UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Univariate and Bivariate Extremes in Meteorology: an application to the Great Plains Low-Level Jet System

Luis Gimeno Sotelo

Mestrado em Estatística e Investigação Operacional

Especialização em Estatística

Dissertação orientada por:

Professora Doutora Patrícia de Zea Bermudez

2021

Acknowledgements

First of all, I would like to thank my advisor, Prof. Patrícia de Zea Bermudez, for the large amount of time that she devoted to help me in this work. I am very grateful to her for the advice in all aspects of this thesis, from helping me to understand theoretical topics that were new for me, to an in-depth process of text reviewing that I very much appreciate.

I would also like to thank the research group EPhysLab (University of Vigo, Spain) for introducing me to a very interesting meteorological problem and providing me with the data that were used in this thesis. Moreover, I am also very grateful for their help in the meteorological interpretation of the results of this work.

Additionally, I thank the professors of this master's degree for their high teaching quality, with a special mention to Prof. Maria Isabel Fraga Alves. Apart from having learnt a lot from the two curricular units that she taught, I really appreciate that she supplied me with very useful materials to study Extreme Value Theory.

Last but not least, I would like to thank my family for their support during all the years of hard study. They have always helped me in everything I have needed.

This dissertation was elaborated under the project with title “ Data Fusion and Calibration Methods for Spatial Risk Analysis ” with reference: PTDC/MATSTA/28649/2017, supported by FCT, Fundação para a Ciência e Tecnologia, Portugal.

Resumo

Nesta tese, colocamo-nos no contexto do sistema do *Great Plains Low-Level Jet* (GPLLJ), que é um sistema de ventos muito fortes na troposfera inferior que transporta uma enorme quantidade de humidade do Golfo do México para as Grandes Planícies Americanas e está principalmente activo nos meses de Verão.

Este trabalho tem dois objectivos: primeiro, analisar o comportamento extremo da humidade transportada da região de origem do GPLLJ para o domínio do *jet*; segundo, nos casos de humidade transportada baixa e alta, estudar a dependência global e extrema entre a cauda superior da precipitação na região do sumidouro do GPLLJ e a cauda inferior da estabilidade troposférica na região do sumidouro do GPLLJ (ómega).

Para este efeito, são utilizadas as séries de observações diárias de humidade transportada, precipitação e "ómega" de todos os períodos de Junho-Julho-Agosto de 1980 a 2017, o que corresponde a 3496 observações. As observações de precipitação e "ómega" foram separadas em dois grupos consoante os valores observados de humidade transportada. Um valor de humidade transportada é considerado baixo, se for inferior ao correspondente quantil empírico de probabilidade 0.25 e alto, se exceder o quantil de probabilidade 0.75 dessa variável.

No que diz respeito à parte teórica desta dissertação, em primeiro lugar são apresentados os conceitos fundamentais da Teoria Univariada de Valores Extremos. É relevante salientar a importância dos modelos de limiar, que são essenciais para ambos os objectivos da tese: realizar a análise univariada dos extremos de humidade transportada e como passo prévio necessário ao estudo dos extremos bivariados.

A seguir, é possível encontrar alguns dos tópicos-chave da Teoria Bivariada de Valores Extremos, que foi utilizada para abordar a abordagem de extremos que constitui o segundo objectivo desta tese. São apresentadas as noções probabilísticas fundamentais e alguns dos modelos paramétricos mais importantes dos extremos bivariados, para além da metodologia estatística mais comum neste contexto. Particularmente importante é o método de verosimilhança censurada, que é utilizado para ajustar o Modelo de Limiar Bivariado de Excessos na modelação dos dados. Também é abordado o conceito de *independência assintótica*, que é uma situação que deve ser analisada quando se utiliza a metodologia apresentada nesta tese.

A fim de obter uma imagem global da estrutura de dependência no contexto do segundo objectivo deste trabalho são utilizadas cópulas. É apresentado um resumo dos aspectos mais importantes da Teoria de Cópulas, tanto do ponto de vista probabilístico como do ponto de vista estatístico. Nomeadamente, introduzimos o conceito de *copula*, alguns dos modelos de cópulas mais comuns e apresentamos diferentes métodos de estimação (embora o foco seja a abordagem semi-paramétrica, que é usada na parte prática da tese), bem como algumas breves considerações sobre a selecção de modelos e testes de ajustamento das cópulas.

Posteriormente, são apresentados o procedimento e os resultados relativamente à análise univariada de extremos de humidade transportada. Depois de realizar uma breve análise exploratória a fim de compreender melhor a série em estudo, são utilizados modelos de limiar para estudar o comportamento dos valores extremos dessa série. Utiliza-se a abordagem conhecida como *Peaks Over Threshold* (POT), traduzido para português como “Picos acima do limiar”. Através de dois dos métodos mais habituais de selecção de limiares, é decidido que $u = 2$ (mm/dia) é um limiar adequado. Como é claramente visível que os excessos em relação ao limiar escolhido não são independentes, é feito um processo de *declustering* (usando o método de "*run-declustering*") a fim de eliminar o mais possível essa dependência. O *declustering* foi realizado considerando quatro valores diferentes de *run length* (r), nomeadamente 1,2,3 e 4. Graficamente e por testes estatísticos, chegamos à conclusão de que o modelo exponencial é mais apropriado do que o modelo de Pareto Generalizado (GPD, pelo seu acrónimo em inglês) para modelar os máximos de *clusters* de excessos acima do limiar escolhido, para todos os valores de r considerados. Além disso, prova-se que no caso de $r = 4$ o modelo exponencial não estacionário é mais adequado do que o estacionário, no sentido em que se demonstra que o parâmetro de escala do modelo exponencial decresce com o tempo. Por esta razão e porque é o valor que garante melhor a independência entre os excessos, conclui-se que $r = 4$ é a melhor escolha. Também são calculados os níveis de retorno estimados de 38 anos, 50 anos e 100 anos para a série de humidade transportada utilizando o modelo exponencial não estacionário ajustado aos máximos dos *clusters* de excessos. É interessante referir que nesta abordagem o período "um ano" corresponde a "um verão" (meses de Junho, Julho e Agosto). Os resultados desses cálculos mostram que os três níveis de retorno estimados foram diminuindo com o tempo e que a diferença entre eles se tornou menor. Por conseguinte, é possível dizer que esperamos observar valores extremos mais baixos de humidade transportada no futuro.

Por outro lado, são analisados os extremos bivariados de (-ómega,precipitação) nos casos de humidade transportada baixa e alta. Note-se que o sinal de "ómega" é trocado porque, em termos meteorológicos, o interesse é estudar o comportamento conjunto da cauda superior de precipitação e da cauda inferior de "ómega". As séries de precipitação e "-ómega" são desfasadas 1 dia em relação à série de humidade transportada devido à natureza temporal do sistema do GPLLJ. Após uma análise preliminar dos dados em estudo, inicia-se o processo de ajustamento do Modelo de Limiar Bivariado de Excessos. Para tal, é necessário previamente ajustar modelos univariados de limiar às margens. Conclui-se que, tanto nos casos de humidade transportada baixa como alta, um limiar adequado para "-ómega" é $u_1 = 0.03$ (Pa/s) e, para precipitação, é adequado escolher $u_2 = 5.2$ (mm/dia). Usando esses limiares, o modelo GPD é mais apropriado do que o exponencial no caso de "-ómega", verificando-se o contrário no caso da precipitação, tanto nos casos de humidade transportada baixa como alta. Tendo escolhido essas distribuições para os excessos acima do respectivo limiar em cada margem, é utilizado o método de verosimilhança censurada considerando oito diferentes modelos paramétricos. É demonstrado que, para todos esses modelos, a dependência extrema entre "-ómega" e precipitação é mais forte no caso de humidade transportada alta do que quando ela é baixa. Os valores do critério de informação de Akaike (AIC, pelo seu acrónimo em inglês) correspondentes a cada um desses modelos são também calculados e o modelo mais parcimonioso no caso de humidade transportada baixa é o bilogístico, enquanto que no caso de ela ser alta, é o logístico. É apresentada a informação mais relevante sobre o modelo bilogístico ajustado no caso de humidade transportada baixa e o modelo logístico ajustado para o caso de ela ser alta. Apresentam-se as estimativas de máxima verosimilhança dos seus coeficientes, as suas correspondentes funções de dependência de Pickands, bem como algumas curvas de quantis estimadas que foram construídas utilizando estes modelos. Além disso, através dos gráficos destinados a esse fim, chega-se à conclusão de que podemos assumir que as variáveis são assimptoticamente dependentes, e portanto os modelos que são apresentados na parte teórica da tese são apropriados para este par de variáveis, tanto nos casos de humidade transportada baixa como alta.

