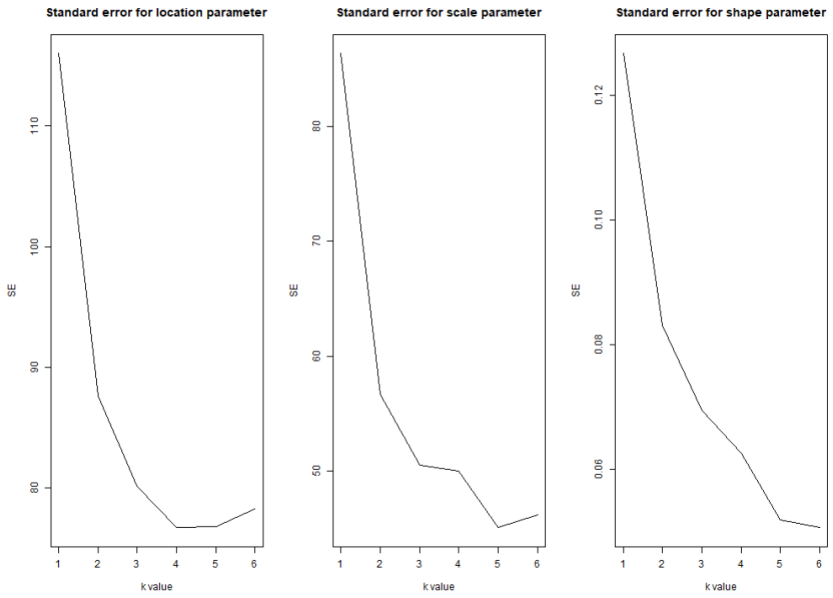


Figure B.58: The MLE standard errors for the GEV_k distribution fit on the hydrological annual top k extreme observations of the time series T_{MRD} in the P2 time horizon.

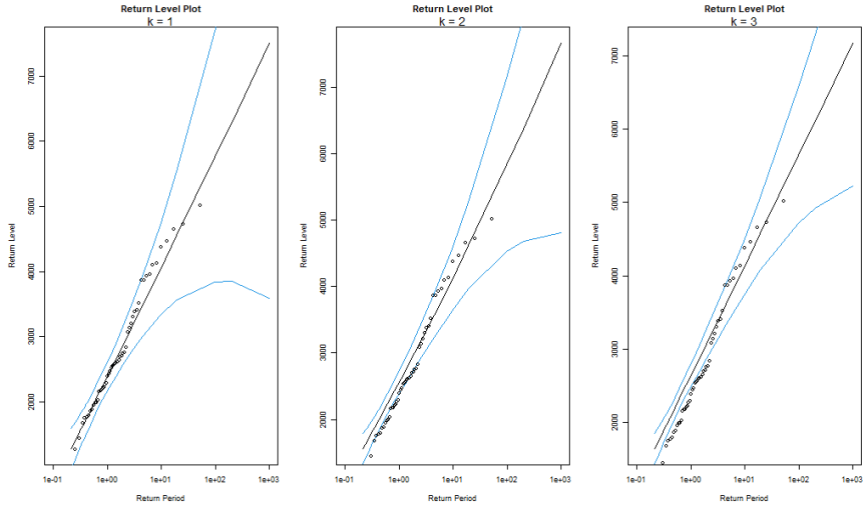


Figure B.59: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{1, 2, 3\}$ largest order statistics of the time series T_{MRD} in the P2 time horizon.

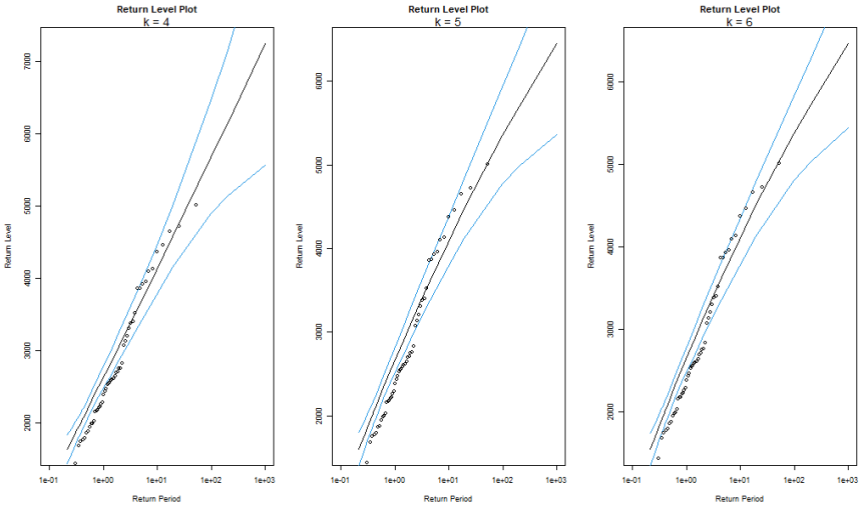


Figure B.60: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{4, 5, 6\}$ largest order statistics of the time series T_{MRD} in the P2 time horizon.

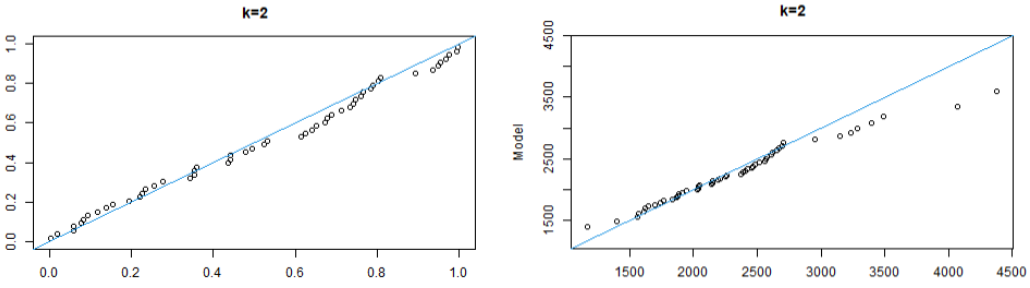


Figure B.61: The model diagnostics plots for the hydrological annual $k = 2$ largest order statistics observations of the time series T_{MRD} in the P1 time horizon. The left column contains the probability plots, and the right column contains the quantile plots.

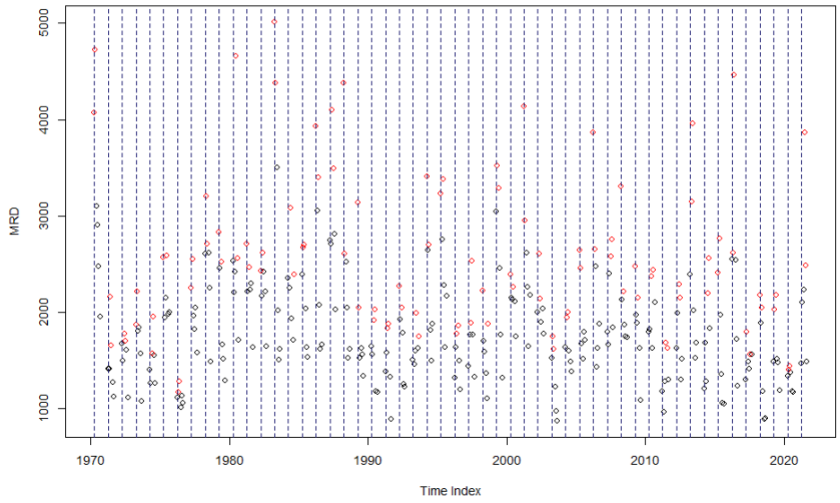


Figure B.62: The MRD hydrological annual top 2 extreme observations are shown in red, while the hydrological annual blocks are shown in blue for the P2 time horizon..

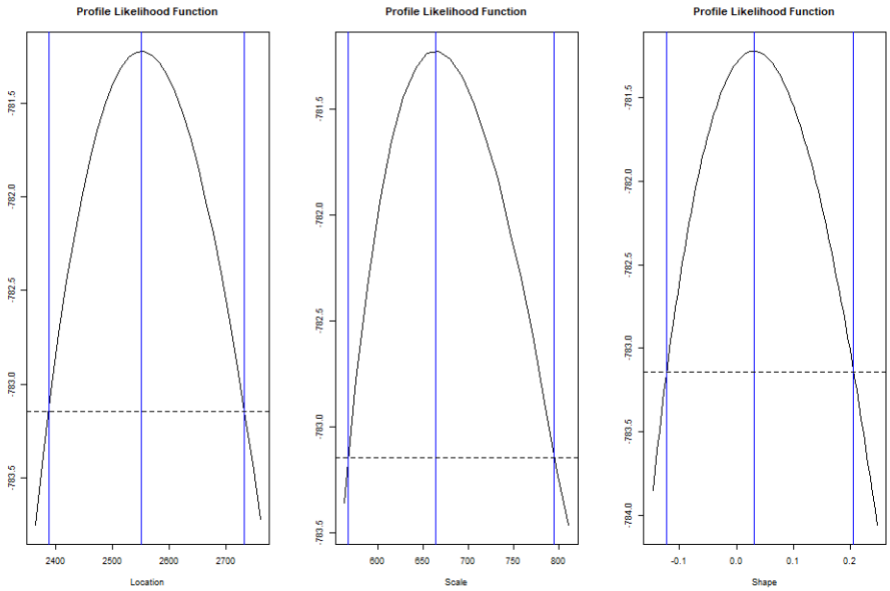


Figure B.63: The KLOS profile log-likelihood plots for each parameter estimate of the GEV_2 distribution fit on the hydrological annual top 2 extreme observations of the time series T_{MRD} in the P2 time horizon.

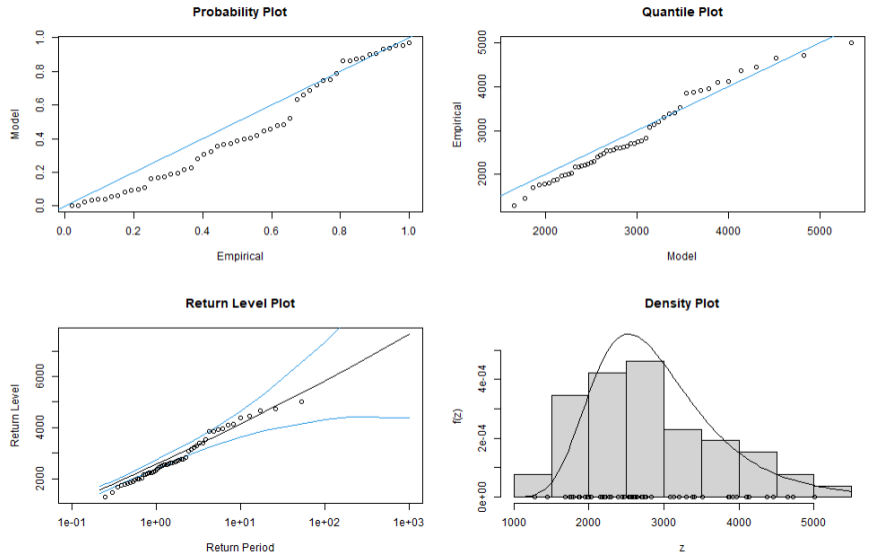


Figure B.64: The diagnostic plots of the GEV_2 distribution fit on the hydrological annual top 2 extreme observations of the time series T_{MRD} in the P2 time horizon.

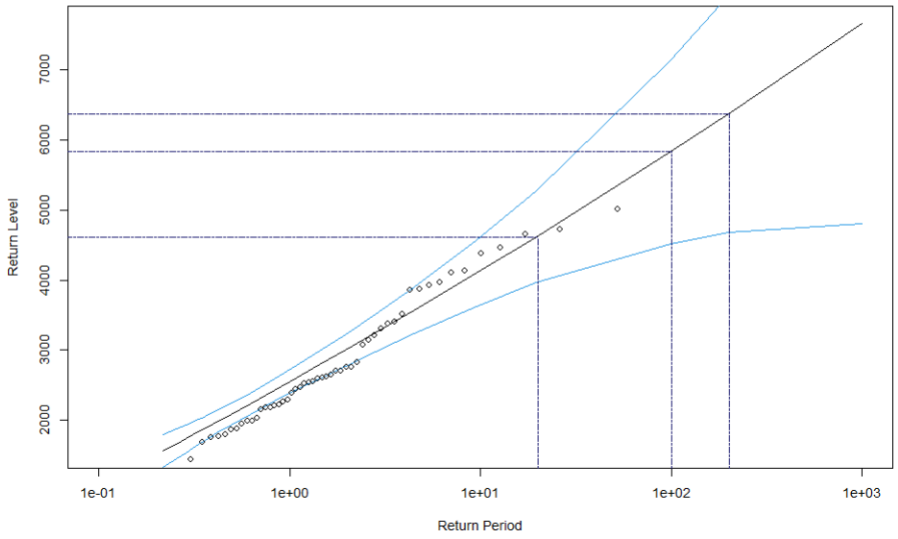


Figure B.65: The KLOS return level curve, with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles for the hydrological annual $k = 2$ largest order statistics of the time series T_{MRD} in the P2 time horizon. The estimate flood scenarios shown in Table B.6 are displayed by the blue dotted lines.

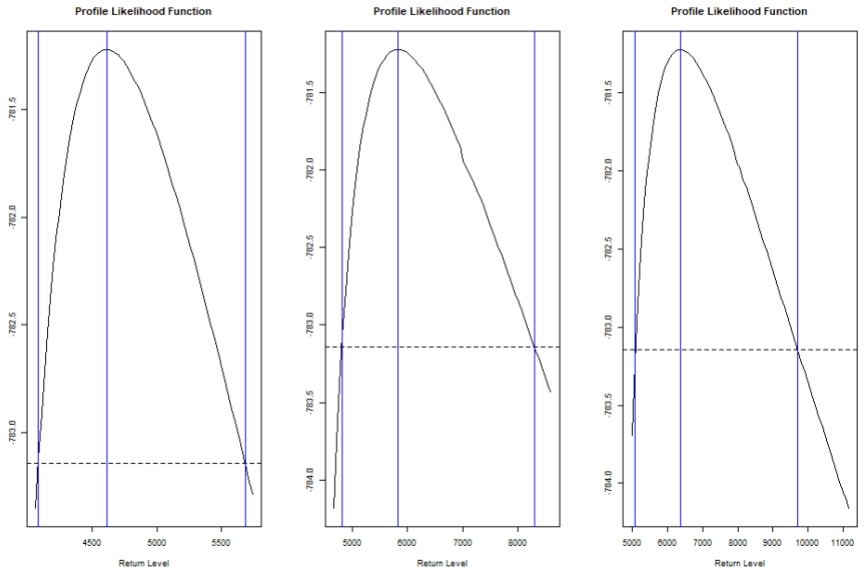


Figure B.66: The KLOS profile log-likelihood plots for the three return periods considered for the top 2 hydrological annual extreme observations of the time series T_{MRD} in the P2 time horizon.

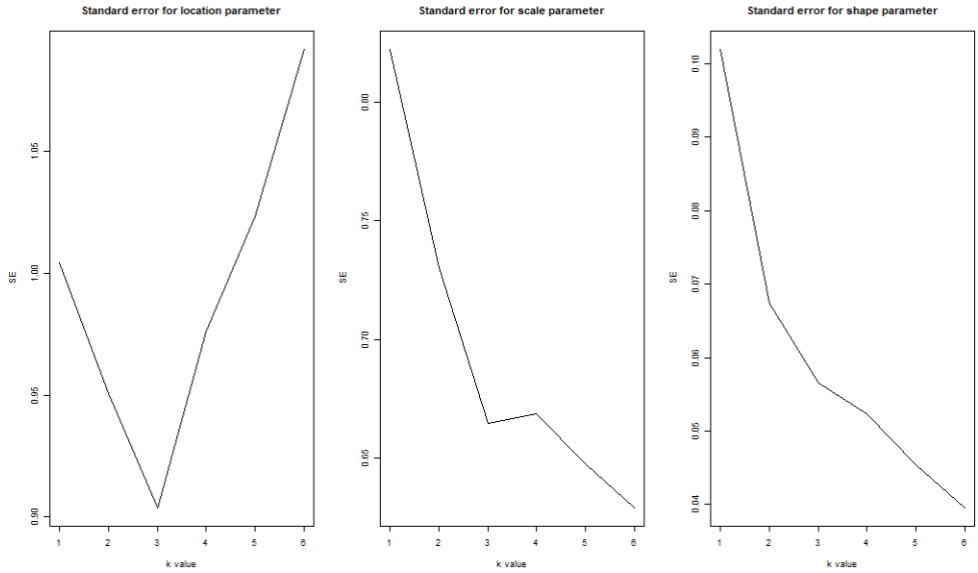


Figure B.67: The MLE standard errors for the GEV_k distribution fit on the hydrological annual top k extreme observations of the time series T_{MRS} in the P1 time horizon.

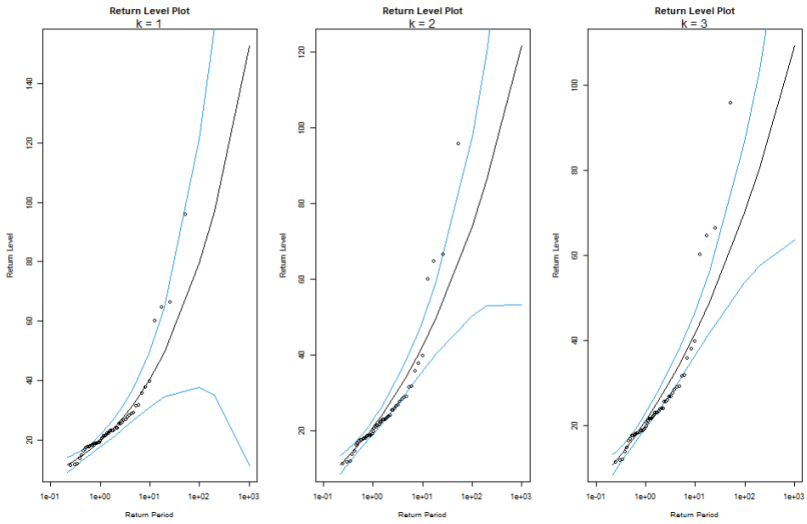


Figure B.68: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{1, 2, 3\}$ largest order statistics of the time series T_{MRS} in the P1 time horizon..

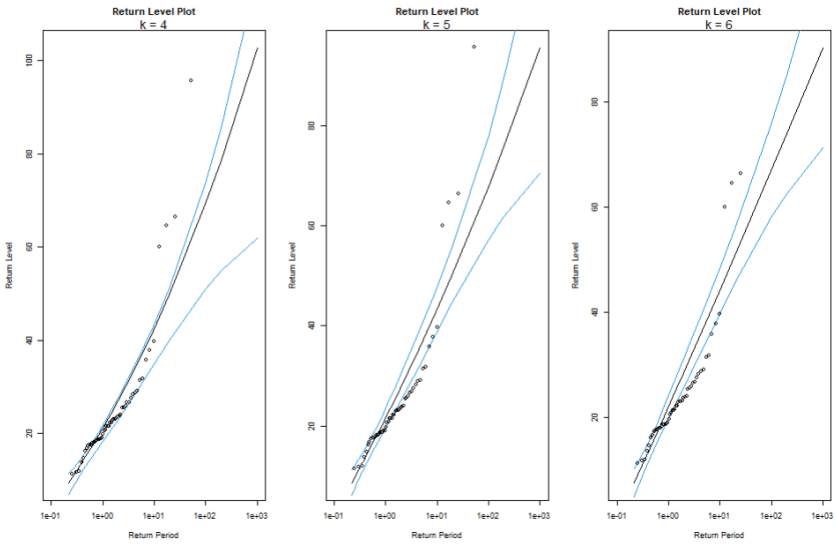


Figure B.69: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{4, 5, 6\}$ largest order statistics of the time series T_{MRS} in the P1 time horizon.

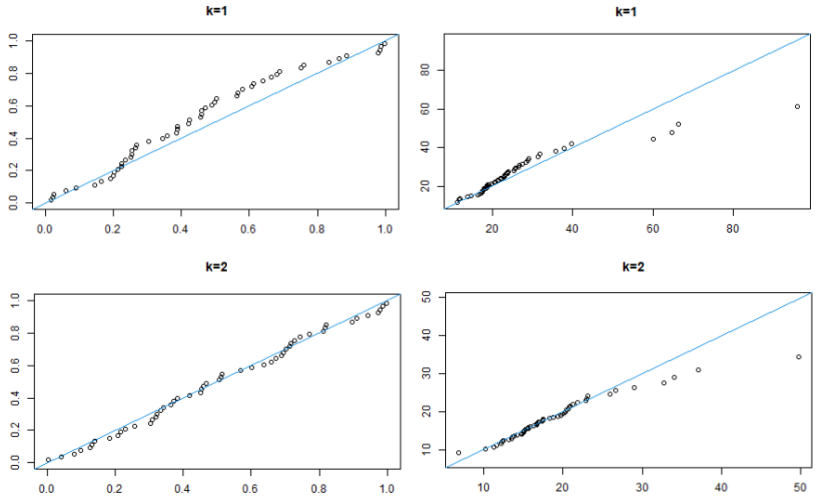


Figure B.70: The model diagnostics plots for the hydrological annual $k \in \{1, 2\}$ largest order statistics observations of the time series T_{MRS} in the P1 time horizon. The left column contains the probability plots, and the right column contains the quantile plots.

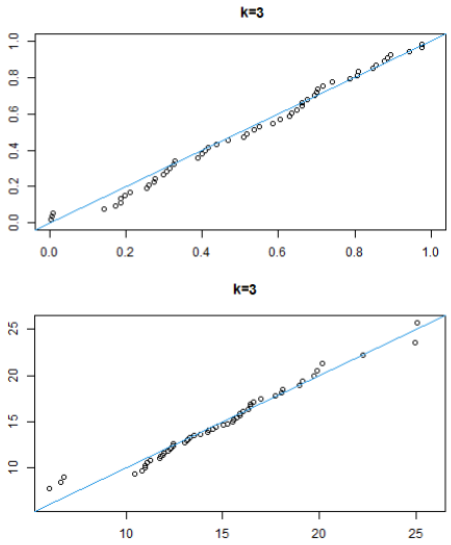


Figure B.71: The model diagnostic plots for the 3^{rd} ordered maxima of the time series T_{MRS} in the P1 time horizon. The top row contains the probability plots, and the bottom row contains the quantile plots.

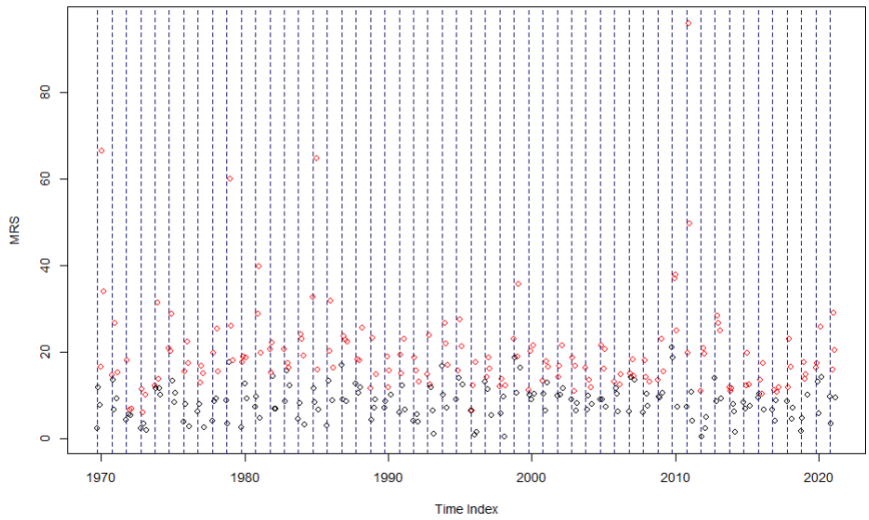


Figure B.72: The MRS hydrological annual top 3 extreme observations are shown in red, while the hydrological annual blocks are shown in blue for the P1 time horizon.

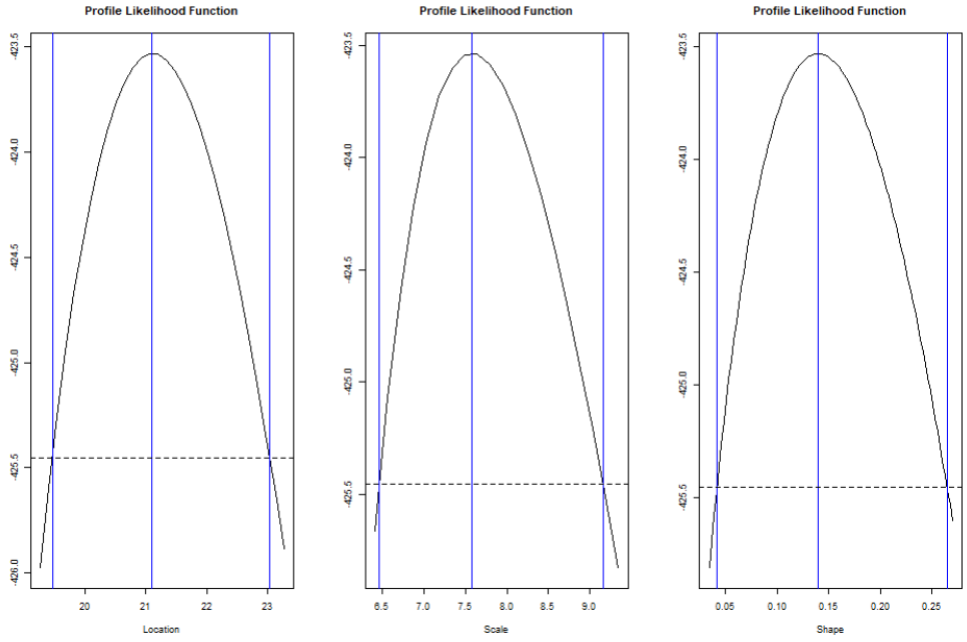


Figure B.73: The KLOS profile log-likelihood plots for each parameter estimate of the GEV_3 distribution fit on the hydrological annual top 3 extreme observations of the time series T_{MRS} in the P1 time horizon.

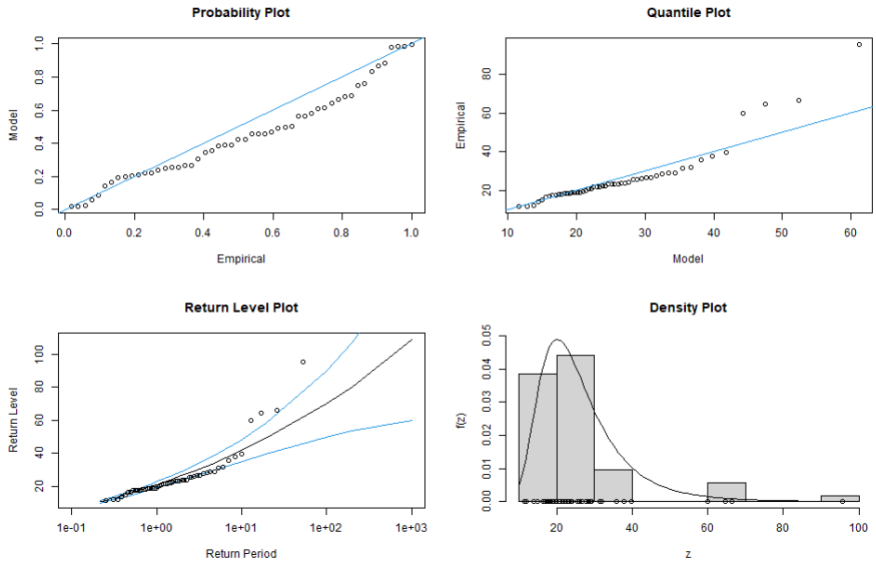


Figure B.74: The diagnostic plots of the GEV_3 distribution fit on the hydrological annual top 3 extreme observations of the time series T_{MRS} in the P1 time horizon.

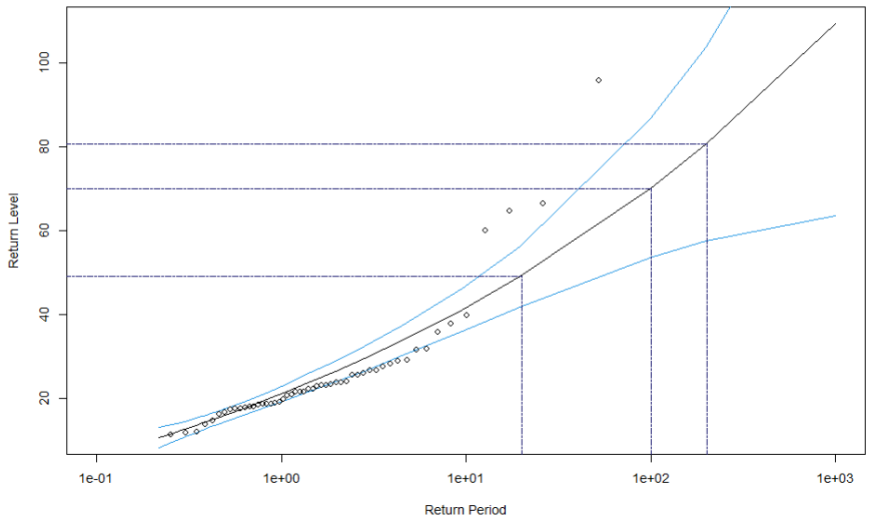


Figure B.75: The KLOS return level curve, with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles for the hydrological annual $k = 3$ largest order statistics of the time series T_{MRS} in the P1 time horizon. The estimate flood scenarios shown in Table B.6 are displayed by the blue dotted lines.

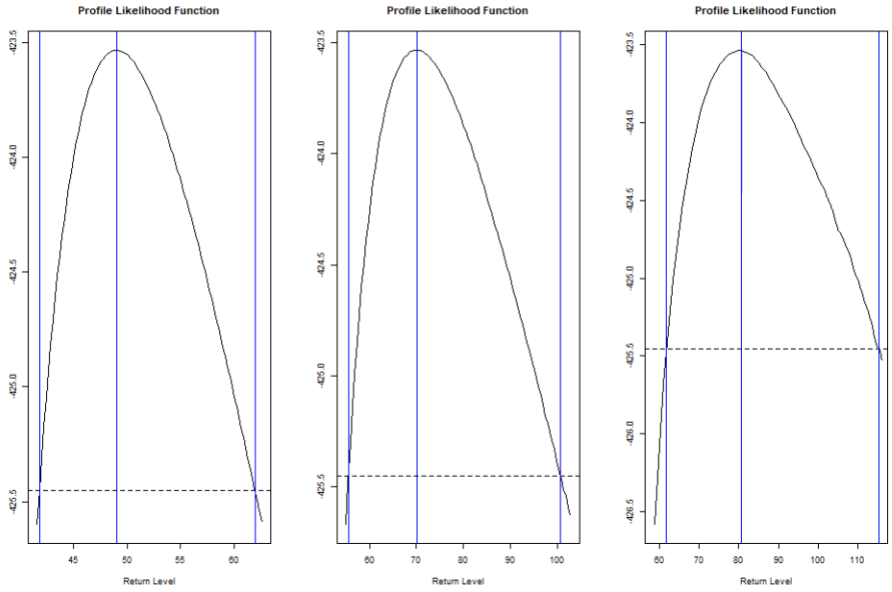


Figure B.76: The KLOS profile log-likelihood plots for the three return periods considered for the hydrological annual top 3 extreme observations of the time series T_{MRS} in the P1 time horizon.

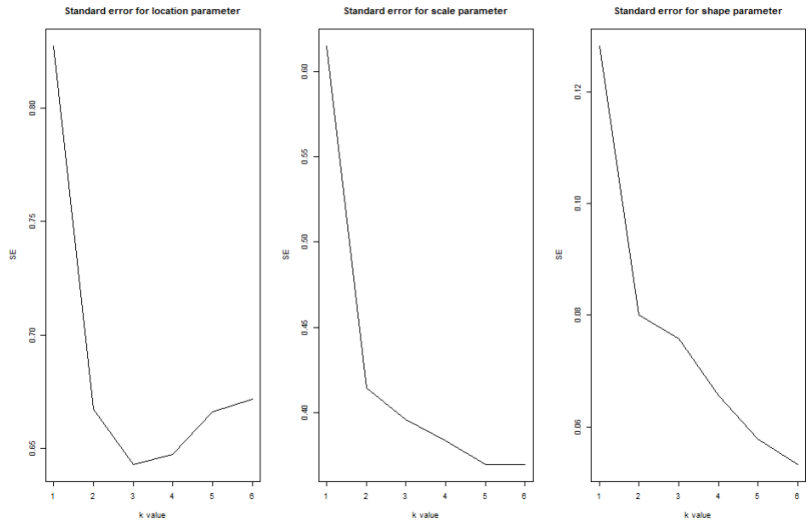


Figure B.77: The MLE standard errors for the GEV_k distribution fit on the hydrological annual top k extreme observations of the time series T_{MRS} in the P2 time horizon.

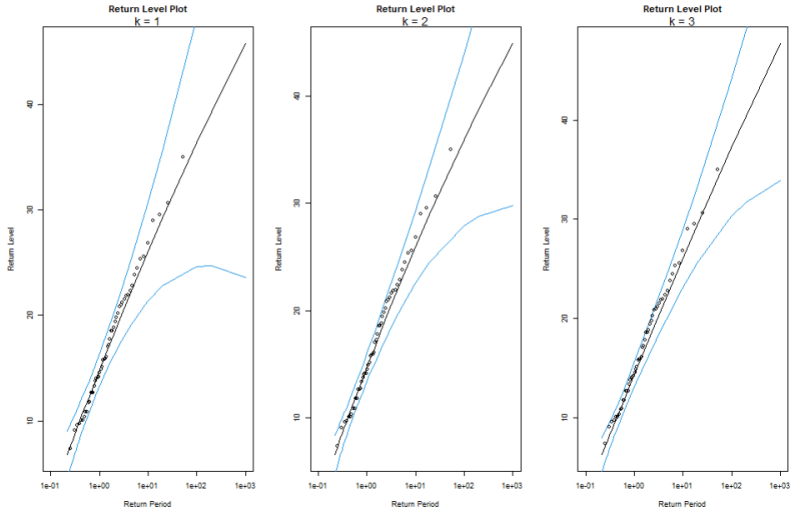


Figure B.78: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{1, 2, 3\}$ largest order statistics of the time series T_{MRS} in the P2 time horizon.

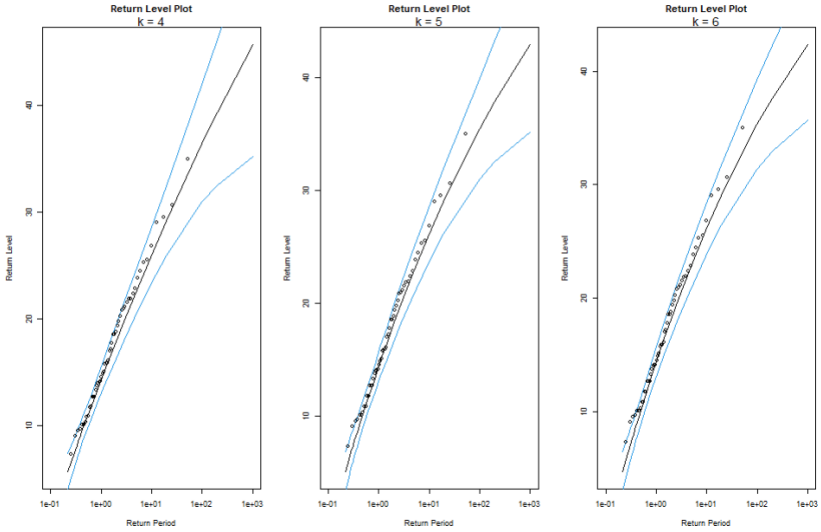


Figure B.79: The KLOS return level curves with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles, are shown for the $k \in \{4, 5, 6\}$ largest order statistics of the time series T_{MRS} in the P2 time horizon.

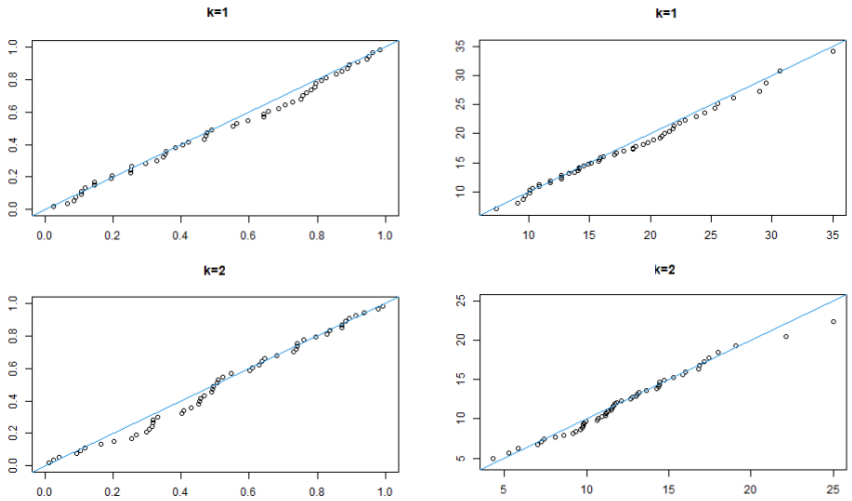


Figure B.80: The model diagnostics plots for the hydrological annual $k \in \{1, 2\}$ largest order statistics observations of the time series T_{MRS} in the P2 time horizon. The left column contains the probability plots, and the right column contains the quantile plots.

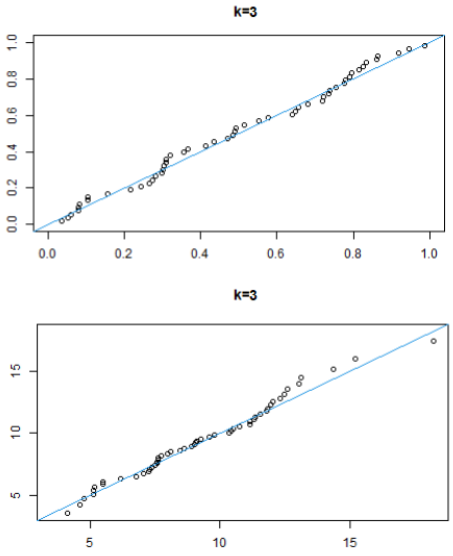


Figure B.81: The model diagnostic plots for the 3^{rd} ordered maxima of the time series T_{MRS} in the P2 time horizon. The top row contains the probability plots, and the bottom row contains the quantile plots.

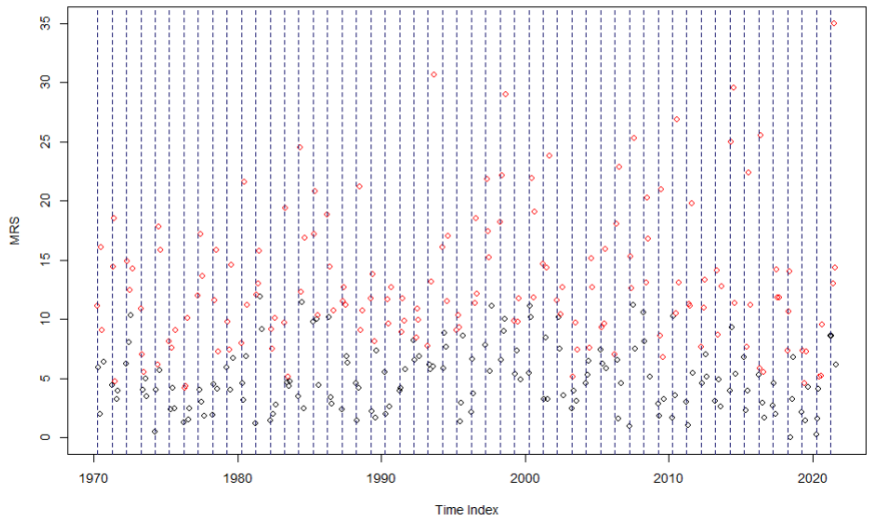


Figure B.82: The MRS hydrological annual top 3 extreme observations are shown in red, while the hydrological annual blocks are shown in blue for the P2 time horizon.

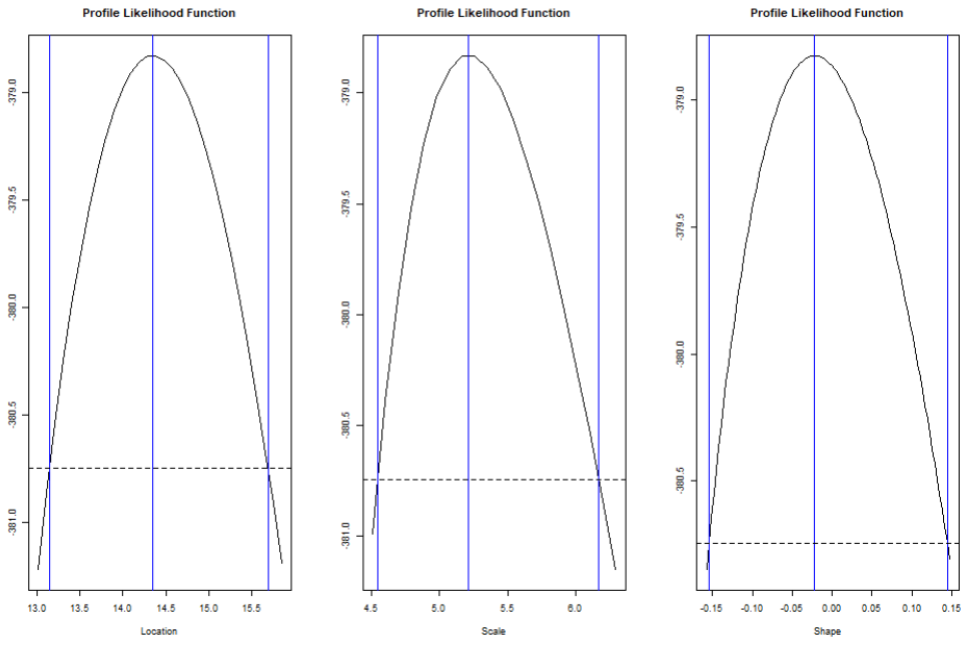


Figure B.83: The KLOS profile log-likelihood plots for each parameter estimate of the GEV_3 distribution fit on the hydrological annual top 3 extreme observations of the time series T_{MRS} in the P2 time horizon.