Por fim, são ajustadas cópulas ao par (-ómega,precipitação), tanto nos casos de humidade transportada baixa como alta. Chega-se à conclusão de que a dependência global entre "-ómega" e precipitação é mais forte no caso de humidade transportada alta do que quando ela é baixa. Além disso, através das estimativas dos coeficientes de dependência de cauda, vemos que a dependência superior de cauda entre "-ómega" e precipitação é mais forte no caso de humidade transportada alta do que quando ela é baixa, resultado que está em consonância com as conclusões obtidas a partir do estudo dos extremos bivariados. Além disso, utilizando cópulas, chega-se à mesma conclusão no que se refere à dependência inferior de cauda. De acordo com os testes de ajustamento realizados às cópulas *t* de Student e Gumbel em cada caso de humidade transportada baixa e alta, esses modelos mostraram ser apropriados em ambos os casos. De acordo com os valores de AIC, a cópula *t* de Student é o modelo mais adequado no caso de humidade transportada baixa e a cópula Gumbel quando ela é alta. Finalmente, usando estas duas cópulas ajustadas, traçaram-se as funções de densidade dos modelos ajustados. A comparação das pseudo-observações com dados simulados a partir das cópulas ajustadas permite-nos pensar que os modelos são adequados.

Palavras-chave: Extremos Univariados, *Declustering*, Extremos Bivariados, Dependência, Cópulas.

Abstract

The Great Plains Low-Level Jet (GPLLJ) system consists of very strong winds in the lower troposphere that transport a huge amount of moisture from the Gulf of Mexico to the American Great Plains and is mainly active during summer. The two main objectives of this thesis are: to study the univariate extremes of the Transported Moisture from the GPLLJ source region to the jet domain; and, in the cases of low and high Transported Moisture, to analyze the global and extremal dependence between the upper tail of the precipitation in the GPLLJ sink region and the lower tail of the tropospheric stability in the GPLLJ sink region (omega). For this purpose, we use the series of daily observations of Transported Moisture, Precipitation and "omega" of the periods June-July-August from 1980 to 2017. Observations of Precipitation and "omega" were separated into two groups according to the observed values of Transported Moisture: a value of Transported Moisture is considered low if it is lower than the corresponding 0.25 empirical probability quantile and high if it exceeds the 0.75 probability quantile of that variable. In order to work on the first objective of the thesis, the fundamental concepts of Univariate Extreme Value Theory are presented. Regarding the second objective, the key topics of Bivariate Extreme Value Theory and Copula Theory are also explained.

With respect to the univariate extremes of Transported Moisture, the *Peaks Over Threshold* (POT) approach is applied. With the aim of dealing with the dependence between the excesses, a declustering scheme is performed. We come to the conclusion that a non-stationary Exponential model is the most appropriate for the cluster maxima of excesses. Using that model, it is shown that the estimated 38-year, 50-year and 100-year return levels of the Transported Moisture series decrease over the period considered.

Taking into account that, when dealing with bivariate extremes, interest focuses in the lower tail of "omega", we studied the extremal behaviour of the pair (-omega,precipitation) in the cases of low and high Transported Moisture. Eight different parametric models are used for fitting the so-called Bivariate Threshold Excess Model. For all those models, we were able to understand that the extremal dependence between "-omega" and precipitation is stronger in the case of high Transported Moisture.

The same result is obtained using copulas. Additionally, it is shown that both the global and the lower tail dependence are also stronger when the Transported Moisture is high.

Keywords: Univariate Extremes, Declustering, Bivariate Extremes, Dependence, Copulas

Contents

1 Introduction 1

1.1 The GPLLJ System 1

1.2 Data 3

2 Basic Concepts of Univariate Extreme Value Theory 5

2.1 Asymptotic Models for Maxima 5

2.1.1 The Generalized Extreme Value distribution 5

2.1.2 The Block Maxima approach 7

2.2 Threshold Models 9

2.2.1 The Generalized Pareto Distribution and the POT approach 9

2.2.2 Threshold Selection 10

2.2.3 Model Assessment 11

2.2.4 Dealing with Dependent Sequences 13

3 Bivariate Extreme Value Theory 14

3.1 Asymptotic Characterization of Componentwise Maxima 14

3.2 Parametric Models 17

3.3 Statistical Modelling of Componentwise Block Maxima 18

3.4 Excesses Over A Threshold 19

3.5 Asymptotic Independence and Asymptotic Dependence 21

4 Copulas 24

4.1 The Concept of Copula 24

4.2 Copula Classes 25

4.3 Estimation 27

4.4 Model Selection and Goodness-of-Fit 28

5 Procedure and Results 29

5.1 Univariate analysis of Transported Moisture 29

5.1.1 A brief exploratory analysis 29

5.1.2 Threshold Models Approach 30

5.2 Bivariate analysis of Precipitation and "-omega" 40

5.2.1 Preliminary analysis 40

5.2.2 Fitting the Bivariate Threshold Excess Model 41

5.3 Copula analysis of Precipitation and "-omega" 53

6 Comments, Conclusions and Future Work **61**

References **64**

Appendices **69**

A Goodness-of-fit Tests Tables **70**

B Scripts **71**

B.1 Univariate extremes 71

B.2 Bivariate extremes and copulas 81

B.2.1 Theoretical example of Pickands Dependence Functions 81

B.2.2 Practical part 82

List of Figures

1.1 Climatology of the Great Plains Low-Level Jet System for June, July and August. 2

1.2 Key regions associated to the GPLLJ 3

3.1 Pickands dependence function for the logistic model with $\alpha = 0.9$, $\alpha = 0.7$ and $\alpha = 0.5$. 19

5.1 Plot of the series of TM from the GPLLJ source region to the jet domain 29

5.2 Histogram and kernel density estimate for the TM series 30

5.3 Boxplots for the observations of the first and second half of the TM series (Years 1980-1998 and Years 1999-2017); and boxplots for the observations of each summer from 1980 to 2017 for the TM series. 31

5.4 Estimated Mean Excess Function of Transported Moisture from the GPLLJ source region to the jet domain 31

5.5 Maximum Likelihood estimates for the shape and modified scale parameters of the GPD models fitted to the Transported Moisture series, as a function of the chosen threshold u . 32

5.6 Exceedances of the TM series above the threshold $u = 2$. 33

5.7 Cluster maxima of excesses of the TM series above $u = 2$, performing declustering with *run length* (r) equal to 1,2,3 and 4. 34

5.8 *Profile log-likelihood* function for the shape parameter of the GPD model fitted to the cluster maxima of excesses of the TM series, considering $u = 2$ and $r = 4$ 35

5.9 Exponential and GPD QQ-Plots fitted for the cluster maxima of excesses of the TM series, for threshold $u = 2$ and *run length* $r = 4$. 35

5.10 Return level plots for the stationary Exponential model for the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and *run length* (r) equal to 1,2,3 and 4. 38

5.11 Estimated return levels for the TM series, considering a non-stationary Exponential model for the cluster maxima of excesses above the threshold $u = 2$, having performed declustering with *run length* (r) equal to 4. 39

5.12 Plots of the "-omega" and precipitation series 41

5.13 Boxplots of "-omega" and precipitation for low TM and high TM 42

5.14 Histograms and kernel density estimates of "-omega" and precipitation for low TM and high TM 43

5.15 Estimated Mean Excess Function of "-omega" for low TM and high TM 44

5.16 Maximum Likelihood estimates for the shape and modified scale parameters of the GPD models fitted to the "-omega" series (for the cases of low TM and high TM), as a function of the chosen threshold u 45

5.17 Estimated Mean Excess Function of precipitation for low TM and high TM 46

5.18 Maximum Likelihood estimates for the shape and modified scale parameters of the GPD models fitted to the precipitation series (for low and high TM), as a function of the chosen threshold u. 47

5.19 *Profile log-likelihood* functions for the shape parameter of the GPD model fitted to the "-omega" and precipitation series in the cases of low TM and high TM, considering $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation. 48

5.20 Exponential and GPD QQ-Plots fitted for the excesses above the threshold $u_1 = 0.03$ for "-omega" in both the cases of low TM and high TM. 49

5.21 Exponential and GPD QQ-Plots fitted for the excesses above the threshold $u_2 = 5.2$ for precipitation in both the cases of low TM and high TM. 50

5.22 Scatterplots for (-omega,precipitation) in the cases of low TM and high TM. 56

5.23 *Chi plots* and *chi bar plots* for (-omega,precipitation) in the cases of low TM and high TM. 57

5.24 Pickands dependence functions corresponding to the fitted Bilogistic model for (-omega,precipitation) in the case of low TM ; and corresponding to the fitted Logistic model for (-omega,precipitation) in the case of high TM 58

5.25 Estimated quantile curves at lower tail probabilities $p = 0.95, 0.975, 0.99$ for the joint distribution function of (-omega,precipitation) in the cases of low and high TM, using the fitted Bilogistic model in the case of low TM and the Logistic model in the case of high TM. 58

5.26 Pseudo-observations for (-omega,precipitation) in the cases of low and high TM. 59

5.27 Copula density function for the most parsimonious fitted copula model for (-omega,precipitation) in each case of low and high TM. 59

5.28 Pseudo-observations and simulated data from the most parsimonious fitted copula model for (-omega,precipitation) in each case of low and high TM. 60

List of Tables

4.1 Tail dependence for some of the most used copulas. 27

5.1 Summary statistics for the TM series 30

5.2 Results of the POT analysis with declustering for the TM series considering the threshold $u = 2$, regarding the number of clusters obtained, the estimate of the extremal index and the ML estimates for the parameters of the GPD and the Exponential model, with their corresponding standard errors. 33

5.3 R^2 of the linear model fitted to the Exponential and to the GPD QQ-Plots for the cluster maxima of excesses the TM series, taking $u = 2$ and declustering *run length* of $r =$ 1,2,3 and 4. 34

5.4 Observed value of the CvM, AD, LRT and LcKS statistics and corresponding p-values for the TM series, with $u = 2$ and declustering *run length* equal to 1, 2, 3 and 4 36

5.5 Maximum Likelihood estimates (with their corresponding standard errors) for the parameters of the non-stationary Exponential model fitted to the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and performing declustering with *run length* equal to 1, 2, 3 and 4. 37

5.6 Observed value of the LRT statistic and corresponding p-value. This LRT is used for comparing the stationary Exponential model and the non-stationary one, fitted to the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and performing declustering with *run length* (r) equal to 1, 2, 3 and 4, respectively for each column. . . . 37

5.7 Estimated 38-year, 50-year and 100-year return levels for the TM series in the first and the last summer of the period considered, using a non-stationary Exponential model for the cluster maxima of excesses above the threshold $u = 2$, having performed declustering with *run length* (r) equal to 4. 39

5.8 Summary statistics for the "-omega" and precipitation series. 41

5.9 Standard deviation, skewness and Pearson's measure of kurtosis for the "-omega" and precipitation series in the cases of low TM and high TM. 42

5.10 Results of the POT analysis for the "-omega" and precipitation series in the cases of low TM and high TM, considering thresholds $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation. 44

5.11 R^2 of the linear models associated with the Exponential and GPD QQ-Plots for the excesses above the threshold $u_1 = 0.03$ for "-omega" and the excesses above the threshold $u_2 = 5.2$ for precipitation, for low TM and high TM. 46

5.12 Observed values of the CvM, AD and LRT statistics and corresponding p-values for the "-omega" and precipitation series, for the thresholds $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation. 49

5.13 ML estimate for the scale parameter of the Exponential distribution and observed value of the LcKS statistic (with the approximate p-value of that test) for the precipitation series (in the cases of low and high TM), choosing threshold $u_2 = 5.2$ in both cases. . . . 50

5.14 AIC and *Dependence* coefficient for the parametric models that we fitted to model the joint distribution function of (-omega,precipitation) in the cases of low and high TM within the region $x_1 > 0.03$, $x_2 > 5.2$ in both cases. 52

5.15 ML estimates (with their standard errors) for the marginal and dependence parameters of the bilogistic model for (-omega,precipitation) in the case of low TM; and the logistic model for (-omega,precipitation) in the case of high TM. 52

5.16 Information about the Student-t and the Gumbel copulas fitted to the pair (-omega,precipitation) in the cases of low and high TM. 54

A.1 Simulated quantiles of asymptotic level p for the Cramér-von Mises Statistic (GPD with both parameters unknown) 70

A.2 Simulated quantiles of asymptotic level p for the Anderson-Darling Statistic (GPD with both parameters unknown) 70

List of Acronyms

WCRP World Climate Research Programme

IPCC Intergovernmental Panel on Climate Change

LLJ Low-Level Jet

ARs Atmospheric Rivers

GPLLJ Great Plains Low-Level Jet

TM Transported Moisture

CPC Climate Prediction Center

i.i.d. independent identically distributed

GEV Generalized Extreme Value (distribution)

EVI Extreme Value Index

POT Peaks Over Threshold

GPD Generalized Pareto Distribution

ML Maximum Likelihood

PWM Probability Weighted Moments

MEF Mean Excess Function

CvM Cramér-von Mises

AD Anderson-Darling

LRT Likelihood Ratio Test

LcKS Lilliefors-corrected Kolmogorov-Smirnov

IFM *Inference Functions for Margins*

MPLE Maximum Pseudo-Likelihood Estimator

AIC Akaike's Information Criterion

AA Azores Anticyclone

1 | Introduction

1.1 The GPLLJ System

The World Climate Research Programme (WCRP), which is the United Nations programme defining climate research priorities, identifies Weather and Climate extremes as one of the big challenges: it has been included as an independent chapter in all Intergovernmental Panel on Climate Change (IPCC) reports (e.g. Field et al. (2012); Qin et al. (2014); Masson-Delmotte et al. (2018)). Within the study of these extremes, the analysis of combined events, defined as "the combination of multiple climate drivers that contributes to societal or environmental risk", has gained great importance, being multiple the publications devoted to them in high-impact journals due to their enormous socioeconomic importance (e.g. Raymond et al. (2020); Ridder et al. (2020); Zscheischler et al. (2020)). Initially focused on the analysis of the simultaneous or consecutive occurrence of local phenomena, such as droughts and heat waves, the studies involving precipitation as one of the variables have been abundant. However, studies trying to link precipitation extremes to large-scale atmospheric circulation patterns have been much less frequent and, up to our knowledge, the role of the large-scale moisture transport has never been considered from this perspective.

Moisture transport from oceans to continents is the primary component of the atmospheric branch of the water cycle and forms the link between evaporation from the ocean and precipitation over the continents (Gimeno et al., 2012). There has been an important number of studies on the role of anomalies in the transport of moisture during natural hydrometeorological hazards, extreme drought (e.g., Drumond et al. (2019)) or intense precipitation (e.g. Stohl and James (2004)). The close relation between moisture transport and extreme precipitation events is maximized when this is studied in the areas of influence of the two major global mechanisms of atmospheric moisture transport, namely Low-Level Jet (LLJ) systems and Atmospheric Rivers (ARs), two large-scale dynamical/meteorological structures, the former being key in tropical and subtropical regions and the latter in extratropical regions (Gimeno et al., 2016).

A LLJ is a system of very strong winds in the lower troposphere, typically in the first 1000 meters height (Stensrud, 1996). As water vapour is mainly confined in the lower troposphere, LLJs are major mechanisms of moisture transport at planetary scale. When LLJs are active, they transport a huge amount of moisture favoring high precipitation in the downwind regions. In contrast, in periods when LLJs are absent, downwind regions can suffer from drought events (Gimeno et al., 2016). Within these systems, the Great Plains Low-Level Jet (GPLLJ) is the most studied one because of its socioeconomic effects. It transports a huge amount of moisture from the Gulf of Mexico to the American Great Plains and it is mainly active in the summer (Burrows et al., 2019). Broadly speaking, the GPLLJ carries one-third of all water vapour entering continental United States (Helfand and Schubert, 1995), and it is associated with 10 % – 45 % of the summer precipitation of the American Great Plains region (Hodges and Pu, 2019). In Figure 1.1 it is possible to see the climatology of the Great Plains Low-Level Jet System for the months of June, July and August.