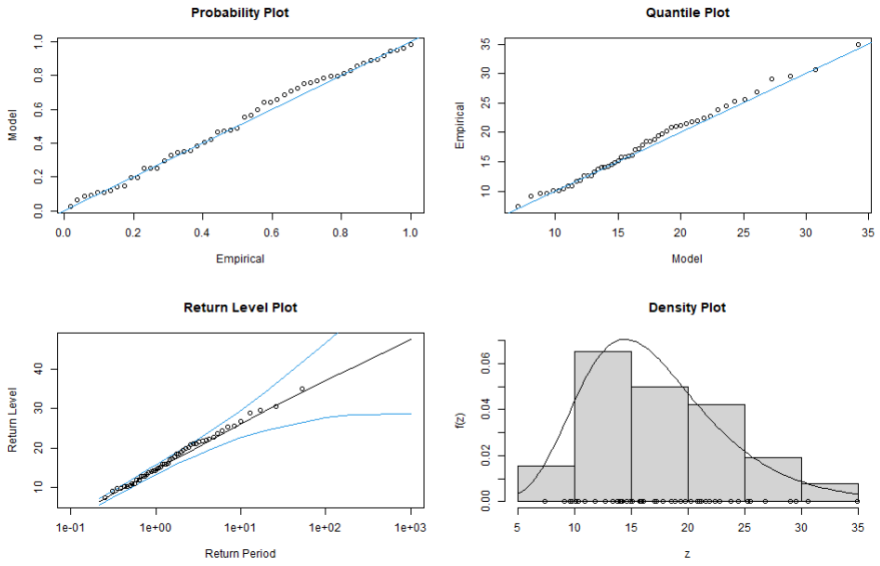


Figure B.84: The diagnostic plots of the GEV_3 distribution fit on the hydrological annual top 3 extreme observations of the time series T_{MRS} in the P2 time horizon.

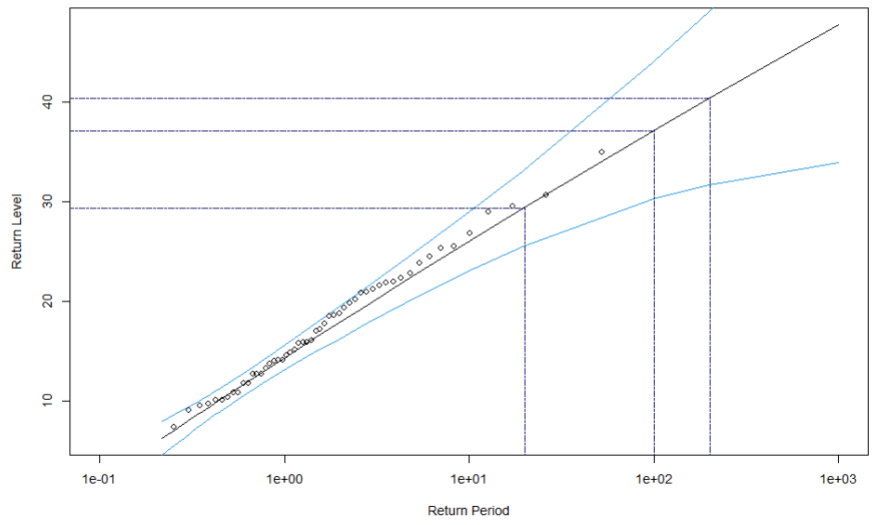


Figure B.85: The KLOS return level curve, with the 95% profile likelihood confidence intervals represented by blue lines and the empirical return levels denoted by black circles for the hydrological annual $k = 3$ largest order statistics of the time series T_{MRS} in the P2 time horizon. The estimate flood scenarios shown in Table B.6 are displayed by the blue dotted lines.

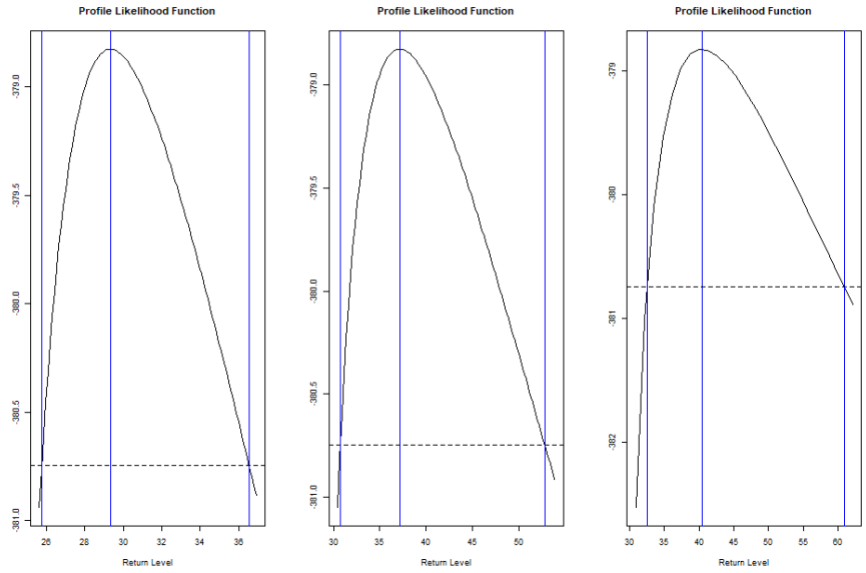


Figure B.86: The KLOS profile log-likelihood plots for the three return periods considered for the hydrological annual top 3 extreme observations of the time series T_{MRS} in the P2 time horizon.

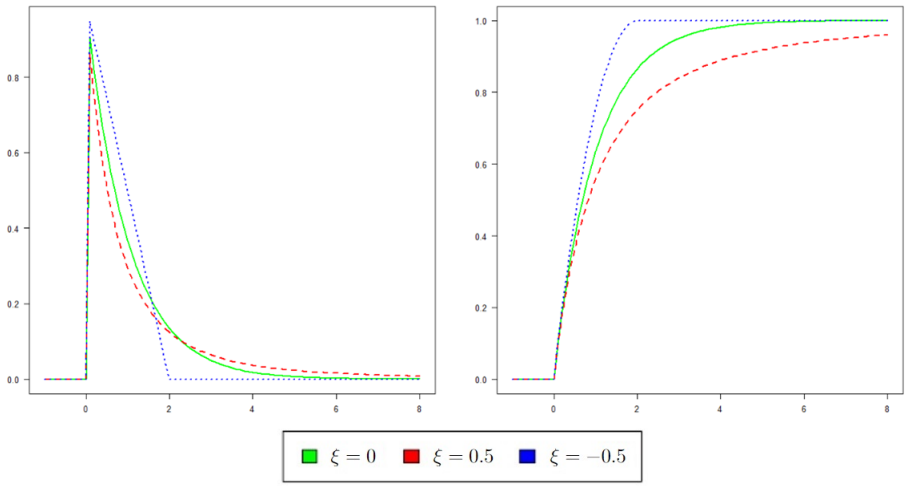


Figure B.87: The family of the GP probability density functions and cumulative distribution functions.

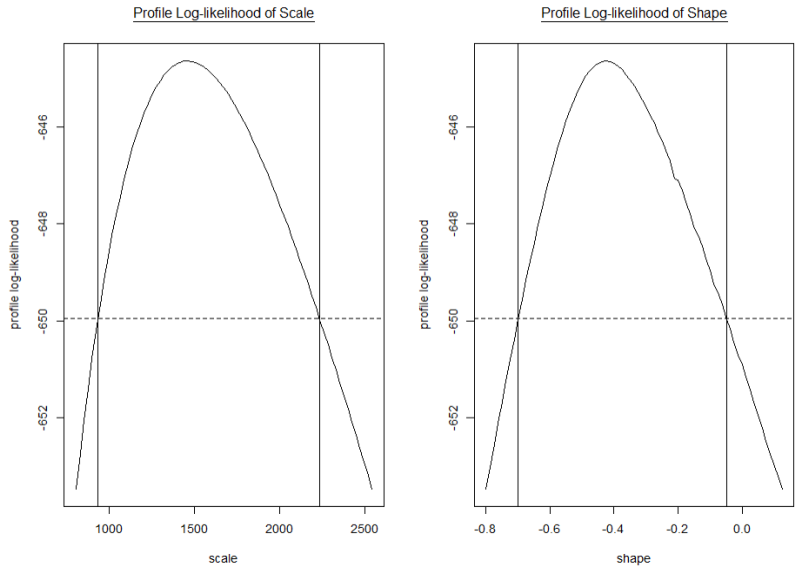


Figure B.88: The POT profile log-likelihood plots for each parameter estimate of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRD} in the P1 time horizon.

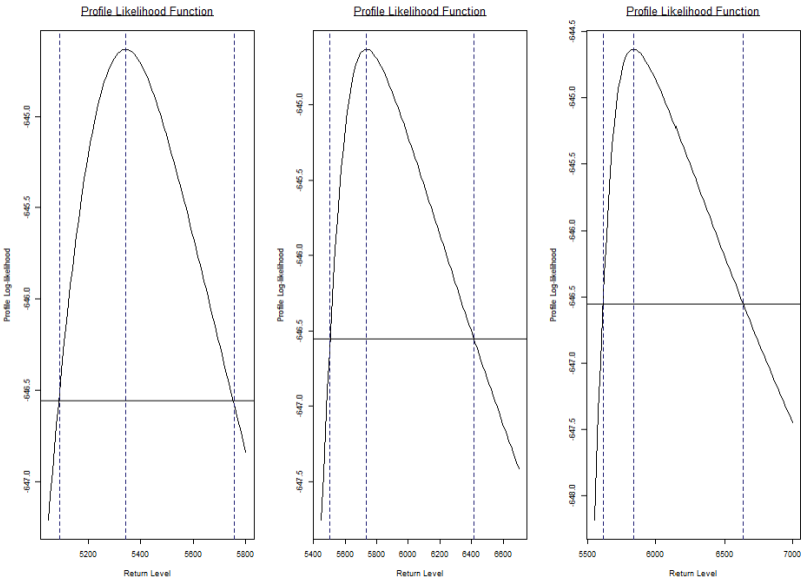


Figure B.89: The POT profile log-likelihood plots for the three return periods considered for the threshold excess of the extreme observations of the time series T_{MRD} in the P1 time horizon.

		Parameters	95% lower CI	Estimate	95% upper CI
T_{MRD}	P1	ξ	-0.7	-0.43	-0.05
		σ	936.60	1460.98	2233.56
	P2	ξ	-0.22	0.16	0.66
		σ	246.06	522.86	717.84
T_{MRS}	P1	ξ	0.11	0.49	1.12
		σ	3.21	5.73	9.80
	P2	ξ	-0.36	-0.18	0.11
		σ	4.80	6.73	9.16

Table B.7: Summary of the GP parameter estimates and their corresponding confidence intervals, evaluated on threshold excess of the extreme observations of the time series T_{MRD} and T_{MRS} for both time horizons.

		Return Period	95% lower CI	Return Levels	95% upper CI
T_{MRD}	P1	1 in 20 year event	5087.88	5342.79	5751.97
		1 in 100 year event	5500.51	5734.75	6409.60
		1 in 200 year event	5608.59	5836.55	6633.84
	P2	1 in 20 year event	4106.06	4720.78	6969.70
		1 in 100 year event	4441.41	6342.96	7517.17
		1 in 200 year event	4372.73	7183.29	11136.36
T_{MRS}	P1	1 in 20 year event	41.61	54.94	107.88
		1 in 100 year event	60.00	108.50	540.00
		1 in 200 year event	72.53	148.20	1137.17
	P2	1 in 20 year event	26.21	28.95	34.90
		1 in 100 year event	30.40	34.18	48.12
		1 in 200 year event	35.10	41.45	47.79

Table B.8: Summary of the POT return levels and their corresponding confidence intervals for the three return periods considered for threshold excess of the extreme observations of the time series T_{MRD} and T_{MRS} for both time horizons.

		Threshold	AIC	BIC
T_{MRD}	P1	2700	1293.27	1293.67
	P2	2510	968.81	968.98
T_{MRS}	P1	22	262.90	262.59
	P2	12	402.45	402.74

Table B.9: Summary of the adequacy of the POT model fit for each time series T_{MRD} and T_{MRS} for both time horizons.

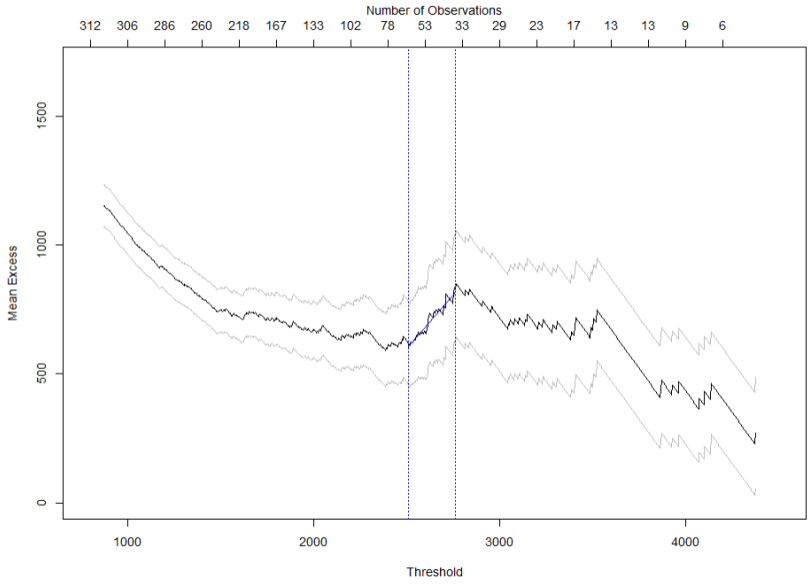


Figure B.90: The empirical MRL plot for the threshold excess of the time series T_{MRD} in the P2 time horizon featuring a solid blue line, indicating the plot's linearity after the threshold $u = 2510$, and two vertical dotted blue lines, which mark the corresponding threshold values and simultaneously the remaining number of observations.

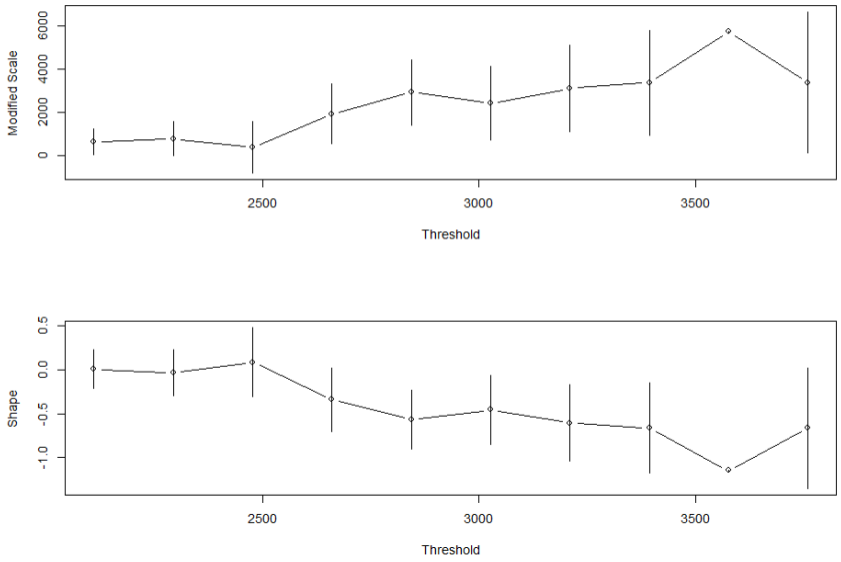


Figure B.91: The parameter stability plots for the threshold excess of the extreme observations of the time series T_{MRD} in the P2 time horizon. The black circles are the parameter estimates and the vertical lines represent the corresponding 95% confidence intervals.

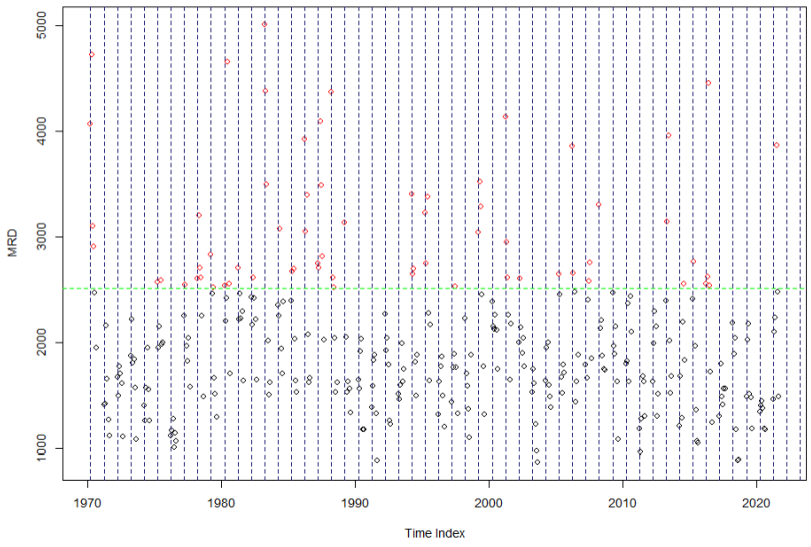


Figure B.92: The MRD extreme observations that are over the threshold $u = 2510$ (green line) are marked with a red circle and the hydrological annual blocks with blue dotted lines for the P2 time horizon

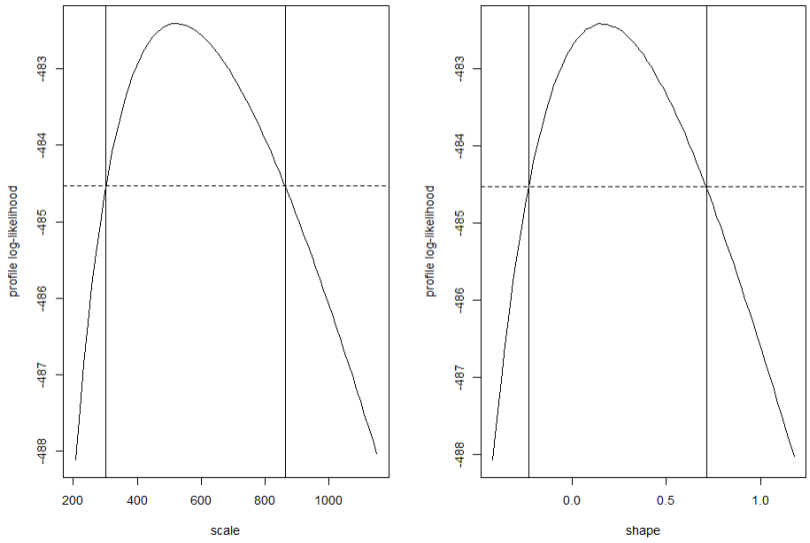


Figure B.93: The POT profile log-likelihood plots for each parameter estimate of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRD} in the P2 time horizon.

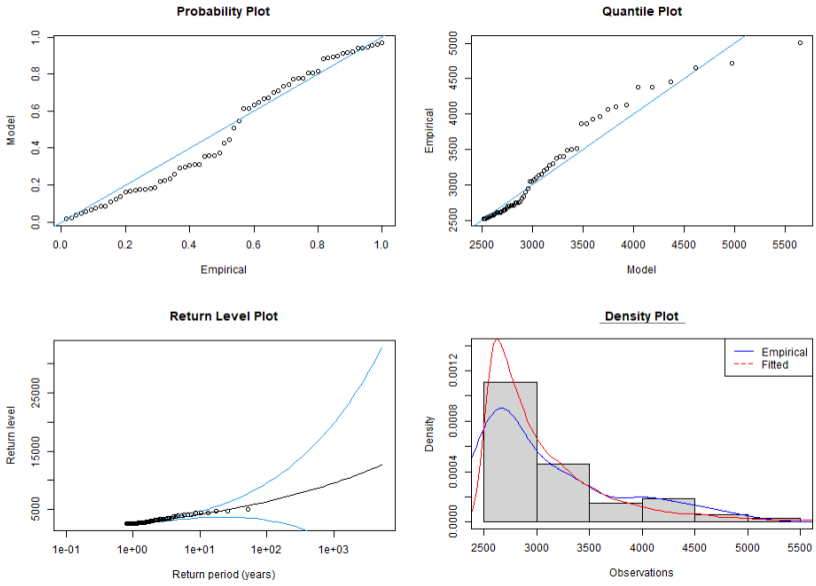


Figure B.94: The diagnostic plots of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRS} in the P2 time horizon.