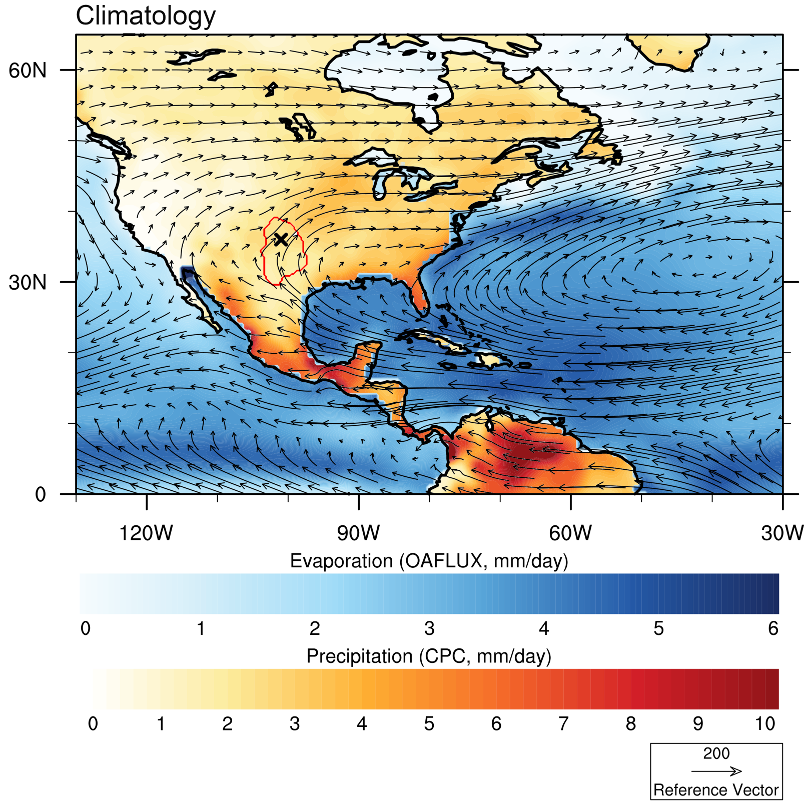


Figure 1.1: **Climatology of the Great Plains Low-Level Jet System for June, July and August.** The region with the highest occurrence of LLJs is inside the red curve, with the cross indicating the point at which the proportion of days on which the LLJ occurs is the highest one. Bluish colors represent the evaporation (mm/day; data from OAFLUX), reddish colors indicate the precipitation (mm/day, data from CPC) and the arrows symbolize the flux of moisture at each point of the grid under consideration ($\mathrm{Kg\,m^{-1}\,s^{-1}}$, data from ERA5). Figure courtesy of Dr. Iago Algarra (University of Vigo, Spain).

The economic importance of the GPLLJ is enormous in the sense that it determines the average and extreme precipitation of a large agricultural region, whose production depends on precipitation, occurring large losses from floods and droughts (Basara et al., 2013). It is also important in the determination of the wind resource and especially in the damage generated by severe weather, as GPLLJ is closely related to the development of mesoscale convective systems (Chen et al., 1998) and they are associated with heavy precipitation, supercelular storms and tornado development (Weaver et al., 2012).

The GPLLJ affects precipitation by increasing its frequency, modifying its spatial distribution and increasing its intensity (Pitchford and London, 1962; Mo et al., 1995; Walters and Winkler, 2001; Schumacher and Johnson, 2009; Squitieri and Gallus, 2016; Squitieri and Gallus Jr, 2016). The underlying mechanism to the relationship between the GPLLJ and the precipitation is a strong moisture and heat transport at low levels from the Gulf of Mexico. Moreover, wind convergence at low levels implies atmospheric instability in the output area of the GPLLJ, favoring upward movement. Therefore, it is evident that transported moisture and atmospheric instability are two factors that play an important role in precipitation.

Hence, after carrying out an univariate extremal analysis of the moisture transported by the GPLLJ, we will apply bivariate Extreme Value Theory in order to jointly analyze the extremes of precipitation and "-omega" (to be next characterized) in the context of the GPLLJ system. These bivariate extremes will be studied in two scenarios: when the transported moisture is low and when it is high. In these situations, we will also resort to copula models to describe the global dependence between the variables. We should stress the fact the term “global” refers to the set of pair (-omega,precipitation) considering the entire sample of low values of transported moisture (25% of the lowest values). The same term is applied for the case of high values of transported moisture (25% of the highest values).

1.2 Data

In a recent paper (Algarra et al., 2019) , a state-of-the-art Lagrangian approach is used in order to identify the main moisture sources and sinks associated to the GPLLJ (Figure 1.2).

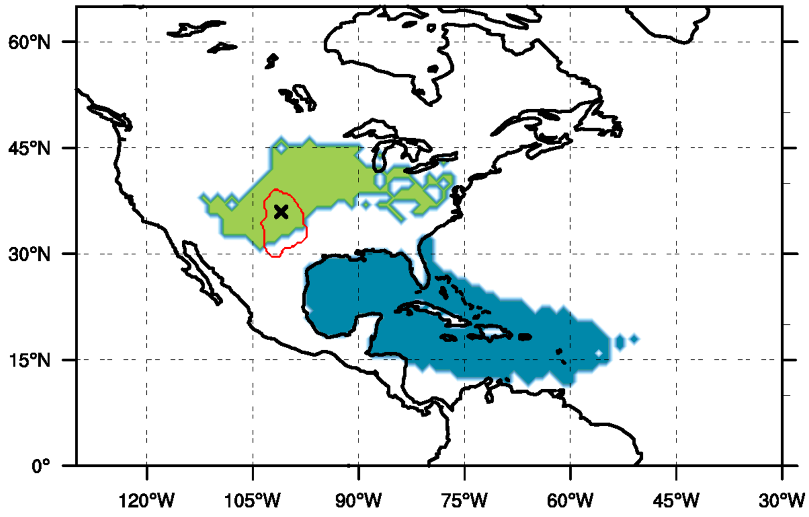


Figure 1.2: **Key regions associated to the GPLLJ:** Region with the highest occurrence of LLJs (inside the red curve, with the cross indicating the point at which the proportion of days on which the LLJ occurs is the highest one); the GPLLJ major oceanic moisture source region (in blue) and its major moisture sink region (in green). Figure courtesy of Dr. Iago Algarra (University of Vigo, Spain).

The area inside the red curve is the jet domain, that is, it is the region with the highest occurrence of LLJs during the period May-October, being the cross the geographical point at which they occur most frequently (36°N, 101°W, 500m height); the area in blue identifies the major oceanic source region for the moisture reaching the jet domain; and the area in green corresponds to the main sink of that moisture, once it has been transported by the jet. So, there are two regions of interest in our analysis: the moisture source and sink regions, connected by the GPLLJ structure in a temporal domain of several days from the evaporation in the source to the precipitation in the sink.

Therefore, the series to analyze, based on the sources and sink areas of moisture linked to the GPLLJ, are:

1. **Transported Moisture (TM)** from the GPLLJ source region to the jet domain (mm/day), as calculated in Algarra et al. (2019). In this study, a Lagrangian approach was used to track air parcels reaching the jet domain from the source region. The TM is then computed by adding the moisture gains of the parcels in the source region before arriving at the jet domain.

2. **Precipitation** in the GPLLJ sink region (mm/day): daily series of precipitation integrated in the whole moisture sink region of the GPLLJ taken from the Climate Prediction Center (CPC) dataset (Xie et al., 2010), which is a state-of-the-art precipitation dataset (see Sun et al. (2018) for a review on gridded precipitation data).

3. Tropospheric Stability in the GPLLJ sink region (**omega**, measured in Pa/s): daily series of vertical velocity computed as the mean of omega at 850 hPa in the sink region, taken from the reanalysis ERA-5 (Hersbach et al., 2020). Omega is defined as the vertical component of velocity in pressure coordinates (these three-dimensional coordinates are defined by replacing the usual z-coordinate by atmospheric pressure (p)). This is, $\omega := \frac{dp}{dt}$, so negative values of ω represent ascending movements and positive values correspond to descending movements. The level of 850 hPa (about 1500 m height) is considered for ω as it represents the vertical movement at the lower troposphere, where the GPLLJ occurs and most of the moisture is confined.

The series consist of 6992 observations daily recorded from 1 May 1980 to 31 October 2017. This period comprises the extended summer periods since the inclusion of satellite data in the reanalysis, which occurred in 1979 [1] . However, in the statistical analysis to be done in this thesis, we will only use the summer months, that is, the June-July-August periods, because they are the most interesting ones meteorologically speaking (in these months, the GPLLJ is more active, with occurrence close to 70% of the days). Therefore, we will initiate our study with series that have 3496 observations, although in fact we have 874 observations in each of the groups of TM.

[1] *Extended summer* refers to the period of May to October

2 | Basic Concepts of Univariate Extreme Value Theory

This chapter introduces important concepts of univariate Extreme Value Theory that we will keep in mind throughout the practical part of this thesis and are essential to understand before starting to study the bivariate case. Extensive information about the topics presented in this chapter can be found in Coles (2001) and Beirlant et al. (2004).

2.1 Asymptotic Models for Maxima

2.1.1 The Generalized Extreme Value distribution

Let $X_1, X_2, ...$ be independent identically distributed (i.i.d.) random variables, each with distribution function F (density function f), and consider $M_n = \max\{X_1, X_2, ..., X_n\}$. The distribution of M_n can be obtained in an exact way from F (taking into account the independence and identical distribution property):

$$F_{M_n}(x) = P(M_n \leq x) = P(X_1 \leq x, ..., X_n \leq x) = P(X_1 \leq x) \times ... \times P(X_n \leq x) = (F(x))^n \tag{2.1}$$

and its density function is obtained by differentiation:

$$f_{M_n}(x) = n\,(F(x))^{n-1} f(x). \tag{2.2}$$

When $n \to \infty$, the previously calculated distribution function converges to 0 in case $F(x) < 1$ and to 1 in case $F(x) = 1$. This is, M_n converges in distribution to $x^F := \sup\{x : F(x) < 1\}$, which is is the right endpoint of the distribution F (there is also almost-sure convergence and convergence in probability). Therefore, in order to obtain a non-degenerate limiting distribution, it is necessary to carry out a normalization. It consists of looking for sequences of constants $\{b_n; n \geq 1\}$ and $\{a_n; n \geq 1\}$ $(a_n > 0)$ such that the distribution of

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{2.3}$$

converges to a non-degenerate distribution when $n \to \infty$, this is,

$$\lim_{n \to \infty} F^n(a_n x + b_n) = G(x). \tag{2.4}$$

Definition 2.1.1 *If there exist sequences of constants $\{a_n; n \geq 1\}$ $(a_n > 0)$ and $\{b_n; n \geq 1\}$ such that 2.4 is verified, then it is said that F belongs to the* ***max-domain of attraction*** *of the distribution G. This is denoted by: $F \in D_M(G)$.*

An important concept is that of *max-stability*:

Definition 2.1.2 *A distribution G is said to be* ***max-stable*** *if, for every $n \in \mathbb{N}$, there are constants $\alpha_n > 0$ and β_n such that:*

$$G^n(\alpha_n x + \beta_n) = G(x) \tag{2.5}$$

In simple words, max-stability is a property that is satisfied by distributions that are identical to the distribution of the sample maximum (for any sample size), apart from possible changes in location and scale. If a limiting distribution for M_n^* exists, that distribution has to be max-stable.

The complete range of limiting distributions that M_n^* may follow is given by the Extremal Types Theorem:

Theorem 2.1.3 (Extremal Types Theorem: Fisher and Tippett (1928)) *If there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that $P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x)$, when $n \to \infty$, with G a non-degenerate distribution function, then G must belong to one of the following families:*

$$I \ \textbf{\textit{Gumbel}}\text{: } G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, \ -\infty < x < \infty$$

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

$$III \ \textbf{\textit{Weibull}}\text{: } G(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{\alpha}\right]\right\}, & x < b, \\ 1, & x \geq b; \end{cases}$$

for parameters $a > 0, b$ and, in the case of families II and III, $\alpha > 0$.

Fréchet, Gumbel and Weibull distributions are known as extreme value distributions and they are the only distributions to which M_n^* can converge. Some common distributions that are in the Fréchet domain of attraction are the Pareto, Loggamma, Student-t and Burr distributions; meanwhile, the Exponential, Gamma, Normal and Lognormal distributions are in the Gumbel domain of attraction; and the Uniform and Beta distributions are in the Weibull domain of attraction.

The three families of extreme value distributions can be combined into a single family, known as **Generalized Extreme Value distribution (GEV)**, with distribution function:

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}\right\}, \tag{2.6}$$

defined on the set $\left\{x : 1 + \frac{\xi\,(x-\mu)}{\sigma} > 0\right\}$. The location, scale and shape parameters satisfy, respectively, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

The shape parameter (ξ) is known as the Extreme Value Index (EVI) and its value determines the family of distributions to which the GEV family corresponds:

- If $\xi < 0$, it corresponds to the Weibull family
- If $\xi = 0$ (limit of 2.6 as $\xi \to 0$), it corresponds to the Gumbel family

- If $\xi > 0$, it corresponds to the Fréchet family

Therefore, Theorem 2.1.3 can be reformulated as follows:

Theorem 2.1.4 (Unified Extremal Types Theorem: Gnedenko (1943)) *If there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that $P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x)$ when $n \to \infty$, with G a non-degenerate distribution function, then G belongs to the GEV family (2.6).*

We will now present an important definition:

Definition 2.1.5 *The* ***right tail*** *of a distribution function F is defined as $\bar{F}(x) := P(X > x) = 1 - F(x)$*

The specification of the EVI determines the behaviour of the tail of the distribution F, since it indicates the speed of decay to 0 of $\bar{F}(x)$ as x approaches the right endpoint x^F:

- If $\xi < 0$, its tail is lighter than an exponential tail and x^F is finite.
- If $\xi = 0$ (limit of 2.6 as $\xi \to 0$), its tail is an exponential tail and x^F is possible to be finite or infinite.
- If $\xi > 0$, its tail is heavier than an exponential tail and x^F is infinite.

Regarding the parameter estimation, the most usual methods to estimate parametrically the parameters of the GEV distribution are Maximum Likelihood (ML) and Probability Weighted Moments (PWM) methods. As for the confidence intervals for the parameters, apart from the asymptotic confidence intervals resulting from the approximation of the ML or PWM estimators to the normal distribution, it is possible to obtain confidence intervals of better quality (although not necessarily centred on the point estimate) using the *profile log-likelihood* function, which is, for each value of the parameter under study, the log-likelihood function, $\log L(.)$, maximized relatively to the other parameters. For example, the *profile log-likelihood* function for ξ is:

$$\log L_p(\xi) := \max_{\mu,\sigma|\xi} \log L(\xi, \mu, \sigma) \tag{2.7}$$

and the $100(1-\alpha)\%$ confidence interval based on that function is given by:

$$CI_\xi = \left\{ \xi : \log L_p(\xi) \geq \log L_p(\hat{\xi}) - \frac{\chi_1^2(1-\alpha)}{2} \right\}, \tag{2.8}$$

where $\hat{\xi}$ is the ML estimate for ξ and $\chi_1^2(.)$ is the inverse distribution function of a χ^2 distribution with 1 degree of freedom.

Analogously, *profile log-likelihood* confidence intervals can be obtained for μ and σ.

For further details about parameter estimation for the GEV family, see Beirlant et al. (2004).

2.1.2 The Block Maxima approach

The GEV family is useful to model the distribution of **block maxima**. The procedure consists of grouping the observations into blocks of equal size and, then, fitting the GEV distribution to the set of the maxima of each of the blocks.

As it is explained in Coles (2001), the main problem with this method lies in the choice of block size, for which a trade-off between bias and variance of the estimators of the model parameters must be found:

the choice of very small blocks leads to a poor approximation of the model, thus increasing the bias in estimation and extrapolation. Conversely, choosing blocks that are too large increases the variance of the estimates. Moreover, some phenomena do not not have a natural time structure, which is an additional problem.

For monthly or daily time series along many years of observation, it is common practice to use blocks of annual length. For example, thinking in m years of daily observations of a phenomenon: for a given year i, we assume independent random variables $X_{i1}, X_{i2}, ..., X_{i365}$ (independence is often unrealistic) and consider $Y_i = max\{X_{i1}, X_{i2}, ..., X_{i365}\}$. Therefore, $Y_1, Y_2, ...Y_m$ is considered as a random sample of $Y = max\{X_1, X_2, ..., X_{365}\}$ and the distribution of Y is approximated by a GEV distribution (even if there is short-range temporal dependence, the assumption that the $Y_1, Y_2, ...Y_m$ are independent is likely to be reasonable (Coles, 2001)).

The estimates of the parameters of the GEV distribution may be used to estimate interesting indicators such as:

- **Exceedance Probability**: It is simply the probability that Y is greater than a predefined high value q, this is, $P(Y > q)$. It can be estimated by: $\widehat{P(Y > q)} = 1 - G_{\hat{\xi}}(q|\hat{\mu}, \hat{\sigma})$, where $G_{\hat{\xi}}(.|\hat{\mu}, \hat{\sigma})$ is the distribution function of the GEV distribution with estimated parameters $\hat{\xi}$, $\hat{\mu}$ and $\hat{\sigma}$.

- **Return Level**: For a given value t, a return level $U(t)$ is defined as $P(Y > U(t)) = \frac{1}{t}$. For instance, consider Y as the annual maximum of a phenomenon. The 50-year return level, this is, $U(50)$, is such that, on average, Y is greater than that quantity once every 50 years. $U(t)$ can be estimated by using the inverse of the distribution function of the estimated GEV model $\left(G_{\hat{\xi}}^{\leftarrow}(y) = \inf\left\{x : G_{\hat{\xi}}(x) \geq y\right\}\right)$, in the way presented as follows: $\widehat{U(t)} = G_{\hat{\xi}}^{\leftarrow}(1 - \frac{1}{t}|\hat{\mu}, \hat{\sigma})$.