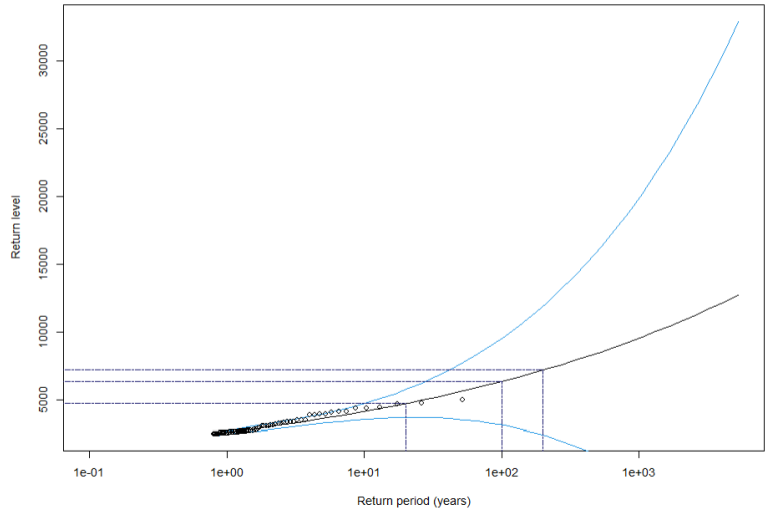


Figure B.95: The POT return level curve, with the 95% profile likelihood confidence intervals (light blue dotted lines) and the empirical return levels (black circles) for the time series T_{MRD} in the P2 time horizon. The estimate flood scenarios shown in Table B.8 are displayed by the blue dotted lines.

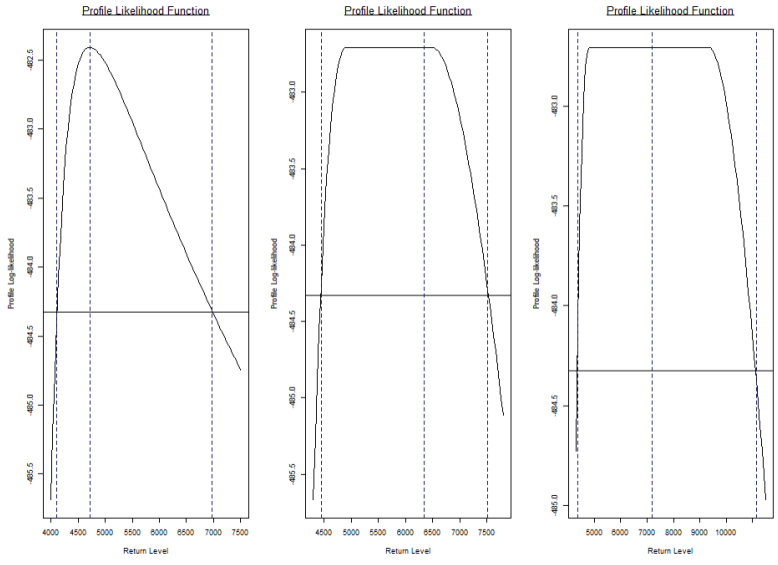


Figure B.96: The POT profile log-likelihood plots for the three return periods considered for the threshold excess of the extreme observations of the time series T_{MRD} in the P2 time horizon.

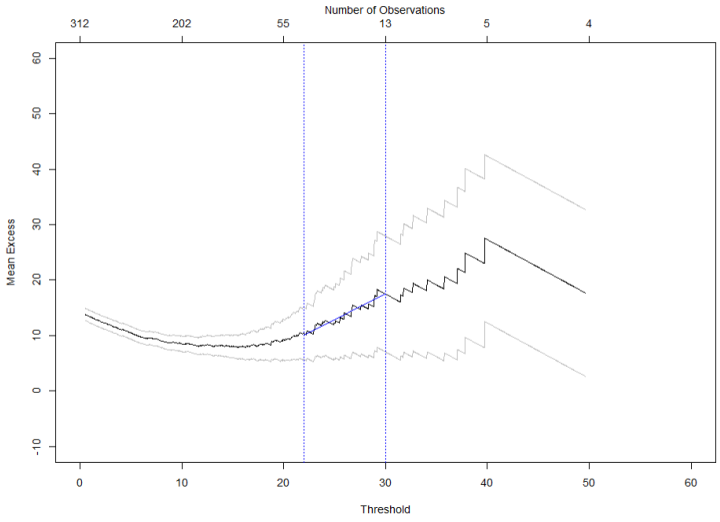


Figure B.97: The empirical MRL plot for the threshold excess of the extreme observations of the time series T_{MRS} in the P1 time horizon featuring a solid blue line, indicating the plot's linearity after the threshold $u = 22$, and two vertical dotted blue lines, which mark the corresponding threshold values and simultaneously the remaining number of observations.

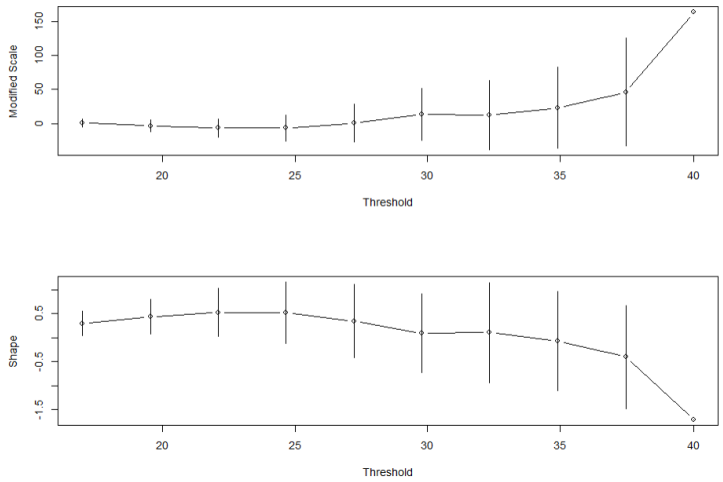


Figure B.98: The parameter stability plot for the threshold excess of the extreme observations of the time series T_{MRS} in the P1 time horizon. The black circles are the parameters estimate and the vertical lines represent the corresponding confidence interval.

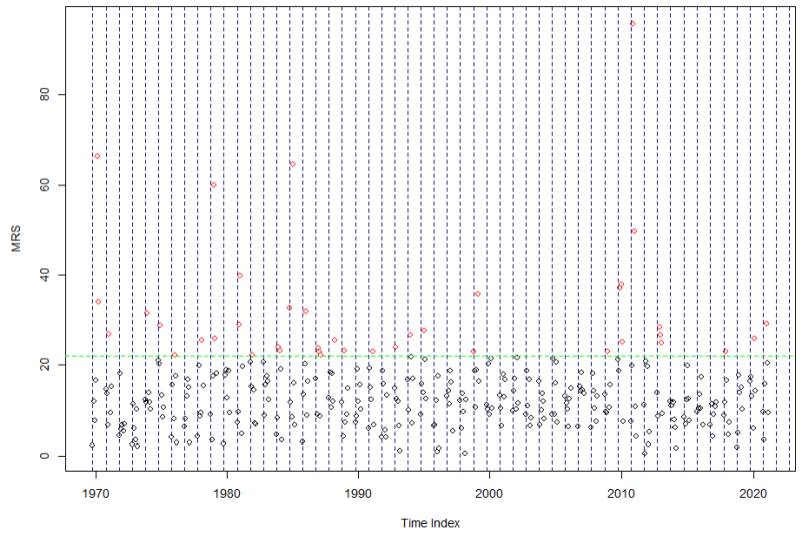


Figure B.99: The MRS extreme observations that are over the threshold $u = 22$ (green line) are marked with a red circle and the hydrological annual blocks with blue dotted lines for the P1 time horizon.

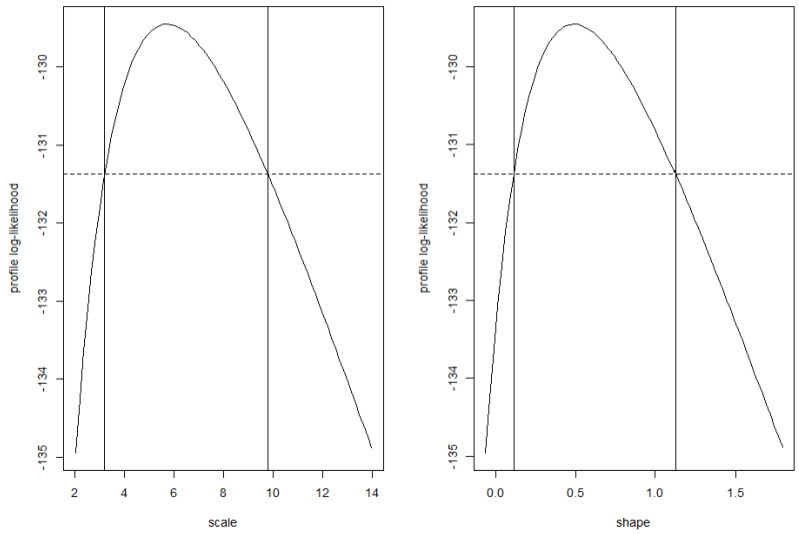


Figure B.100: The POT profile log-likelihood plots for each parameter estimate of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRS} in the P1 time horizon.

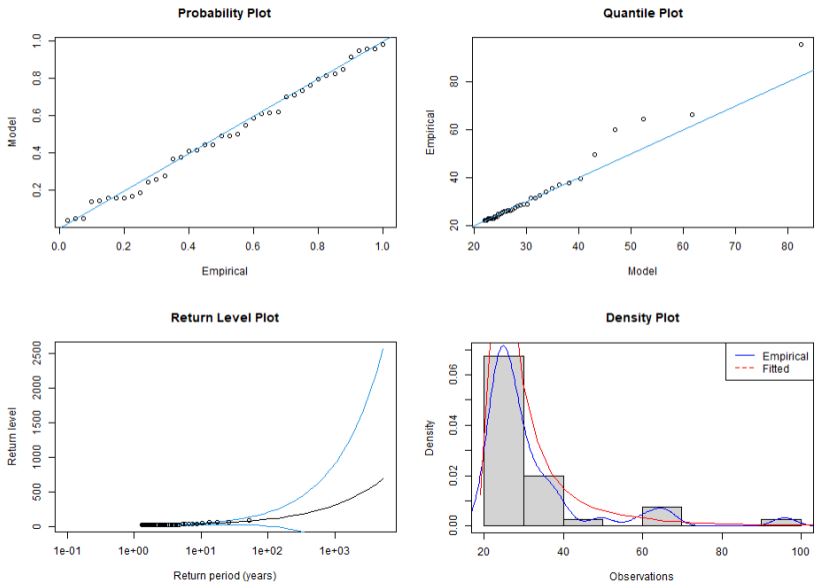


Figure B.101: The diagnostic plots of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRS} in the P1 time horizon.

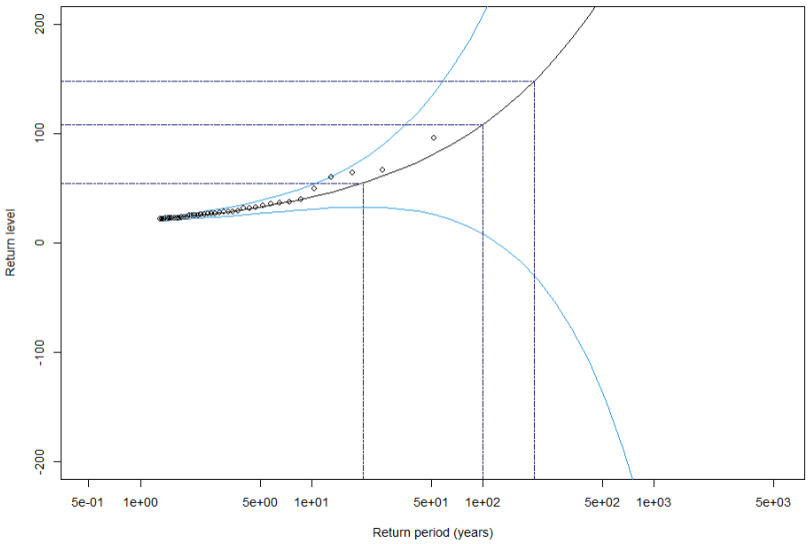


Figure B.102: The POT return level curve, with the 95% profile likelihood confidence intervals (light blue dotted lines) and the empirical return levels (black circles) for the time series T_{MRS} in the P1 time horizon. The estimate flood scenarios shown in Table B.8 are displayed by the blue dotted lines.

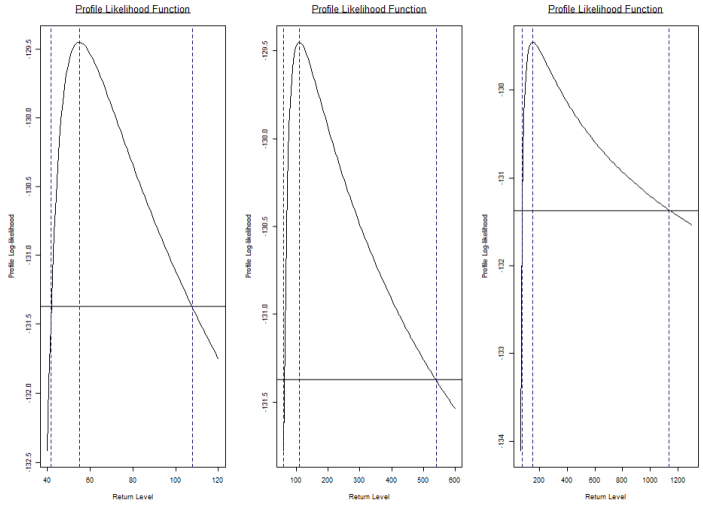


Figure B.103: The POT profile log-likelihood plots for the three return periods considered for the threshold excess of the extreme observations of the time series T_{MRS} in the P1 time horizon.

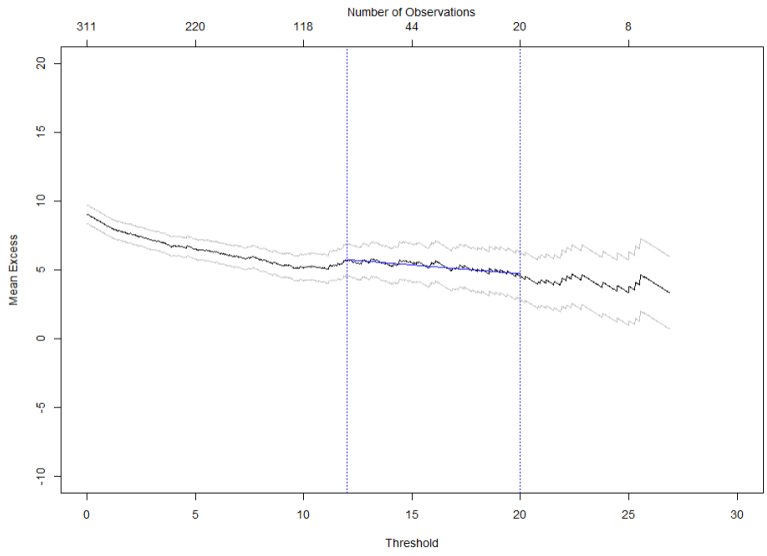


Figure B.104: The empirical MRL plot for the threshold excess of the extreme observations of the time series T_{MRS} in the P2 time horizon featuring a solid blue line, indicating the plot's linearity after the threshold $u = 12$, and two vertical dotted blue lines, which mark the corresponding threshold values and simultaneously the remaining number of observations.

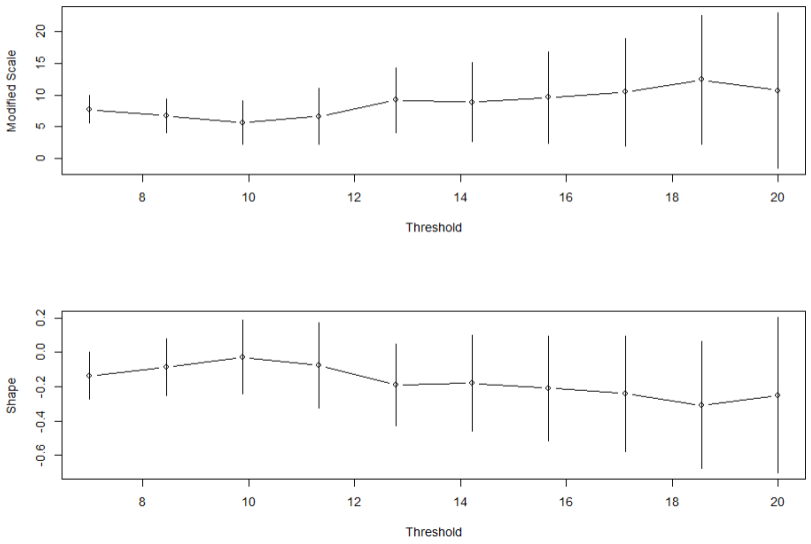


Figure B.105: The parameter stability plots for the threshold excesses of the extreme observations of the time series T_{MRS} in the P2 time horizon. The black circles are the parameter estimates and the vertical lines represent the corresponding 95% confidence intervals.

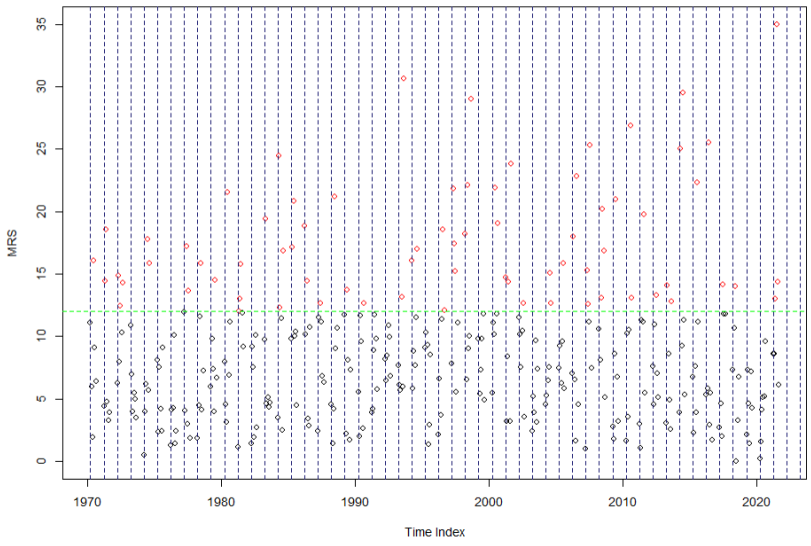


Figure B.106: The MRS extreme observations that are over the threshold $u = 12$ (green line) are marked with a red circle and the hydrological annual blocks with blue dotted lines for the P2 time horizon.

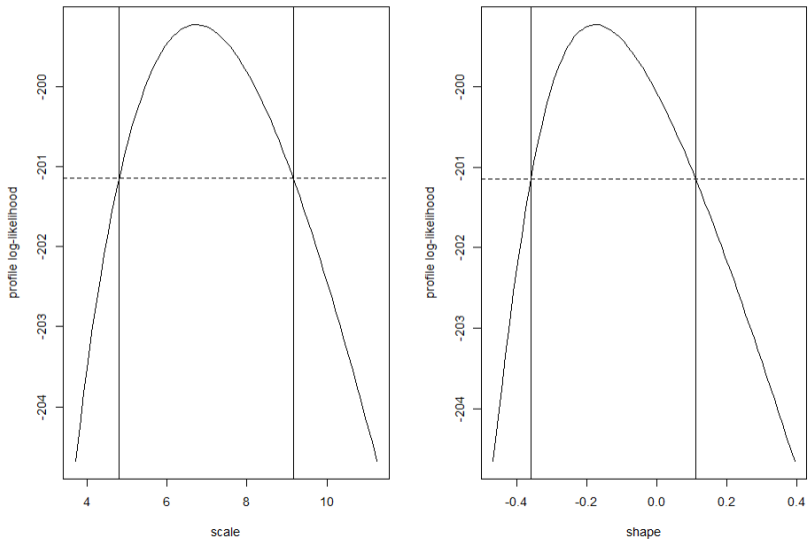


Figure B.107: The POT profile log-likelihood plots for each parameter estimate of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRS} in the P2 time horizon.

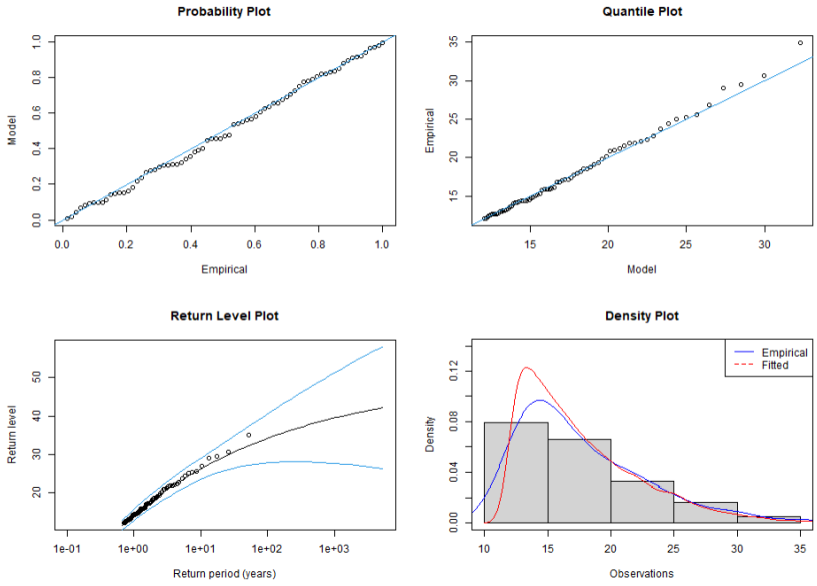


Figure B.108: The diagnostic plots of the GP distribution fit on the threshold excess of the extreme observations of the time series T_{MRS} in the P2 time horizon.

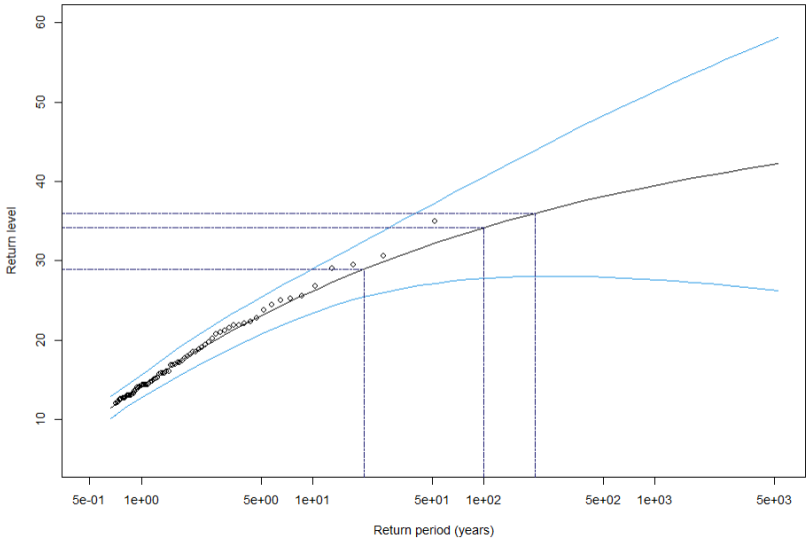


Figure B.109: The POT return level curve, with the 95% profile likelihood confidence intervals (light blue dotted lines) and the empirical return levels (black circles) for the time series T_{MRS} in the P2 time horizon. The estimate flood scenarios shown in Table B.8 are displayed by the blue dotted lines.

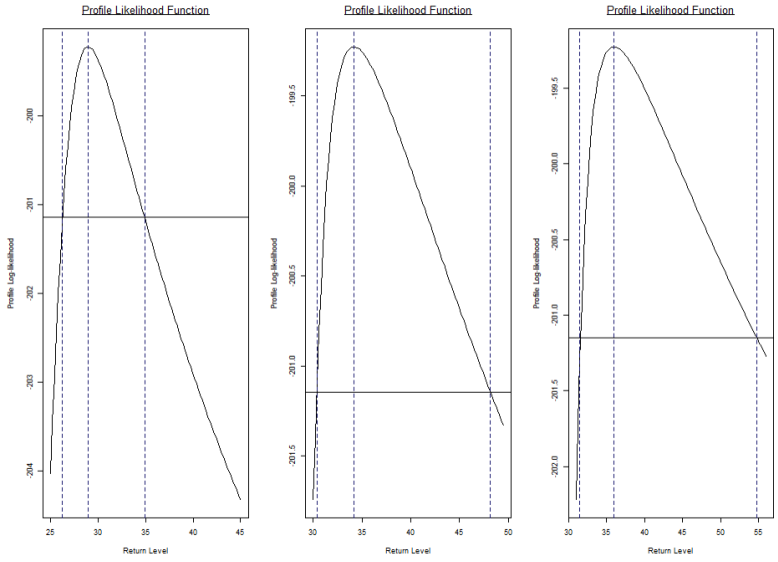


Figure B.110: The POT profile log-likelihood plots for the three return periods considered for the threshold excess of the extreme observations of the time series T_{MRS} in the P2 time horizon.

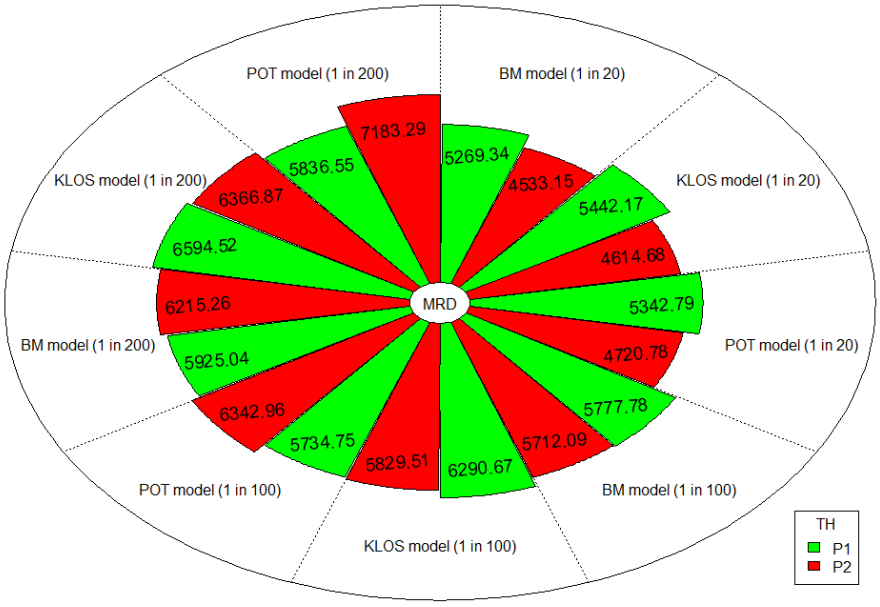


Figure B.111: Comparison of the flood scenarios generated by the three univariate models using the time series T_{MRD} in both time horizons

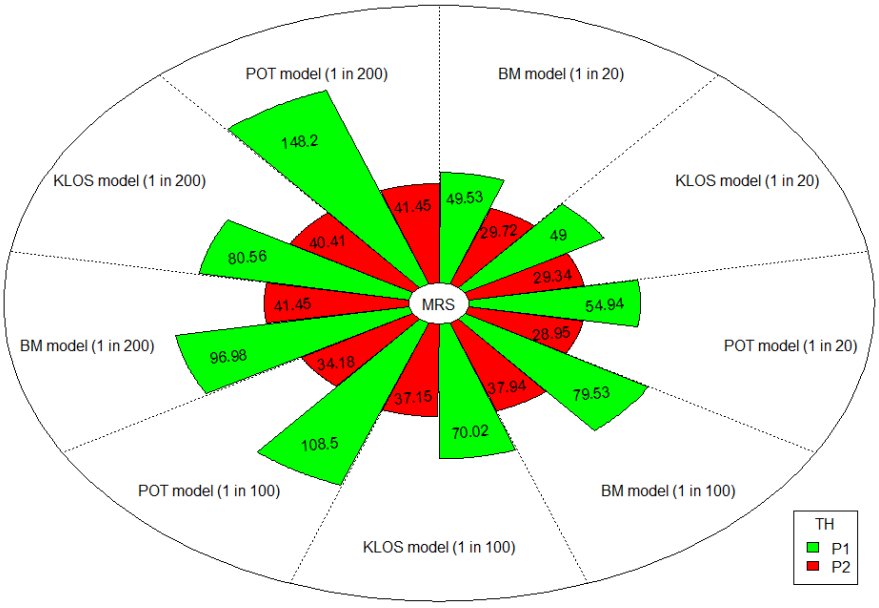


Figure B.112: Comparison of the flood scenarios generated by the three univariate models using the time series T_{MRS} in both time horizons.

B.3 Additional Material for Chapter 4

		AIC	BIC	95% lower CI	Parameter Estimate	95% lower CI
Gumbel Copula	P1	−3.24	−1.29	1.16	1.19	1.23
	P2	−1.49	0.47	1.15	1.19	1.23
Clayton Copula	P1	1.85	3.80	0.65	0.69	0.73
	P2	−4.87	−2.92	0.46	0.48	0.53
Frank Copula	P1	−4.95	−3.00	2.08	2.20	2.37
	P2	−3.04	−1.09	1.60	1.75	1.87

Table B.10: Summary of the adequacy of the CWBM and GCB model fit for both time horizons and the parameter estimates with the corresponding confidence intervals.

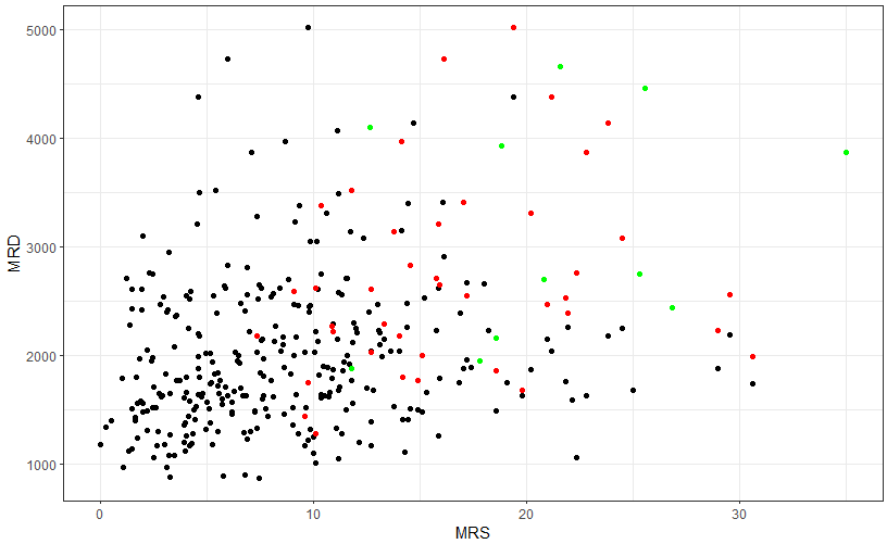


Figure B.113: The scatter plot shows the relation between the extreme values of the time series T_{MRD} and T_{MRS} in the P2 time horizon. The red and green observations are the component-wise maxima, but the green observations correspond to actual observations.

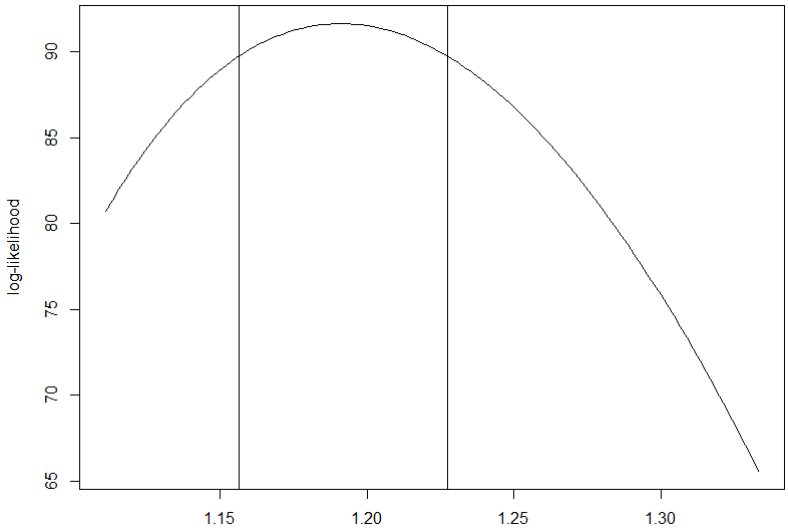


Figure B.114: The CWBM profile log-likelihood function for the Gumbel copula with parameter $\hat{\alpha} = 1.19$ for the extremes of T_{MRD} and T_{MRS} in the P1 time horizon.

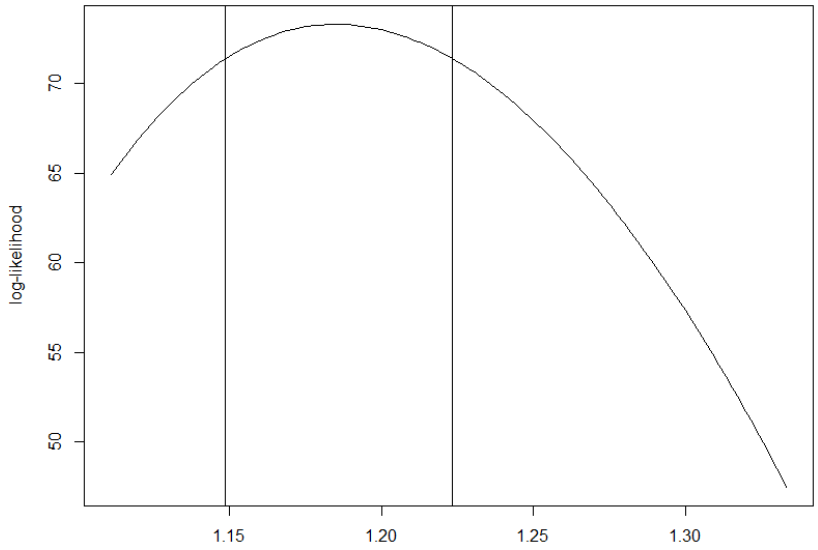


Figure B.115: The CWBM profile log-likelihood function for the Gumbel copula with parameter $\hat{\alpha} = 1.19$ for the extremes of T_{MRD} and T_{MRS} in the P2 time horizon.

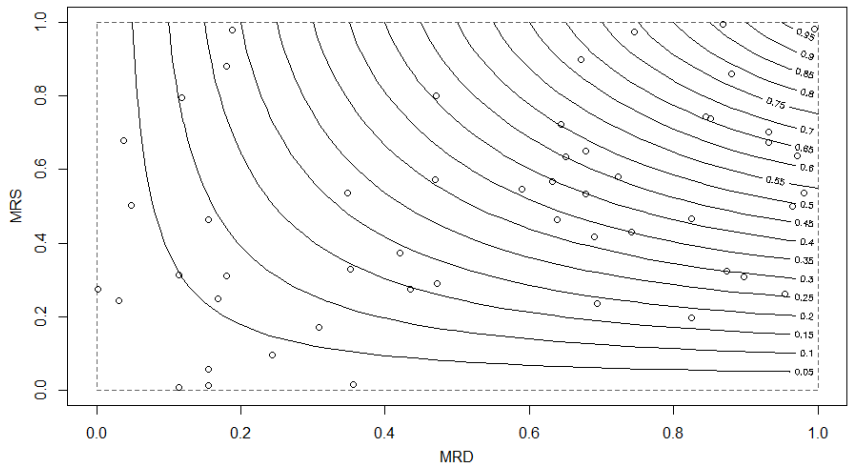


Figure B.116: The critical layers of the Gumbel copula under the OR hazard scenario for the time series T_{MRD} and T_{MRS} in P1 time horizon. The black circles represents the probability transformed component-wise maxima.

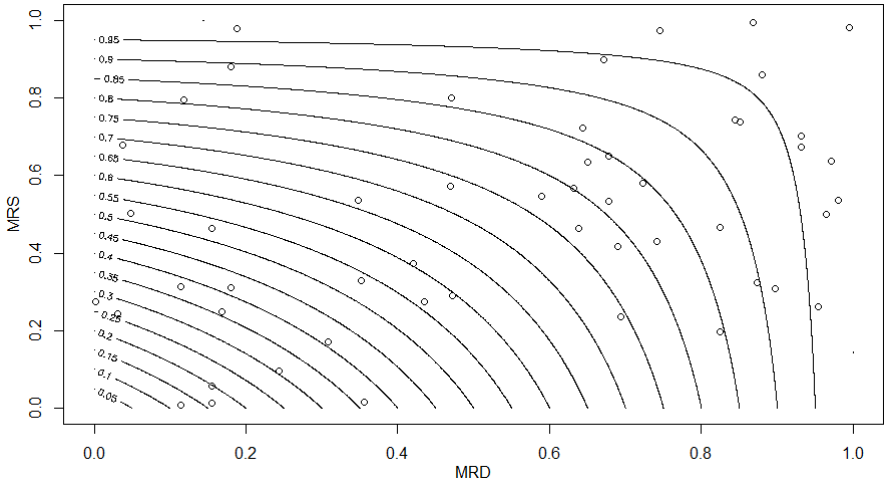


Figure B.117: The critical layers of the Gumbel copula under the AND hazard scenario for the time series T_{MRD} and T_{MRS} in P1 time horizon. The black circles represents the probability transformed component-wise maxima.

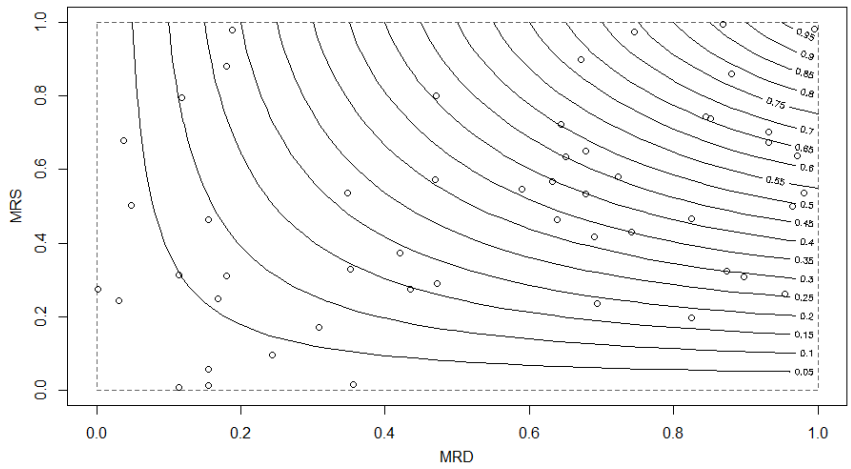


Figure B.118: The critical layers of the Gumbel copula under the OR hazard scenario for the time series T_{MRD} and T_{MRS} in P2 time horizon. The black circles represents the probability transformed component-wise maxima.

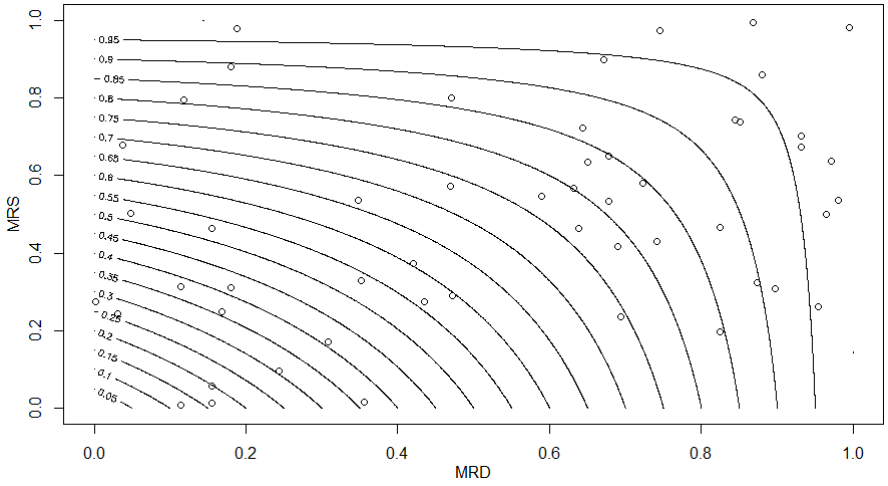


Figure B.119: The critical layers of the Gumbel copula under the AND hazard scenario for the time series T_{MRD} and T_{MRS} in P2 time horizon. The black circles represents the probability transformed component-wise maxima.