- **Return Period**: For a given value q, the return period (denoted by $T(q)$) is the average number of blocks before a higher value than q occurs, this is, $T(q) := \frac{1}{P(Y > q)}$. Also considering Y as the annual maximum of a phenomenon, Y is, on average, greater than q once every $T(q)$ years. It can be estimated by: $\widehat{T(q)} = \frac{1}{\widehat{P(Y > q)}} = \frac{1}{1 - G_{\hat{\xi}}(q|\hat{\mu}, \hat{\sigma})}$. The concept of return period is closely related to the concept of return level, and there exists the following relationship between them: $T(U(t)) = t$.

- **Extremal Quantile of probability p**: It is denoted by χ_p and it is simply the value that is exceeded by Y with probability p (p is usually a very small value). This is, $\chi_p := G_{\xi}^{\leftarrow}(1 - p|\mu, \sigma)$, which can obviously be estimated by: $\widehat{\chi_p} := G_{\hat{\xi}}^{\leftarrow}(1 - p|\hat{\mu}, \hat{\sigma})$.

- **Right endpoint of Y**: If the shape parameter of the GEV distribution is negative, it is possible to estimate $y^* := \sup\left\{y : G_{\xi}(y|\mu, \sigma) < 1\right\}$ by $\widehat{y^*} = \widehat{\chi_0} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}$.

It is possible to estimate analogous indicators for the distribution of the underlying population $X \frown F$ by simply taking into account the relationship $F_{M_n} = F^n \approx G_{\xi}$, as it is explained in Gomes et al. (2013).

2.2 Threshold Models

2.2.1 The Generalized Pareto Distribution and the POT approach

Since using only block maxima is wasteful if other extreme-value data is available, the **Peaks Over Threshold (POT)** approach consists of fitting an asymptotic model to the excesses above a high (enough) threshold u. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each having distribution function F, then the random variable $Y = X - u | X > u$ represents the excesses of X above u. It is straightforward that the distribution function of Y is:

$$F_u(y) = P(X - u \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \qquad 0 < y \leq x^F - u\,, \tag{2.9}$$

where $x^F := \sup\{x : F(x) < 1\}$

If F were known, F_u would also be known. However, since this does not occur in practice, the **Generalized Pareto Distribution (GPD)** is used as an approximation of F_u (as long as u is high enough). The distribution function of the GPD is:

$$H_\xi(y|\sigma_u) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{\frac{-1}{\xi}}, & y \in (0, \infty), \xi > 0, \\ 1 - \exp\left(-\frac{y}{\sigma_u}\right), & y \in (0, \infty), \xi = 0, \\ 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{\frac{-1}{\xi}}, & y \in \left(0, -\frac{\sigma_u}{\xi}\right), \xi < 0; \end{cases} \tag{2.10}$$

The scale and shape parameters satisfy, respectively, $\sigma_u > 0$ and $-\infty < \xi < \infty$. σ_u is used to indicate that the scale parameter depends on the threshold u. It is important to remark that, for $\xi = 0$, $\xi > 0$ and $\xi < 0$, the GPD corresponds to the Exponential, Pareto and Beta distributions, respectively.

The approximation of F_u to the GPD is determined by the following theorem (Pickands, 1975; Balkema and De Haan, 1974), which also establishes a duality between the GEV and the GPD:

Theorem 2.2.1 (Pickands-Balkema-de Haan) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each having distribution function F, and consider $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Then, (i) and (ii) are equivalent, where:*

(i) : *F belongs to the max-domain of attraction of a GEV distribution with shape parameter ξ*

(ii) : $\sup\limits_{0<y<x^F - u} |F_u(y) - H_\xi(y|\sigma_u)|$ *converges to 0 when $u \to x^F$*

This theorem implies that, if the distribution of block maxima can be approximated by a GEV, then threshold excesses follow approximately a GPD. It is worth noting that the shape parameter ξ of the GPD is the same as the shape parameter ξ of the associated GEV. Thereby, for the GPD, ξ is as important in determining the tail behaviour as it is for the GEV, in the sense that:

- If $\xi < 0$, F_u has a light tail with finite right endpoint given by $u - \frac{\sigma_u}{\xi}$.
- If $\xi = 0$, F_u has an exponential tail.
- If $\xi > 0$, F_u has a heavy tail with infinite upper limit.

With regard to the parameter estimation for the GPD, the methods to obtain point estimates and confidence intervals are analogous to those that are used for the GEV distribution (they were mentioned in the previous section). As it was said before, in Beirlant et al. (2004) it is possible to find extensive information about this topic, and see also de Zea Bermudez and Kotz (2010a) and de Zea Bermudez and Kotz (2010b) for even more details.

As we did in the case of the GEV distribution, it is possible to estimate interesting quantities using the estimates of the parameters of the GPD. Let N_u be the number of observations over the threshold u and consider n as the number of observations of the original sample. We will now present the indicators corresponding to the population $X \frown F$ (Only the results will be presented. The details of the deductions can be found in Gomes et al. (2013)):

- Exceedance Probability:

$$\widehat{\overline{F}(x)} := \widehat{P(X > x)} = \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\widehat{\sigma_u}}\right)^{-1/\hat{\xi}} \tag{2.11}$$

- Extremal Quantile of probability p :

$$\widehat{F^{\leftarrow}(1-p)} = u + \frac{\widehat{\sigma_u}}{\hat{\xi}}\left(\left(\frac{np}{N_u}\right)^{-\hat{\xi}} - 1\right) \tag{2.12}$$

- Right endpoint of X (in the case of the shape parameter of the GPD being negative):

$$\widehat{x^F} = u - \frac{\widehat{\sigma_u}}{\hat{\xi}} \tag{2.13}$$

2.2.2 Threshold Selection

A very important question in the POT approach is how to choose the threshold u. The problem is in selecting a value that allows a trade-off between the large variance of the estimators that occurs for too high values of u and the large bias that occurs for too small values of this threshold. In this thesis two methods for threshold selection are presented, although it is important to remark that this problem is an ongoing research topic and there is not a solution that is globally satisfactory.

First method: Mean Excess Function

The Mean Excess Function (MEF) is defined as:

$$e(u) := E\left[X - u | X > u\right], \quad if \quad E\left[X\right] < \infty \tag{2.14}$$

Let $x_1, x_2, ..., x_n$ be the observed sample and $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$ be the ordered sample. Then, the empirical counterpart of the MEF is:

$$\hat{e}_n(u) := \frac{\sum_{i=1}^{\infty} x_i \mathbb{I}_{(u,\infty)}(x_i)}{\sum_{i=1}^{\infty} \mathbb{I}_{(u,\infty)}(x_i)} - u, \quad with \quad \mathbb{I}_{(u,\infty)} = \begin{cases} 1, & if \quad x_i \in (u,\infty) \\ 0, & if \quad x_i \in (-\infty, u] \end{cases} \tag{2.15}$$

The sample mean excess plot is frequently plotted in order to choose an adequate threshold u. For each $u = x_{n-k:n}$, where $x_{n-k:n}$ denotes the $(k+1)^{th}$ largest observation, $\hat{e}_n(u)$ may be written as:

$$\hat{e}_n(x_{n-k:n}) = \frac{\sum_{j=1}^{k} x_{n-j+1:n}}{k} - x_{n-k:n} \tag{2.16}$$

Assuming that $X-u|X>u$ follows a GPD, it is possible to write the MEF as follows:

$$e(u) = E\left[X-u|X>u\right] = E\left[Y|Y>0\right] = \frac{\sigma_u+\xi u}{1-\xi}, \quad if \quad \xi<1 \tag{2.17}$$

this is, the plot of $\hat{e}_n(u)$ against u should be linear (and the line should have intercept $\frac{\sigma_u}{1-\xi}$ and slope $\frac{\xi}{1-\xi}$). Thus, the method proposed by Davison and Smith (1990) consists on identifying a point on the plot for which it is possible to see a reasonable linear pattern to its right, corresponding to an appropriate threshold u.

Second method: Stability of the parameter estimates

According to the characterization of the GPD, if a GPD(ξ,σ_{u_0}) is adequate to model the excesses over a threshold u_0, the excesses over a threshold u that is higher than u_0 would also follow a GPD, with the same shape parameter, while the scale parameter satisfies the following relationship:

$$\sigma_u = \sigma_{u_0} + \xi(u-u_0) \tag{2.18}$$

As we can see from the expression above, if $\xi \neq 0$, the scale parameter changes with u. However, if a modified scale $\sigma^* := \sigma_u - \xi u$ is considered, 2.18 can be rewritten as follows:

$$\sigma^* = \sigma_{u_0} - \xi u_0, \tag{2.19}$$

which is constant with respect to u.