			Return Periods	Quantile	Lower CI	Estimated	Upper CI
P1	OR ($\vee$)	T_{MRD}	1 in 20 year event	0.974	5460.66	5682.65	5691.26
			1 in 100 year event	0.995	5899.52	6054.54	6056.58
			1 in 200 year event	0.997	6156.53	6186.18	6292.36
		T_{MRS}	1 in 20 year event	0.974	65.89	75.67	77.98
			1 in 100 year event	0.995	98.82	188.72	211.77
			1 in 200 year event	0.997	120.71	270.56	276.06
	AND ($\wedge$)	T_{MRD}	1 in 20 year event	0.785	3846.27	4048.76	4279.38
			1 in 100 year event	0.905	4384.83	4537.42	5028.75
			1 in 200 year event	0.9733	4588.06	4721.31	5356.89
		T_{MRS}	1 in 20 year event	0.785	19.53	23.74	24.71
			1 in 100 year event	0.905	20.61	24.52	36.16
			1 in 200 year event	0.933	20.89	24.74	47.70
P2	OR ($\vee$)	T_{MRD}	1 in 20 year event	0.975	5159.11	5395.05	5548.73
			1 in 100 year event	0.994	6420.88	6634.02	6825.09
			1 in 200 year event	0.997	6779.46	6934.77	7449.95
		T_{MRS}	1 in 20 year event	0.975	33.20	34.49	35.43
			1 in 100 year event	0.994	39.44	40.55	42.98
			1 in 200 year event	0.997	43.13	44.64	45.97
	AND ($\wedge$)	T_{MRD}	1 in 20 year event	0.844	2684.40	2990.54	3323.83
			1 in 100 year event	0.916	3283.89	3728.82	4362.08
			1 in 200 year event	0.941	3160.88	3737.95	4678.14
		T_{MRS}	1 in 20 year event	0.844	17.32	19.45	21.55
			1 in 100 year event	0.916	20.12	20.99	28.33
			1 in 200 year event	0.941	21.25	22.05	30.23

Table B.11: The Most-Likely design realizations for each of the three return period scenarios, according to the OR and AND hazard scenarios that are based on the Gumbel copula for both time horizon. The corresponding confidence intervals are also included.

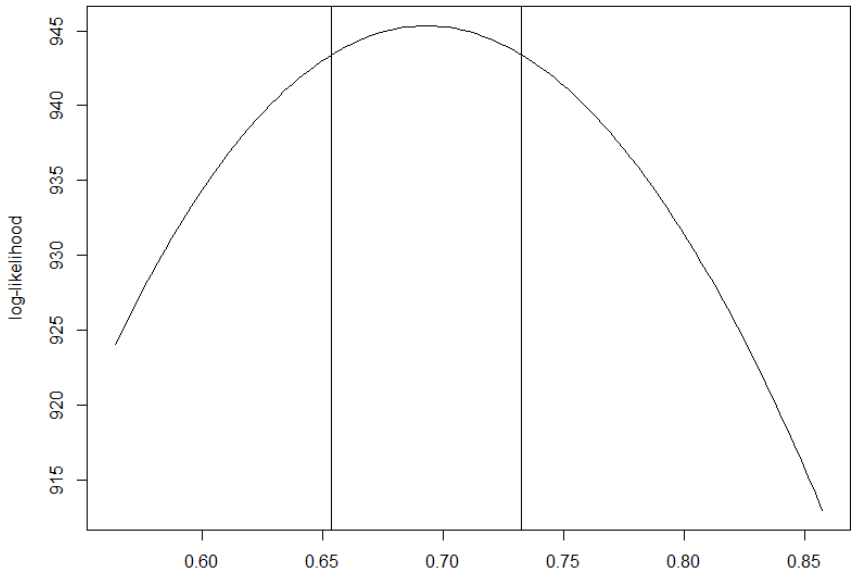


Figure B.120: The GCB profile log-likelihood function for the Clayton copula with parameter $\hat{\beta} = 0.69$ for the extremes of T_{MRD} and T_{MRS} in the P1 time horizon.

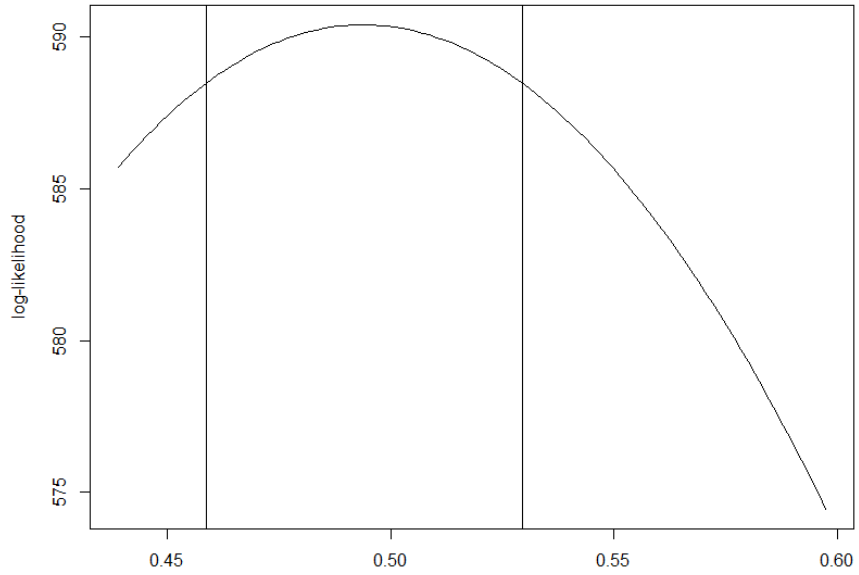


Figure B.121: The GCB profile log-likelihood function for the Clayton copula with parameter $\hat{\beta} = 0.48$ for the extremes of T_{MRD} and T_{MRS} in the P2 time horizon.

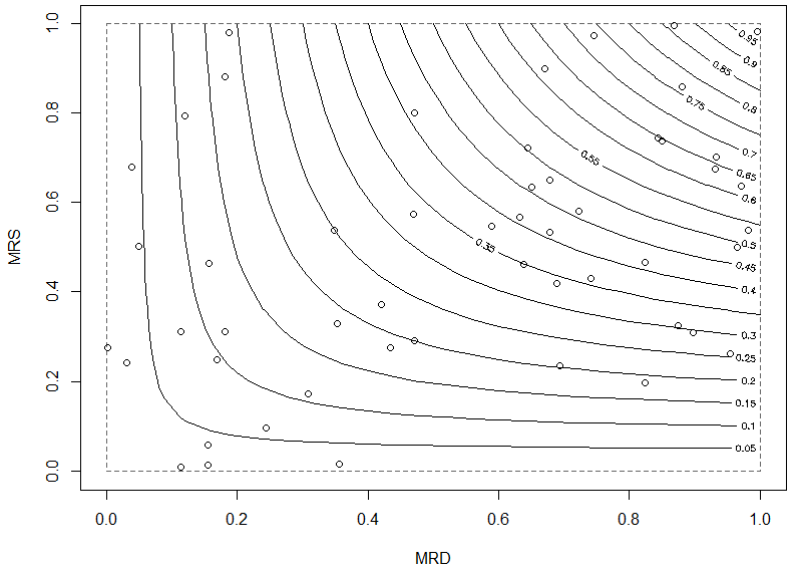


Figure B.122: The critical layers of the Clayton copula under the OR hazard scenario for the time series T_{MRD} and T_{MRS} in P1 time horizon. The black circles represents the empirical transformed vector of extreme points.

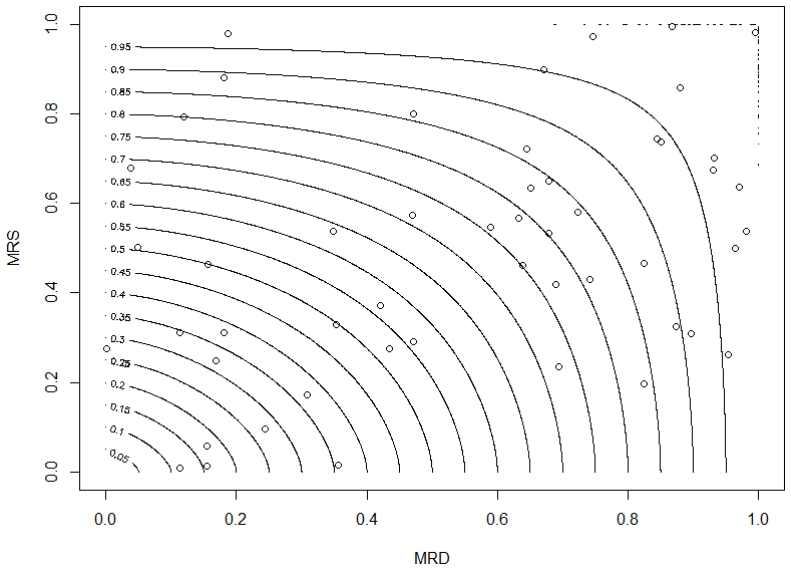


Figure B.123: The critical layers of the Clayton copula under the AND hazard scenario for the time series T_{MRD} and T_{MRS} in P1 time horizon. The black circles represents the empirical transformed vector of extreme points.

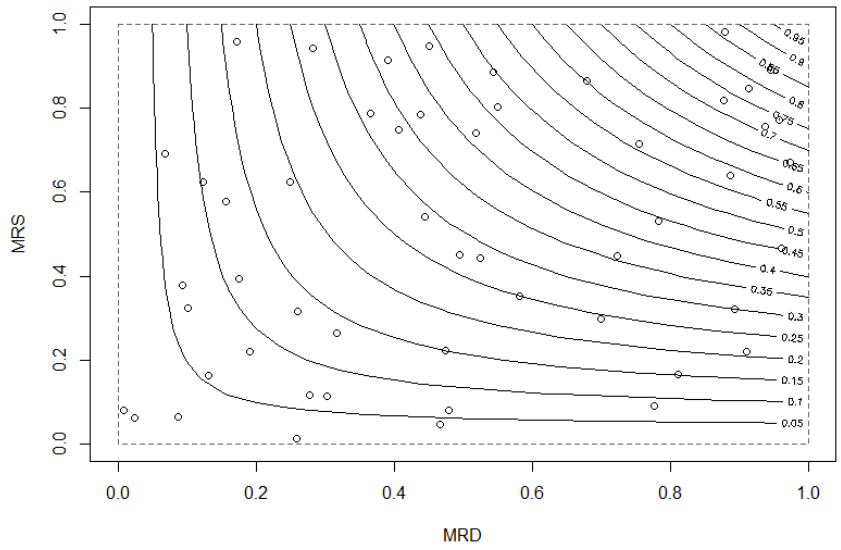


Figure B.124: The critical layers of the Clayton copula under the OR hazard scenario for the time series T_{MRD} and T_{MRS} in P2 time horizon. The black circles represents the empirical transformed vector of extreme points.

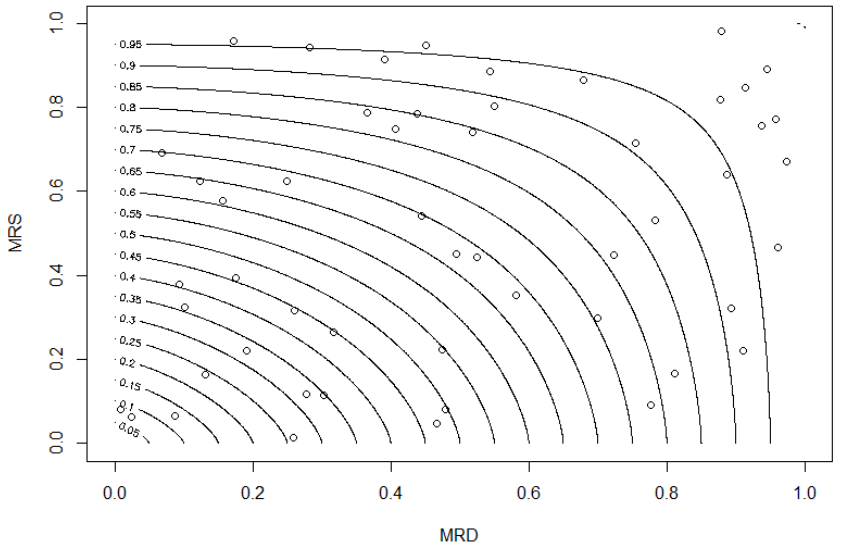


Figure B.125: The critical layers of the Clayton copula under the AND hazard scenario for the time series T_{MRD} and T_{MRS} in P2 time horizon. The black circles represents the empirical transformed vector of extreme points.

			Return Periods	Quantile	Lower CI	Estimated	Upper CI
P1	OR ($\vee$)	T_{MRD}	1 in 20 year event	0.974	5460.66	5682.65	5703.45
			1 in 100 year event	0.995	6035.52	6054.54	6124.54
			1 in 200 year event	0.997	6156.53	6186.19	6292.36
		T_{MRS}	1 in 20 year event	0.974	65.89	75.67	77.98
			1 in 100 year event	0.995	111.33	111.48	112.03
			1 in 200 year event	0.997	120.71	135.77	276.06
	AND ($\wedge$)	T_{MRD}	1 in 20 year event	0.817	3453.07	3688.13	4054.59
			1 in 100 year event	0.921	3925.63	4176.42	4626.83
			1 in 200 year event	0.945	4096.03	4365.90	4822.43
		T_{MRS}	1 in 20 year event	0.817	19.81	20.99	22.33
			1 in 100 year event	0.921	22.28	23.33	24.04
			1 in 200 year event	0.945	23.14	24.09	24.50
P2	OR ($\vee$)	T_{MRD}	1 in 20 year event	0.975	5029.24	5041.63	5125.17
			1 in 100 year event	0.995	6119.12	6208.54	6454.54
			1 in 200 year event	0.997	6779.46	6869.81	6872.52
		T_{MRS}	1 in 20 year event	0.975	51.63	53.13	57.62
			1 in 100 year event	0.995	15.29	66.13	70.44
			1 in 200 year event	0.997	55.28	72.96	73.02
	AND ($\wedge$)	T_{MRD}	1 in 20 year event	0.808	2854.23	2893.40	2910.52
			1 in 100 year event	0.916	3152.98	3283.89	3305.45
			1 in 200 year event	0.941	3423.56	3437.95	3487.12
		T_{MRS}	1 in 20 year event	0.808	18.19	18.25	22.25
			1 in 100 year event	0.916	19.24	20.97	21.47
			1 in 200 year event	0.941	20.85	22.03	22.89

Table B.12: The Most-Likely design realizations for each of the three return period scenarios, according to the OR and AND hazard scenarios that are based on the Clayton copula for both time horizon. The corresponding confidence intervals are also included.

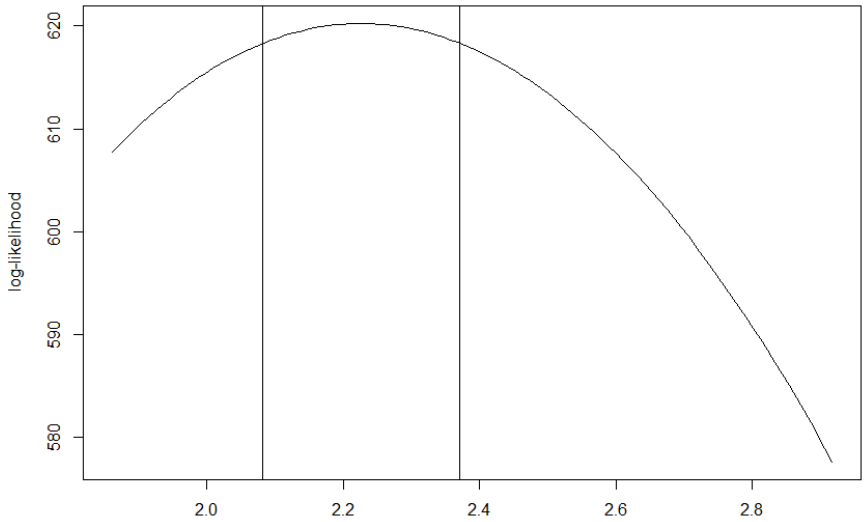


Figure B.126: The GCB profile log-likelihood function for the Frank copula with parameter $\hat{\gamma} = 2.20$ for the extremes of T_{MRD} and T_{MRS} in the P1 time horizon.

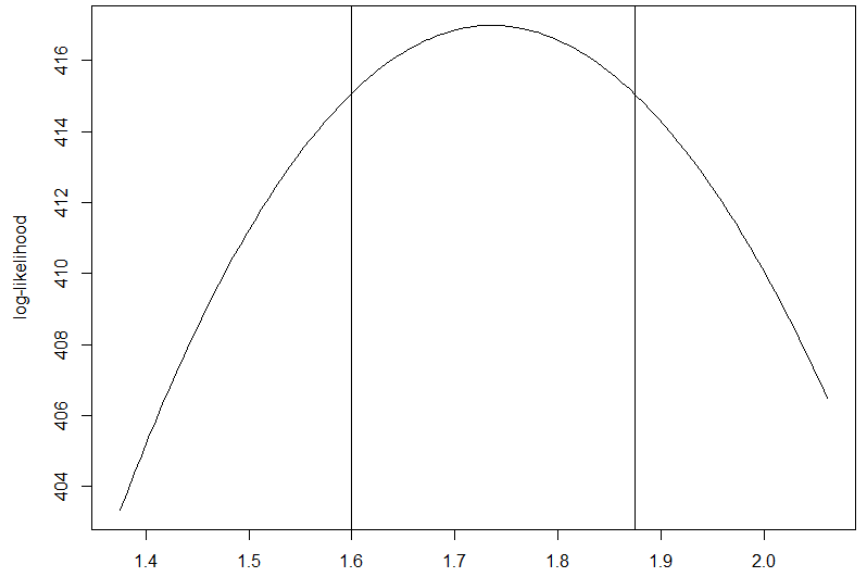


Figure B.127: The GCB profile log-likelihood function for the Frank copula with parameter $\hat{\gamma} = 1.75$ for the extremes of T_{MRD} and T_{MRS} in the P2 time horizon.

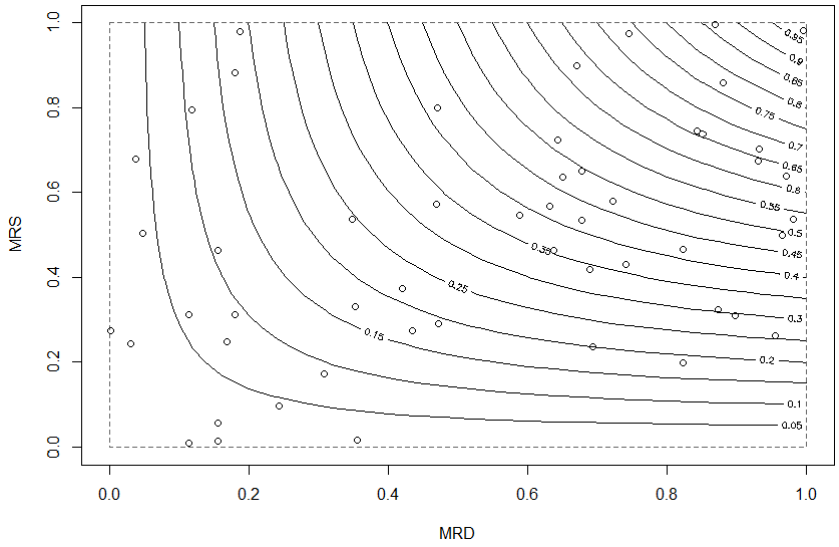


Figure B.128: The critical layers of the Frank copula under the OR hazard scenario for the time series T_{MRD} and T_{MRS} in P1 time horizon. The black circles represents the empirical transformed vector of extreme points.

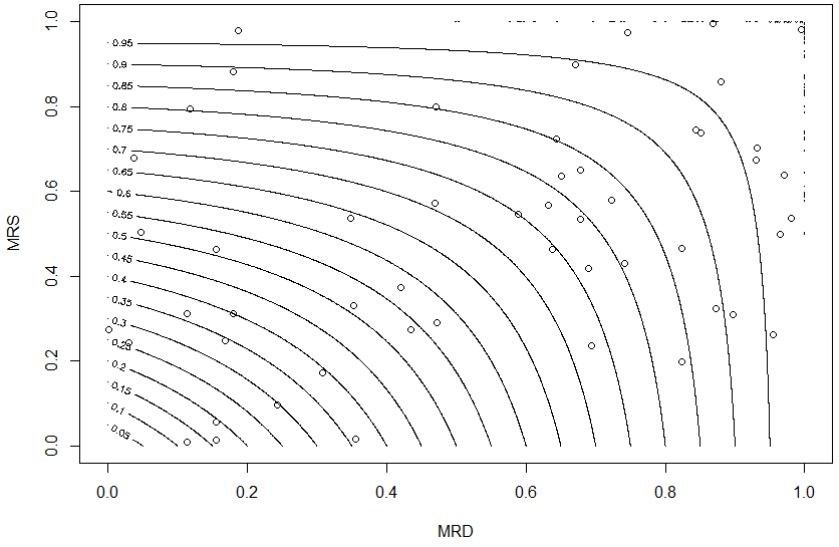


Figure B.129: The critical layers of the Frank copula under the AND hazard scenario for the time series T_{MRD} and T_{MRS} in P1 time horizon. The black circles represents the empirical transformed vector of extreme points.

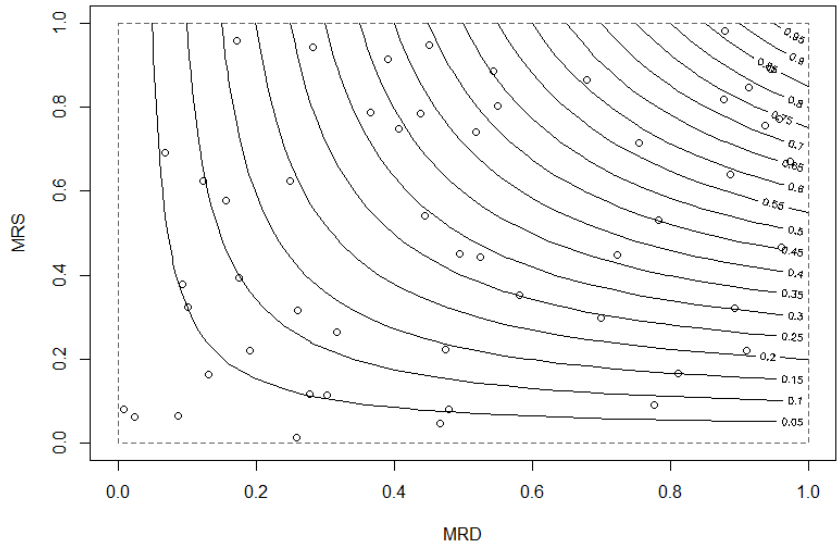


Figure B.130: The critical layers of the Frank copula under the OR hazard scenario for the time series T_{MRD} and T_{MRS} in P2 time horizon. The black circles represents the empirical transformed vector of extreme points.

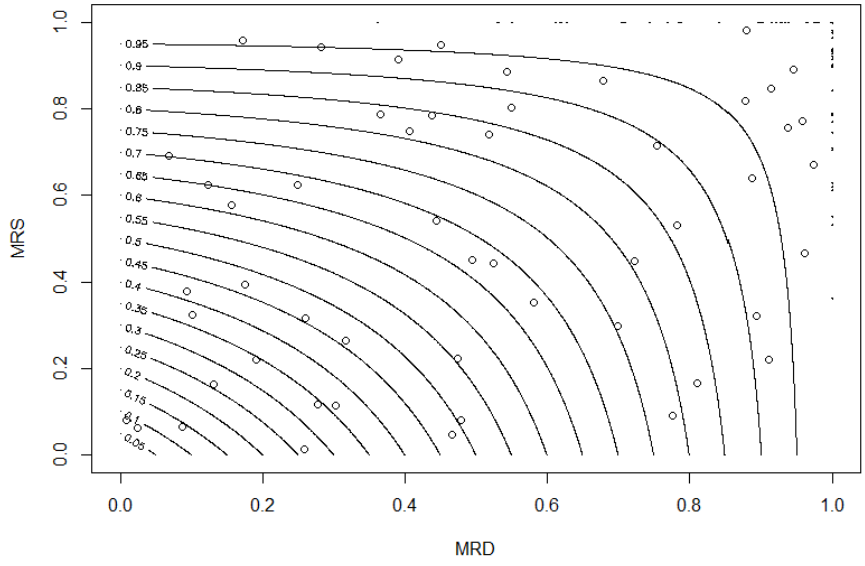


Figure B.131: The critical layers of the Frank copula under the AND hazard scenario for the time series T_{MRD} and T_{MRS} in P2 time horizon. The black circles represents the empirical transformed vector of extreme points.

			Return Periods	Quantile	Lower CI	Estimated	Upper CI
P1	OR ($\vee$)	T_{MRD}	1 in 20 year event	0.974	5460.66	5651.26	5682.65
			1 in 100 year event	0.995	5899.52	6054.54	6056.58
			1 in 200 year event	0.997	6096.23	6186.19	6280.36
		T_{MRS}	1 in 20 year event	0.974	68.57	69.21	70.36
			1 in 100 year event	0.995	98.82	111.30	120.77
			1 in 200 year event	0.997	135.68	135.77	136.22
	AND ($\wedge$)	T_{MRD}	1 in 20 year event	0.835	3292.10	4087.05	4151.17
			1 in 100 year event	0.932	3846.27	4597.58	4616.97
			1 in 200 year event	0.953	4090.04	4779.14	4785.12
		T_{MRS}	1 in 20 year event	0.835	17.48	23.58	24.95
			1 in 100 year event	0.932	18.37	26.23	26.83
			1 in 200 year event	0.953	18.46	27.03	27.12
P2	OR ($\vee$)	T_{MRD}	1 in 20 year event	0.974	5013.00	5041.63	5622.17
			1 in 100 year event	0.995	5899.52	6208.54	6354.54
			1 in 200 year event	0.997	6054.54	6869.81	6883.36
		T_{MRS}	1 in 20 year event	0.974	49.51	51.33	73.99
			1 in 100 year event	0.995	63.57	64.36	82.33
			1 in 200 year event	0.997	49.55	75.26	135.42
	AND ($\wedge$)	T_{MRD}	1 in 20 year event	0.826	2385.27	2976.69	3244.84
			1 in 100 year event	0.927	2601.43	3411.14	3631.90
			1 in 200 year event	0.949	2667.36	3575.97	3784.05
		T_{MRS}	1 in 20 year event	0.826	14.70	18.84	20.71
			1 in 100 year event	0.927	16.22	21.89	23.40
			1 in 200 year event	0.949	16.67	23.05	24.47

Table B.13: The Most-Likely design realizations for each of the three return period scenarios, according to the OR and AND hazard scenarios that are based on the Frank copula for both time horizon. The corresponding confidence intervals are also included.

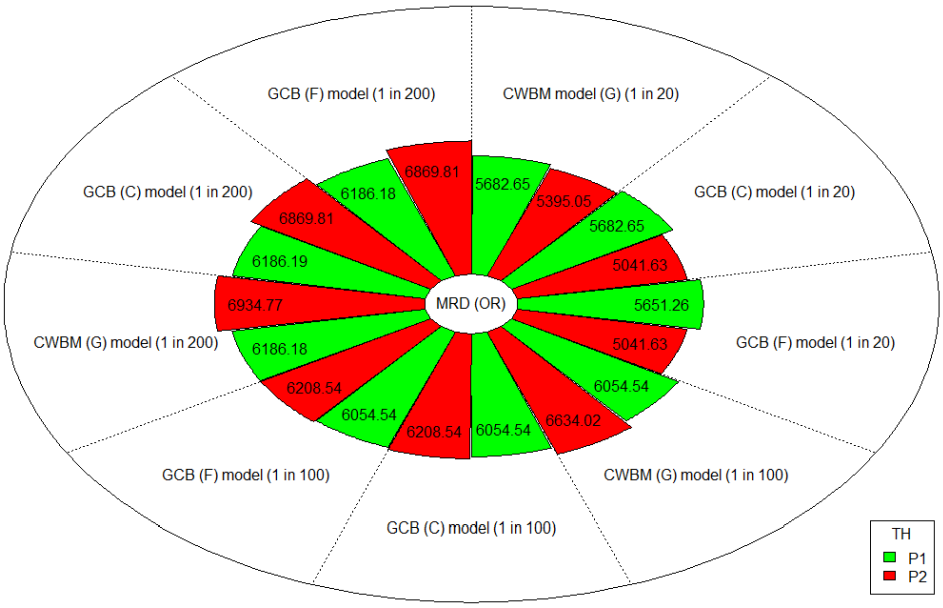


Figure B.132: Comparison of the flood scenarios generated by the three multivariate models under the OR HS using the time series T_{MRD} in both time horizons.

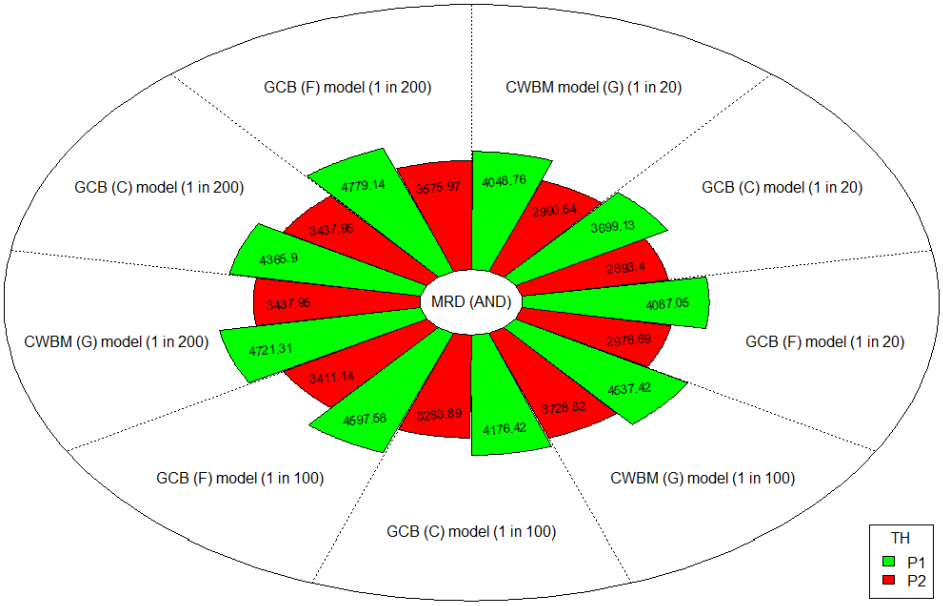


Figure B.133: Comparison of the flood scenarios generated by the three multivariate models under the AND HS using the time series T_{MRD} in both time horizons.

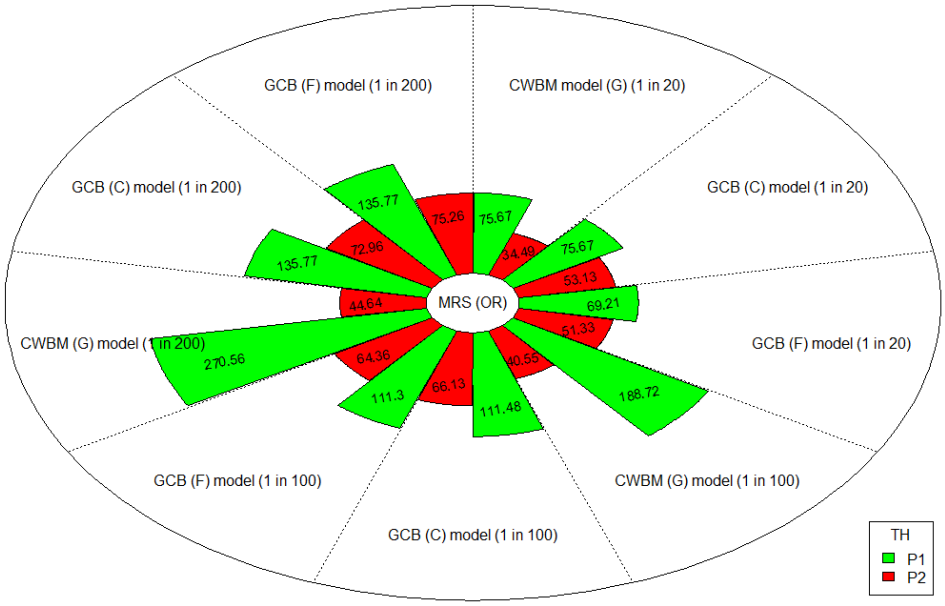


Figure B.134: Comparison of the flood scenarios generated by the three multivariate models under the OR HS using the time series T_{MRS} in both time horizons.

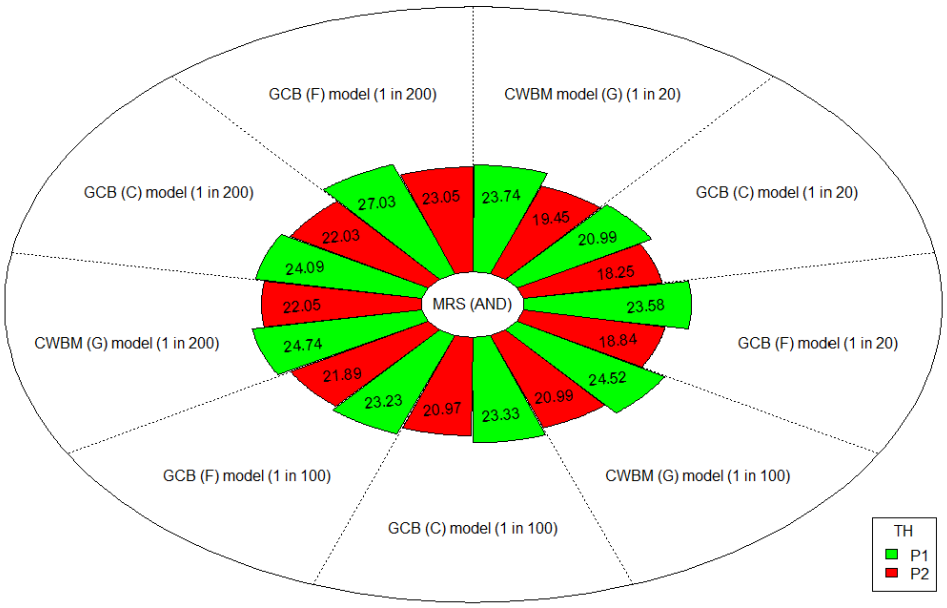


Figure B.135: Comparison of the flood scenarios generated by the three multivariate models under the AND HS using the time series T_{MRS} in both time horizons.

Bibliography

(1) Jonathan Kwaku Afriyie et al. "Evaluating the performance of unit root tests in single time series processes". In: *Mathematics and Statistics* 8.6 (2020), pp. 656–664.

(2) Amir AghaKouchak et al. "Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 California drought". In: *Geophysical Research Letters* 41.24 (2014), pp. 8847–8852.

(3) GN Alexander. "Some aspects of time series in hydrology". In: *J. Inst. Eng. Aust* 26.9 (1954), p. 194.

(4) Heiko Apel, Bruno Merz, and AH Thieken. "Influence of dike breaches on flood frequency estimation". In: *Computers & Geosciences* 35.5 (2009), pp. 907–923.

(5) Barry C Arnold, Narayanaswamy Balakrishnan, and Haikady Navada Nagaraja. *A first course in order statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1992.

(6) Daniel Åsljung, Jonas Nilsson, and Jonas Fredriksson. "Using Extreme Value Theory for Vehicle Level Safety Validation and Implications for Autonomous Vehicles". In: *IEEE Transactions on Intelligent Vehicles* 2.4 (2017), pp. 288–297. DOI: 10.1109/TIV.2017.2768219.

(7) Brian Bader, Jun Yan, and Xuebin Zhang. "Automated selection of r for the r largest order statistics approach with adjustment for sequential testing". In: *Statistics and Computing* 27.6 (2017), pp. 1435–1451.

(8) August A Balkema and Laurens De Haan. "Residual life time at great age". In: *The Annals of Probability* 2.5 (1974), pp. 792–804. ISSN: 00911798.

(9) Nikolina Ban et al. "Analysis of Alpine precipitation extremes using generalized extreme value theory in convection-resolving climate simulations". In: *Climate Dynamics* 55.1 (2020), pp. 61–75. ISSN: 1432-0894. DOI: 10.1007/s00382-018-4339-4.

(10) V Barnet. "The ordering of multivariate data (with discussion)". In: *Journal of the Royal Statistics Society, Series A* 139 (1976), pp. 318–354.

(11) Jules J Beersma and T Adri Buishand. "Joint probability of precipitation and discharge deficits in the Netherlands". In: *Water resources research* 40.12 (2004).

(12) Jan Beirlant et al. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley, 2004. ISBN: 9780471976479.

(13) Emanuele Bevacqua et al. "Multivariate statistical modelling of compound events via pair-copula constructions: analysis of floods in Ravenna (Italy)". In: *Hydrology and Earth System Sciences* 21.6 (2017), pp. 2701–2723.

(14) Ladislaus von Bortkiewicz. "Variationsbreite und mittlerer Fehler". In: *Sitzungsberichte der Berliner Mathematischen Gesellschft* 21 (1922), pp. 3–11.

(15) Laurens M Bouwer. "Observed and projected impacts from extreme weather events: implications for loss and damage". In: *Loss and damage from climate change: concepts, methods and policy options* (2019), pp. 63–82.

(16) B Büchele et al. "Flood-risk mapping: contributions towards an enhanced assessment of extreme events and associated risks". In: *Natural Hazards and Earth System Sciences* 6.4 (2006), pp. 485–503.

(17) Wasserstraßen Schifffahrtsverwaltung des Bundes. *Services*. URL: https://www.gdws.wsv.bund.de/DE/startseite/startseite_node.html (visited on 01/25/2022).

(18) S Caires. "A comparative simulation study of the annual maxima and the peaks-over-threshold methods". In: *Deltares report 1200264-002 for Rijkswaterstaat, Waterdienst* (2009).

(19) Enrique Castillo and Ali S Hadi. "Fitting the generalized Pareto distribution to data". In: *Journal of the American Statistical Association* 92.440 (1997), pp. 1609–1620.

(20) Düsseldorf Information Center. *Office for Statistics and Elections*. URL: `https://www.duesseldorf.de/rheinpegel.html`.

(21) The Global Runoff Data Centre. *Standard Services, River Discharge Data, GRDC Data Download*. URL: `https://www.bafg.de/GRDC/EN/Home/homepage_node.html` (visited on 01/25/2022).

(22) Fateh Chebana and Taha BMJ Ouarda. "Multivariate quantiles in hydrological frequency analysis". In: *Environmetrics* 22.1 (2011), pp. 63–78.

(23) Jiangpeng Chen et al. "Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang, China". In: *PloS one* 10.2 (2015), e0118521.

(24) Mark Cioc. *The Rhine: an eco-biography, 1815-2000*. University of Washington Press, 2002.

(25) Jonathan P. Cohen. "Fitting Extreme Value Distributions to Maxima". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 50.1 (1988), pp. 74–97. ISSN: 0581572X.

(26) Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. 1. Springer, London, 2001, p. 209.

(27) Stuart G Coles and Jonathan A Tawn. "A Bayesian analysis of extreme rainfall data". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 45.4 (1996), pp. 463–478.

(28) Stuart G Coles and Jonathan A Tawn. "Modelling extreme multivariate events". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.2 (1991), pp. 377–392.

(29) Environmental Commissariat. URL: `https://environment.ec.europa.eu/topics/water/floods_en` (visited on 01/19/2023).

(30) Daniel Cooley. "Return periods and return levels under climate change". In: *Extremes in a changing climate*. Springer, 2013, pp. 97–114.

(31) D.R. Cox and O.E. Barndorff-Nielsen. *Inference and Asymptotics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994. ISBN: 9780412494406.

(32) David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

(33) Alireza Daneshkhah et al. "Probabilistic modeling of flood characterizations with parametric and minimum information pair-copula model". In: *Journal of Hydrology* 540 (2016), pp. 469–487.

(34) A. C. Davison and R. L. Smith. "Models for Exceedances over High Thresholds". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 52.3 (1990), pp. 393–442. ISSN: 00359246.

(35) Anthony C Davison. "Modelling excesses over high thresholds, with an application". In: *Statistical extremes and applications*. Springer, 1984, pp. 461–482.

(36) Laurens De Haan and Sidney I Resnick. "Limit theory for multivariate sample extremes". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 40.4 (1977), pp. 317–337.

(37) Francesco De Leo et al. "Non-stationary extreme value analysis of sea states based on linear trends. Analysis of annual maxima series of significant wave height and peak period in the Mediterranean Sea". In: *Coastal Engineering* 167 (2021), p. 103896.

(38) Jan Dhaene and Marc J Goovaerts. "Dependency of Risks and Stop-Loss Order1". In: *ASTIN Bulletin: The Journal of the IAA* 26.2 (1996), pp. 201–212.

(39) P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: for Insurance and Finance*. Applications of Mathematics. Springer Berlin Heidelberg, 1997. ISBN: 978-3-642-08242-9.

(40) H Engel. "The flood events of 1993/1994 and 1995 in the Rhine River basin". In: *IAHS Publications-Series of Proceedings and Reports-Intern Assoc Hydrological Sciences* 239 (1997), pp. 21–32.

(41) "Extremal distribution". In: *Universidade de Lisboa. Revista da Faculdade de Ciências. 2a Série. A: Ciências Matemáticas* 7 (1958), pp. 215–227.

(42) "Extreme value modeling under power normalization". In: *Applied Mathematical Modelling* 37.24 (2013), pp. 10162–10169. ISSN: 0307-904X.

(43) M. Falk, J. Husler, and R.D. Reiss. *Laws Of Small Numbers: Extremes And Rare Events*. Seminare der Deutschen Mathematiker-Vereinigung Series. Springer Basel AG, 2004. ISBN: 9783764324162.

(44) Simone A. Falk Michaeland Padoan and Florian Wisheckel. "Generalized Pareto copulas: A key to multivariate extremes". In: *Journal of Multivariate Analysis* 174 (2019), p. 104538.

(45) A-C Favre, André Musy, and Stephan Morgenthaler. "Two-site modeling of rainfall based on the Neyman-Scott process". In: *Water Resources Research* 38.12 (2002), pp. 43–1.

(46) Jean-David Fermanian. "Goodness-of-fit tests for copulas". In: *Journal of multivariate analysis* 95.1 (2005), pp. 119–152.