Therefore, the method consists of fitting the GPD to a range of increasing thresholds, producing the plots of the estimates of ξ (shape) and σ^* (modified scale) against u. If u_0 is an appropriate threshold, the estimates of both parameters should be constant above u_0. Of course, as in practice the sample of excesses changes as u increases, we will be looking for estimates of ξ and σ^* which are approximately constant.

2.2.3 Model Assessment

Here we will describe two ways of assessing if the threshold model we fitted is appropriate for our data: the Quantile-Quantile plots (QQ-Plots) and the goodness of fit tests. The QQ-plots are not really a proper assessment tool. Besides being part of the exploratory analysis, they help us to evaluate graphically the suitability of the model. As usual, in order to find statistical evidence that a certain GPD model fits the data, hypothesis tests are necessary, and thus two goodness-of-fit tests for the GPD will be presented.

Preliminary analysis: QQ-Plots

In a QQ-plot, the ascending ordered observations, $(x_{1:n}, x_{2:n}, ..., x_{n:n})$, are plotted against the model quantile function, $Q(p) = F^{\leftarrow}(p)$. The points of a QQ-plot have the form $(F^{\leftarrow}(p_{i:n}), x_{i:n})$, $i = 1,2,...,n$, where $p_{i:n}$ are the plotting positions. In the literature, there are several possible choices of plotting positions; in this work we will consider $p_{i:n} = \frac{i}{n+1}, i = 1,2,...,n$. If a model is reasonable for the observed data, then the corresponding QQ-plot should consist of points that are approximately linear. If the QQ-plot is non-linear, it shows that the data has a heavier or a lower tail than the model considered.

In order to study the distribution of the threshold excesses, the first approach is to construct an Exponential QQ-Plot. Let u be the threshold used and $y_{1:N_u} \leq ... \leq y_{N_u:N_u}$ the corresponding ordered sample of excesses. An Exponential QQ-Plot consists of the points:

$$\left\{ \left(-\log(1-\frac{i}{N_u+1}), y_{i:N_u} \right), i=1,...,N_u \right\} \tag{2.20}$$

Moreover, it is possible to assess graphically the fit of a GPD model to the excesses by using a GPD QQ-Plot. Considering $\hat{H}$ as the estimated GPD model, a GPD QQ-Plot is made up of the points:

$$\left\{ \left(\hat{H}^{\leftarrow} \left(\frac{i}{N_u+1} \right), y_{i:N_u} \right), i=1,...,N_u \right\}, \tag{2.21}$$

where

$$\hat{H}^{\leftarrow}(y) = \begin{cases} u + \frac{\hat{\sigma}}{\hat{\xi}} \left[y^{-\hat{\xi}} - 1 \right], & \hat{\xi} \neq 0, \\ -\hat{\sigma} \log(1-y), & \hat{\xi} = 0; \end{cases} \tag{2.22}$$

Goodness-of-fit Tests

Letting $y_1, y_2, ..., y_{N_u}$ be the excesses over a threshold u, there are several statistical tests that can be used to assess the fit of a GPD to these data. The null hypothesis of these goodness-of-fit tests is $H_0 : y_1, y_2, ..., y_{N_u}$ come from a GPD. In this thesis we will focus our attention to the Cramér-von Mises (CvM) and Anderson-Darling (AD) tests for the case in which both parameters ξ and σ are not known. Their test statistics are as follows:

- Cramér-von Mises Statistic:

$$W^2 = \sum_{i=1}^{N_u} \left(H_{\hat{\xi}}(Y_{i:N_u}|\widehat{\sigma_u}) - \frac{2i-1}{2N_u} \right)^2 + \frac{1}{12N_u} \tag{2.23}$$

- Anderson-Darling Statistic:

$$A^2 = -N_u - \frac{1}{N_u} \sum_{i=1}^{N_u} \left\{ (2i-1) \log\left(H_{\hat{\xi}}(Y_{i:N_u}|\widehat{\sigma_u}) \right) + (2N_u+1-2i) \log\left(1 - H_{\hat{\xi}}(Y_{i:N_u}|\widehat{\sigma_u}) \right) \right\} \tag{2.24}$$

where $H_{\xi}(.)$ is the distribution function of the GPD and $(\hat{\xi}, \widehat{\sigma_u})$ are the ML estimates for its shape and scale parameters, respectively. The AD statistic is a modification of the CvM statistic, in the sense that the AD statistic gives more weight to observations in the tail of the distribution. These tests were adapted to the GPD hypothesis testing by Choulakian and Stephens (2001).

The procedure that must be followed begins by calculating the test statistics W^2 (2.23) and A^2 (2.24) using the ML estimates of ξ and σ_u. Afterwards, looking at the tables that are presented and explained in Appendix A , it is possible to decide if the null hypothesis should be rejected or not at some level of significance α.

Exponential model vs GPD model

In Statistics, it is recommendable to use the simplest models (principle of parsimony). The GPD model (2.10), which has a shape parameter ξ and scale parameter σ, reduces to the Exponential model when $\xi = 0$. Since these models are nested, it is possible to use a Likelihood Ratio Test (LRT) to see if there is statistical evidence that $\xi \neq 0$, that is, if the GPD model is more appropriate than the Exponential one. Let $l_0(M_0)$ and $l_1(M_1)$ be the maximized log-likelihoods for the Exponential model (denoted by M_0) and the GPD model (denoted by M_1), respectively. The test statistic is the difference of *deviances*, that is, $L = D_0 - D_1 = -2\{l_0(M_0) - l_1(M_1)\}$, which follows approximately a chi-square distribution with 1 degree of freedom. At some level of significance α, the null hypothesis $H_0 : \xi = 0$ is rejected if $L_{obs} > \chi^2_{1;1-\alpha}$, where $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2 distribution with 1 degree of freedom.

If the model chosen for the excesses over a threshold is the Exponential one, the most popular goodness-of-fit test for that distribution is the Lilliefors-corrected Kolmogorov-Smirnov (LcKS) test with null hypothesis H_0 : The excesses follow an Exponential distribution. The Lilliefors correction is used when the parameter of the Exponential distribution is not known and needs to be estimated through the excess data. Using the *R* package *KScorrect*, it is possible to obtain approximate *p*-values via simulation. The interested reader can see Lilliefors (1969) for details about this statistical test.

2.2.4 Dealing with Dependent Sequences

So far, we have assumed that the excesses above a threshold u are i.i.d. However, this assumption is often unrealistic when working with time series, as there is usually, at least, short-term temporal dependence that may affect our analysis. In the literature there are several ways of addressing this issue (see Fawcett and Walshaw, 2008). The most popular one is to carry out a **declustering** process. In this thesis, we will be using the "runs-declustering", which is explained in Coles (2001). This method consists on fitting a GPD model to the sample of the maxima of each cluster of excesses, where clusters are defined as follows: exceedances (observations above u) separated by less than r non-exceedances are included in the same cluster. The *run length* (r) is an integer number selected by the user. This value should be chosen carefully, in the sense that if r is too low, there are too many clusters and the problem of short-term dependence may not be solved; if r is too large, the sample of cluster maxima is too small to make reliable inferences.

At this point, it is important to present a parameter called *extremal index*, denoted by θ, which is approximately equal to the inverse of the mean cluster size. As it is evident, $\theta \in (0,1]$ and the lower the value of θ, the higher the level of clustering within the sample of excesses.

In applications, it is often interesting to estimate, for a high number m, the m-observation return level (x_m), which satisfies $P(X > x_m) = p$, where $p = \dfrac{1}{m}$. That is, x_m is exceeded once in every m observations. It can be estimated as follows:

$$\widehat{x_m} = u + \frac{\widehat{\sigma}}{\widehat{\xi}} \left[\left(m \frac{N_u}{n} \hat{\theta} \right)^{\widehat{\xi}} - 1 \right], \tag{2.25}$$

where $\widehat{\sigma}$ and $\widehat{\xi}$ are the estimates of the parameters of the GPD model fitted to the cluster maxima, N_u is the number of excesses above the threshold u, n is the total number of observations of the series and $\hat{\theta} = \dfrac{N_c}{N_u}$ is the estimate of the extremal index, with N_c being the number of clusters of excesses.