(47) Ronald Fisher and Leonard Tippett. "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (1928), pp. 180–190. DOI: 10.1017/S0305004100015681.

(48) Bastien François and Mathieu Vrac. "Time of Emergence of compound events: contribution of univariate and dependence properties". In: *Natural Hazards and Earth System Sciences* 23.1 (2023), pp. 21–44.

(49) Maurice Fréchet. "Sur la loi de probabilité de l'écart maximum". In: *Annales De La Société Polonaise De Mathématique* 6 (1927), pp. 93–116.

(50) Mauritius Fréchet. "On correlation tables with given margins". In: 14 (1951), pp. 53–77.

(51) G Friedrich and D Müller. "Rhine. Ecology of European Rivers". In: (1984), pp. 265–315.

(52) Hans Friesland and Franz-Josef Löpmeier. "The performance of the model AMBAV for evapotranspiration and soil moisture on Müncheberg data". In: *Modelling water and nutrient dynamics in soil–crop systems*. Springer, 2007, pp. 19–26.

(53) Jean Geffroy. "Contributions à la théorie des valeurs extrêmes". In: *Publ. Inst. Statist. Univ. Paris* 7 (1958), pp. 37–185.

(54) Christian Genest, Kilani Ghoudi, and L-P Rivest. "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions". In: *Biometrika* 82.3 (1995), pp. 543–552.

(55) Christian Genest and Bruno Rémillard. "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 44. 6. 2008, pp. 1096–1127.

(56) Christian Genest and Louis-Paul Rivest. "A characterization of Gumbel's family of extreme value distributions". In: *Statistics and Probability Letters* 8.3 (1989), pp. 207–211.

(57) Christian Genest and Louis-Paul Rivest. "Statistical inference procedures for bivariate Archimedean copulas". In: *Journal of the American statistical Association* 88.423 (1993), pp. 1034–1043.

(58) B. Gnedenko. "Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire". In: *Annals of Mathematics* 44.3 (1943), pp. 423–453. ISSN: 0003486X.

(59) Vidyadhar P Godambe. "An optimum property of regular maximum likelihood estimation". In: *The Annals of Mathematical Statistics* 31.4 (1960), pp. 1208–1211.

(60) Scott D. Grimshaw. "Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution". In: *Technometrics* 35.2 (1993), pp. 185–191. DOI: 10.1080/00401706.1993.10485040.

(61) Pavel Ya Groisman et al. "Changes in the probability of heavy precipitation: important indicators of climatic change". In: *Weather and climate extremes*. Springer, 1999, pp. 243–283.

(62) Emil Julius Gumbel. *La durée extrême de la vie humaine*. Vol. 520. Hermann et cie, 1937.

(63) Emil Julius Gumbel. "Les intervalles extrêmes entre les émissions radioactives. II". In: *Journal de Physique et le Radium* 8.11 (1937), pp. 446–452.

(64) Emil Julius Gumbel. "On the plotting of flood-discharges". In: *Transactions, American Geophysical Union* 24.2 (1944), pp. 699–719.

(65) Emil Julius Gumbel. *Statistics of extremes*. Columbia university press, 1958. DOI: doi:10.7312/gumb92958.

(66) Emil Julius Gumbel. "The Return Period of Flood Flows". In: *The Annals of Mathematical Statistics* 12.2 (1941), pp. 163–190. ISSN: 00034851.

(67) L. de Haan. *Sample Extremes: An Elementary Introduction*. ERASMUS UNIVERSITY, ROTTERDAM. INSTITUTE FOR ECONOMIC RESEARCH. DISCUSSION PAPER. Erasmus Universiteit Rotterdam. Econometrisch Instituut, 1976.

(68) Laurens de Haan. "A form of regular variation and its application to the domain of attraction of the double exponential distribution". In: *Wahrscheinlichkeitstheorie verw Gebiete* 17 (1971), pp. 241–258.

(69) Laurens de Haan. "On Regular Variation and Its Application to the Weak Convergence of Sample Extremes". In: *Mathematical Centre Tracts* 32 (1970).

(70) Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research. Springer, 2006. ISBN: 9780387239460.

(71) Zengchao Hao et al. "A multivariate approach for statistical assessments of compound extremes". In: *Journal of hydrology* 565 (2018), pp. 87–94.

(72) Janet Heffernan and Jonathan Tawn. "A Conditional Approach for Multivariate Extreme Values". In: *Journal of the Royal Statistical Society Series B* 66.3 (2004), pp. 497–546. DOI: 10.1111/j.1467-9868.2004.02050.x.

(73) Mark Hegnauer, Jaap Kwadijk, and Frans Klijn. *The plausibility of extreme high discharges in the river Rhine*. Deltares Delft, The Netherlands, 2015.

(74) J. R. M. Hosking. "Algorithm AS 215: Maximum-Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34.3 (1985), pp. 301–310. ISSN: 00359254, 14679876.

(75) J. R. M. Hosking, J. R. Wallis, and E. F. Wood. "Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments". In: *Technometrics* 27.3 (1985), pp. 251–261. ISSN: 00401706.

(76) Jonathan RM Hosking. "Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34.3 (1985), pp. 301–310.

(77) Jonathan RM Hosking and James R Wallis. "Parameter and quantile estimation for the generalized Pareto distribution". In: *Technometrics* 29.3 (1987), pp. 339–349.

(78) Philip Hougaard. *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer New York, 2000. ISBN: 9780387988733.

(79) James Pickands III. "Statistical inference using extreme order statistics". In: *The Annals of statistics* 3.1 (1975), pp. 119–131. DOI: 10.1214/aos/1176343003.

(80) Pickands III. "Multivariate extreme value distributions". In: *Proceedings of the 43rd session of the International Statistical Institute, Vol. 2 (Buenos Aires)*. Vol. 29. 1981, pp. 859–878, 894–902.

(81) Swiss Re Institute. *Sigma*. URL: https://www.swissre.com/dam/jcr:8a467970-b803-4783-8b92-63f62685d93b/swiss-re-institute-natcat-sigma-2022.pdf (visited on 09/24/2022).

(82) Mohammad Tariquel Islam and Kumer Pial Das. "Predicting Bitcoin Return Using Extreme Value Theory". In: *American Journal of Mathematical and Management Sciences* 40.2 (2021), pp. 177–187. DOI: 10.1080/01966324.2021.1950086.

(83) Robert Jane et al. "Multivariate statistical modelling of the drivers of compound flood events in south Florida". In: *Natural Hazards and Earth System Sciences* 20.10 (2020), pp. 2681–2699.

(84) P. Jaworski et al. *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*. Lecture Notes in Statistics. Springer Berlin Heidelberg, 2010. ISBN: 9783642124655.

(85) AF Jenkinson. "Statistics of extremes". In: *Estimation of maximum floods, WMO* 233 (1969), pp. 183–228.

(86) Arthur F Jenkinson. "The frequency distribution of the annual maximum (or minimum) values of meteorological elements". In: *Quarterly Journal of the Royal Meteorological Society* 81.348 (1955), pp. 158–171.

(87) H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1997. ISBN: 9780412073311.

(88) Harry Joe. "Asymptotic efficiency of the two-stage estimation method for copula-based models". In: *Journal of multivariate Analysis* 94.2 (2005), pp. 401–419.

(89) Harry Joe. "Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters". In: *Lecture Notes-Monograph Series* (1996), pp. 120–141.

(90) Harry Joe and James Jianmeng Xu. "The estimation method of inference functions for margins for multivariate models". In: (1996).

(91) F Kaspar et al. "Monitoring of climate change in Germany–data, products and services of Germany's National Climate Data Centre". In: *Advances in Science and Research* 10.1 (2013), pp. 99–106.

(92) Richard W Katz and Barbara G Brown. "Extreme events in a changing climate: variability is more important than averages". In: *Climatic change* 21.3 (1992), pp. 289–302.

(93) RW Katz. "Extreme value theory for precipitation: Implications for climate change". In: *AMS Proceedings of the Ninth Symposium on Global Change Studies, Phoenix, Arizona*. 1998, pp. 11–16.

(94) SF Kew et al. "The simultaneous occurrence of surge and discharge extremes for the Rhine delta". In: *Natural hazards and earth system sciences* 13.8 (2013), pp. 2017–2029.

(95) RR Kinnison. *Applied extreme-value statistics*. Tech. rep. Pacific Northwest Lab., Richland, WA (USA), 1983.

(96) Anna Kiriliouk et al. "Peaks over thresholds modeling with multivariate generalized Pareto distributions". In: *Technometrics* 61.1 (2019), pp. 123–135.

(97) S. Kotz and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. Imperial College Press, 2000. ISBN: 9781860942242.

(98) DF Kroekenstoel and R Lammersen. "Transboundary effects of extreme floods on the Lower Rhine". In: *NCR-days 2004* (2005).

(99) Amelie Krug et al. "On the temporal variability of widespread rain-on-snow floods". In: *Meteorol. Z.* 29.2 (2020), pp. 147–163.

(100) Kristian Kumbier et al. "Investigating compound flooding in an estuary using hydrodynamic modelling: a case study from the Shoalhaven River, Australia". In: *Natural Hazards and Earth System Sciences* 18.2 (2018), pp. 463–477.

(101) Sang Kyu Kwak and Jong Hae Kim. "Statistical data preparation: management of missing values and outliers". In: *Korean journal of anesthesiology* 70.4 (2017), p. 407.

(102) M.R. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer New York, 1983. ISBN: 9780387907314.

(103) Dennis P Lettenmaier and Stephen J Burges. "Gumbel's extreme value I distribution: a new look". In: *Journal of the Hydraulics Division* 108.4 (1982), pp. 502–514.

(104) Jan-frederik Mai and Matthias Scherer. *Simulating Copulas: Stochastic Models, Sampling Algorithms, And Applications (Second Edition)*. Series In Quantitative Finance. World Scientific Publishing Company, 2017. ISBN: 9789813149267.

(105) Michael Marcus and Mark Pinsky. "On the domain of attraction of e- e- x". In: *Journal of Mathematical Analysis and Applications* 28.2 (1969), pp. 440–449.

(106) Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Business Pro collection. Princeton University Press, 2005. ISBN: 9780691122557.

(107) William Q Meeker and Luis A Escobar. "Teaching about approximate confidence regions based on maximum likelihood estimation". In: *The American Statistician* 49.1 (1995), pp. 48–53.

(108) David G Mejzler. "On a theorem of BV Gnedenko". In: *Sb. Trudov Inst. Mat. Akad. Nauk. Ukrain. RSR* 12 (1949), pp. 31–35.

(109) Vivian Meløysund et al. "Predicting snow density using meteorological data". In: *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling* 14.4 (2007), pp. 413–423.

(110) Philippe Naveau et al. "Statistical methods for the analysis of climate extremes". In: *Comptes Rendus Geoscience* 337.10-11 (2005), pp. 1013–1022.

(111) R.B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2007. ISBN: 9780387286785.

(112) "North America drought: A paleo perspective." In: *National Oceanic and Atmospheric Administration Climatology Program* (2000).

(113) T Okubo and N Narita. "On the distribution of extreme winds expected in Japan". In: *National Bureau of Standards Special Publication* 560.1 (1980).

(114) Pardeep Pall, Lena M Tallaksen, and Frode Stordal. "A climatology of rain-on-snow events for Norway". In: *Journal of Climate* 32.20 (2019), pp. 6995–7016.

(115) Andrew J Patton. "Estimation of multivariate models for time series of possibly different lengths". In: *Journal of applied econometrics* 21.2 (2006), pp. 147–173.

(116) Debbie J. Dupuis PhD and FCIA Bruce L. Jones PhD FSA. "Multivariate Extreme Value Theory And Its Usefulness In Understanding Risk". In: *North American Actuarial Journal* 10.4 (2006), pp. 1–27. DOI: 10.1080/10920277.2006.10597411.

(117) Jeroen Aerts Philip Bubeck Aline te Linde. *Assessment of upstream flood risk in the Rhine Basin (HSGR02)*. URL: https://edepot.wur.nl/319229.

(118) UFZ DATAINVESTIGATION PORTAL. *Water fluxes and states dataset for Germany from 1951 to 2010*. URL: http://www.ufz.de/record/dmp/archive/4376 (visited on 01/25/2022).

(119) P Prescott and AT Walden. "Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples". In: *Journal of Statistical Computation and Simulation* 16.3-4 (1983), pp. 241–250.

(120) P. Prescott and A. T. Walden. "Maximum likelihood estimation of the parameters of the generalized extreme-value distribution". In: *Biometrika* 67.3 (Dec. 1980), pp. 723–724. ISSN: 0006-3444. DOI: 10.1093/biomet/67.3.723.

(121) Ali Razmi et al. "Time-Varying Univariate and Bivariate Frequency Analysis of Nonstationary Extreme Sea Level for New York City". In: *Environmental Processes* 9.1 (2022), p. 8. ISSN: 2198-7505.

(122) Joeri B Reinders and Samuel E Muñoz. "Improvements to Flood Frequency Analysis on Alluvial Rivers Using Paleoflood Data". In: *Water Resources Research* 57.4 (2021).

(123) R.D. Reiss. *Approximate Distributions of Order Statistics*. Springer New York, 1989.

(124) R.D. Reiss and M. Thomas. *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Springer Basel AG, 2007. ISBN: 9783764372309.

(125) A. Rényi. "On measures of Dependence". In: 10 (1959), pp. 441–451.

(126) S.I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Applied probability. Springer, 2008. ISBN: 9780387759524.

(127) R Riaman et al. "Analysing the decision making for agricultural risk assessment: An application of extreme value theory". In: *Decision Science Letters* 10.3 (2021), pp. 351–360.

(128) Riaman et al. "Determining the premium of paddy insurance using the extreme value theory method and the operational value at risk approach". In: *Journal of Physics: Conference Series* 1722.1 (2021), p. 012059. DOI: 10.1088/1742-6596/1722/1/012059.

(129) Holger Rootzén, Johan Segers, and Jennifer L. Wadsworth. "Multivariate peaks over thresholds models". In: *Extremes* 21.1 (2018), pp. 115–145. ISSN: 1572-915X. DOI: 10.1007/s10687-017-0294-4.

(130) Holger Rootzén, Johan Segers, and Jennifer L. Wadsworth. "Multivariate generalized Pareto distributions: Parametrizations, representations, and properties". In: *Journal of Multivariate Analysis* 165 (2018), pp. 117–131. ISSN: 0047-259X.

(131) Mojtaba Sadegh et al. "Multihazard scenarios for analysis of compound extreme events". In: *Geophysical Research Letters* 45.11 (2018), pp. 5470–5480.

(132) G Salvadori and Carlo De Michele. "Frequency analysis via copulas: Theoretical aspects and applications to hydrological events". In: *Water resources research* 40.12 (2004).

(133) G Salvadori and Carlo De Michele. "Multivariate multiparameter extreme value models and return periods: A copula approach". In: *Water resources research* 46.10 (2010).

(134) G. Salvadori et al. *Extremes in Nature: An Approach Using Copulas*. Water Science and Technology Library. Springer Netherlands, 2007. ISBN: 9781402044151.

(135) Gianfausto Salvadori, Carlo De Michele, and Fabrizio Durante. "On the return period and design in a multivariate framework". In: *Hydrology and Earth System Sciences* 15.11 (2011), pp. 3293–3305.

(136) Gianfausto Salvadori et al. "A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities". In: *Water Resources Research* 52.5 (2016), pp. 3701–3721.

(137) Philipp.J. Schönbucher. *Credit Derivatives Pricing Models: Models, Pricing and Implementation*. The Wiley Finance Series. John Wiley & Sons, 2003. ISBN: 9780470868171.

(138) Berthold Schweizer and Edward F Wolff. "On Nonparametric Measures of Dependence for Random Variables". In: *The Annals of Statistics* 9.4 (1981), pp. 879–885. ISSN: 00905364.

(139) Francesco Serinaldi. "Dismissing return periods!" In: *Stochastic environmental research and risk assessment* 29 (2015), pp. 1179–1189.

(140) Masaaki Sibuya et al. "Bivariate extreme statistics". In: *Annals of the Institute of Statistical Mathematics* 11.2 (1960), pp. 195–210.

(141) Vijendra P Singh and Huizhen Guo. "Parameter estimation for 3-parameter generalized Pareto distribution by the principle of maximum entropy (POME)". In: *Hydrological sciences journal* 40.2 (1995), pp. 165–181.

(142) Chandrasekaran Sivapragasam and Nitin Muttil. "Discharge rating curve extension–a new approach". In: *Water Resources Management* 19.5 (2005), pp. 505–520.

(143) Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. "Climate Change 2022: Impacts, Adaptation and Vulnerability". In: *Cambridge University Press. Cambridge University Press* (2022). DOI: 10.1017/9781009325844.

(144) A Sklar. "Fonctions de répartition à n dimensions et leurs marges". In: *Publ. inst. statist. univ. Paris* 8 (1959), pp. 229–231.

(145) M Sklar. "Distribution functions an dimensions and their margins". In: *Publ. inst. statist. univ. Paris* 8 (1959), pp. 229–231.

(146) M Sklar. "Distribution functions an dimensions and their margins". In: *Publ. inst. statist. univ. Paris* 8 (1959), pp. 229–231.

(147) Richard L Smith. "Extreme value theory based on the r largest annual events". In: *Journal of Hydrology* 86.1-2 (1986), pp. 27–43.

(148) Richard L Smith. "Maximum likelihood estimation in a class of nonregular cases". In: *Biometrika* 72.1 (1985), pp. 67–90.

(149) Richard L Smith and Ishay Weissman. "Estimating the extremal index". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.3 (1994), pp. 515–528.

(150) U Strasser. "Snow loads in a changing climate: new risks?" In: *Natural Hazards and Earth System Sciences* 8.1 (2008), pp. 1–8.

(151) Jonathan A Tawn. "An extreme-value theory model for dependent observations". In: *Journal of Hydrology* 101.1-4 (1988), pp. 227–250.

(152) Jonathan A Tawn. "Bivariate extreme value theory: models and estimation". In: *Biometrika* 75.3 (1988), pp. 397–415.

(153) Tetra Tech. "Statistical analysis for monotonic trends". In: (2011).

(154) Willem HJ Toonen et al. "Lower Rhine historical flood magnitudes of the last 450 years reproduced from grain-size measurements of flood deposits using End Member Modelling". In: *Catena* 130 (2015), pp. 69–81.

(155) Urs F Uehlinger et al. *The Rhine river basin*. 2009.

(156) A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN: 9780521784504.

(157) Owen Vella. *Univariate-and-Multivariate-Extreme-Value-Analysis-of-Dusseldorf-Hydrological-Data*. `https://github.com/OwenkVella/Univariate-and-Multivariate-Extreme-Value-Analysis-of-Dusseldorf-Hydrological-Data.git`. 2023.

(158) Willem W Verstraeten, Frank Veroustraete, and Jan Feyen. "Assessment of evapotranspiration and soil moisture content across different scales of observation". In: *Sensors* 8.1 (2008), pp. 70–117.

(159) Ivan Vorobevskii, Rico Kronenberg, and Christian Bernhofer. "Estimation of joint return periods of compound precipitation-discharge extremes for small catchments". In: *Hydrology and Earth System Sciences Discussions* (2020), pp. 1–23.

(160) Conrad Wasko, Rory Nathan, and Murray C Peel. "Trends in global flood and streamflow timing based on local water year". In: *Water Resources Research* 56.8 (2020).

(161) Ishay Weissman. "Estimation of parameters and large quantiles based on the k largest observations". In: *Journal of the American Statistical Association* 73.364 (1978), pp. 812–815.

(162) Deutscher Wetterdienst. *Climate & Enviroment, CDC (Climate Data Center)*. URL: `https://www.dwd.de/EN/Home/home_node.html` (visited on 01/25/2022).

(163) Chuanhao Wu et al. "Copula-based risk evaluation of global meteorological drought in the 21st century based on CMIP5 multi-model ensemble projections". In: *Journal of Hydrology* 598 (2021), pp. 126–265.

(164) Sheng Yue, Taha BMJ Ouarda, and Bernard Bobée. "A review of bivariate gamma distributions for hydrological application". In: *Journal of Hydrology* 246.1-4 (2001), pp. 1–18.

(165) Ran Zhang, Claudia Czado, and Aleksey Min. "Efficient maximum likelihood estimation of copula based meta t-distributions". In: *Computational statistics & data analysis* 55.3 (2011), pp. 1196–1214.

(166) Feifei Zheng, Seth Westra, and Scott A Sisson. "Quantifying the dependence between extreme rainfall and storm surge in the coastal zone". In: *Journal of hydrology* 505 (2013), pp. 172–187.

(167) Matthias Zink et al. "A high-resolution dataset of water fluxes and states for Germany accounting for parametric uncertainty". In: *Hydrology and Earth System Sciences* 21.3 (2017), pp. 1769–1790.

(168) Jakob Zscheischler et al. "Future climate risk from compound events". In: *Nature Climate Change* 8.6 (2018), pp. 469–477.