For an Exponential model with estimated scale parameter $\hat{\sigma}$ and keeping the notation presented above, expression (2.25) reduces to:

$$\widehat{x_m} = u + \hat{\sigma} \log \left(m \frac{N_u}{n} \hat{\theta} \right). \tag{2.26}$$

3 | Bivariate Extreme Value Theory

In this chapter our explanations will essentially be based on Coles (2001), which provides a very intuitive approach to bivariate Extreme Value Theory. In Beirlant et al. (2004) it is possible to find more advanced details about these theoretical issues and consequently it is also an important reference at this point.

3.1 Asymptotic Characterization of Componentwise Maxima

In the study of extremes of two or more variables, the simplest approach is to model each of them individually using univariate techniques. However, it is generally more interesting and useful to analyse the relationships that may exist between them.

Let $\mathbf{X}_i = (X_{i,1},...,X_{i,p}), i \in \{1,...,n\}$ be a sequence of i.i.d random vectors, with joint distribution function F and marginal distributions F_1, F_2, ... , F_p. The vector of maxima, which will be denoted by $\mathbf{M}$, is defined as:

$$\mathbf{M} = (M_1,...,M_p), \tag{3.1}$$

where $M_j = \max\limits_{i\in\{1,...,n\}} X_{i,j}$, for $j = 1,...,p$.

Note that vector $\mathbf{M}$ does not necessarily correspond to an observed vector in the original series; see Coles (2001, page 143).

Let $\mathbf{x} = (x_1,x_2,...,x_p)$ and $\mathbf{y} = (y_1,y_2,...,y_p)$ and consider that the relation $\mathbf{x} \leq \mathbf{y}$ is defined as $x_j \leq y_j$ for all $j \in \{1,...,p\}$. The exact distribution function of $\mathbf{M}$ is given by:

$$P(\mathbf{M} \leq \mathbf{x}) = P(\mathbf{X}_1 \leq \mathbf{x},...,\mathbf{X}_n \leq \mathbf{x}) = F^n(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^p \tag{3.2}$$

Likewise in the univariate case, the vector of maxima conveniently centered and scaled converges to a non-degenerate distribution. As such, Definition 2.1.1 can be extended to the multivariate case as follows:

Definition 3.1.1 (Multivariate Domain of attraction) *If there exist sequences of vectors* $(\mathbf{a}_n)_n > \mathbf{0}$ *and* $(\mathbf{b}_n)_n$ *such that* $\mathbf{a}_n^{-1}(\mathbf{M}_n - \mathbf{b}_n)$ *converges in distribution to a non-degenerate p-variate distribution function G with non-degenerate margins such that:*

$$P(\mathbf{a}_n^{-1}(\mathbf{M}_n - \mathbf{b}_n)) = F^n(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n) \to G(\mathbf{x}), \quad n \to \infty \quad , \tag{3.3}$$

F is in the ***domain of attraction*** *of a multivariate extreme value distribution G (and this is denoted by* $F \in D(G)$*).*

From now on in this chapter, we will only consider the bivariate case ($p = 2$) because, as it is explained in Coles (2001), it allows us to emphasise the main concepts without getting embroiled in the complexity of the notation:

Theorem 3.1.2 *Consider* $\mathbf{M}^* = (M_1^*, M_2^*)$*, where for each* $j \in \{1,2\}$*,* $M_j^* := \dfrac{M_j}{n} = \dfrac{\max\limits_{i \in \{1,\ldots,n\}} X_{i,j}}{n}$*, with* (X_{i1}, X_{i2})*,* $i \in \{1,\ldots,n\}$ *being independent random vectors with standard Fréchet marginal distributions* [1] *,that is, for* $j \in \{1,2\}$*,* $F_{X_j}(x_j) = \exp(-1/x_j), \;\; x_j > 0$ *. If, when* $n \to \infty$*,*

$$P(M_1^* \leq x_1, M_2^* \leq x_2) \xrightarrow{d} G(x_1, x_2), \tag{3.4}$$

where G is a non-degenerate distribution function, then:

$$G(x_1, x_2) = \exp\{-V(x_1, x_2)\}, \quad x_1 > 0,\, x_2 > 0, \tag{3.5}$$

where

$$V(x_1, x_2) = 2\int_0^1 \max\left(\frac{w}{x_1}, \frac{1-w}{x_2}\right) dH(w), \quad x_1 > 0,\, x_2 > 0, \tag{3.6}$$

and H is a measure, known as spectral measure, defined on $[0,1]$ *satisfying the constraint:*

$$\int_0^1 w dH(w) = 1/2. \tag{3.7}$$

The term "**bivariate extreme value distributions**" is used for designating the family of distributions that arise as limits in (3.4). This class of distributions is a one-to-one correspondence with the set of measures H on $[0,1]$ satisfying (3.7).

Provided H is differentiable with density h, it is possible to rewrite (3.6) as follows:

$$V(x_1, x_2) = 2\int_0^1 \max\left(\frac{w}{x_1}, \frac{1-w}{x_2}\right) h(w) dw, \quad x_1 > 0,\, x_2 > 0 \tag{3.8}$$

However, (3.7) may be satisfied in some cases in which H is not differentiable. This situation is usually highlighted by the following two cases that reflect indepencence and perfect dependence, respectively:

- If H places mass 0.5 on $w = 0$ and $w = 1$, the corresponding bivariate extreme value distribution is $G(x_1, x_2) = \exp\left\{-(x_1^{-1} + x_2^{-1})\right\}$, $x_1 > 0$, $x_2 > 0$.
- If H places unit mass on $w = 0.5$, the corresponding bivariate extreme value distribution is $G(x_1, x_2) = \exp\left\{-\max(x_1^{-1}, x_2^{-1})\right\}$, $x_1 > 0$, $x_2 > 0$.

The complete class of bivariate extreme value distributions can be obtained letting

$$\tilde{x}_j = \left[1 + \xi_j \left(\frac{x_j - \mu_j}{\sigma_j}\right)\right]^{1/\xi_j}, \quad j \in \{1,2\}. \tag{3.9}$$

The bivariate extreme value distribution function can then be expressed as follows:

$$G(x_1, x_2) = \exp\{-V(\tilde{x}_1, \tilde{x}_2)\}, \quad x_1 > 0,\, x_2 > 0, \tag{3.10}$$

[1] Standard Fréchet is the most popular option for the marginal distributions. However, other alternatives such as Weibull, Gumbel, exponential and uniform distributions are also present in the literature, as it is explained in Beirlant et al. (2004).

as long as $\left[1+\frac{\xi_1\left(x_1-\mu_1\right)}{\sigma_1}\right]>0$ and $\left[1+\frac{\xi_2\left(x_2-\mu_2\right)}{\sigma_2}\right]>0$, and where V is calculated from (3.6), assuming that the measure H satisfies (3.7). The marginal distributions are GEV (see expression (2.6)) with parameters (μ_1, σ_1, ξ_1) and (μ_2, σ_2, ξ_2), this is, for $j \in \{1,2\}$, $G_j(x_j) = \exp\left(-\frac{1}{\tilde{x}_j}\right)$, $x_j > 0$.

It is also possible to write (3.10) as follows:

$$\begin{aligned} G(x_1,x_2) &= \exp\left[\log\left\{G_1(x_1)G_2(x_2)\right\}A\left(\frac{\log\left\{G_2(x_2)\right\}}{\log\left\{G_1(x_1)G_2(x_2)\right\}}\right)\right] \\ &= \exp\left\{-\left(\frac{1}{\tilde{x}_1}+\frac{1}{\tilde{x}_2}\right)A\left(\frac{\tilde{x}_1}{\tilde{x}_1+\tilde{x}_2}\right)\right\}, \quad x_1>0,\ x_2>0\,. \end{aligned} \tag{3.11}$$

The function $A(.)$ in (3.11) is called **Pickands dependence function** and has the following expression:

$$A(t) = 1 - t + 2\int_0^t H(u)du, \quad t \in [0,1], \tag{3.12}$$

where $H(.)$ is the spectral measure. $A(.)$ is a convex function defined on $[0,1]$ with $\max(t, 1-t) \leq A(t) \leq 1$ for all $0 \leq t \leq 1$. This definition of $A(.)$ that we present here is also used in Beirlant et al. (2004), and it is Pickands´ original definition. However, in the *R* package *evd* (Stephenson, 2018), the alternative definition $A^*(t) = A(1-t)$ for $t \in [0,1]$ is the default option.

In case of a perfect dependence between X_1 and X_2, $A(t) = \max(t, 1-t)$, $\forall t \in [0,1]$, whereas $A(t) = 1$, $\forall t \in [0,1]$ if the random variables are independent.

There are some extremal coefficients that can be calculated using Pickands dependence function, for example in Beirlant et al. (2004) the coefficient $\theta = 2A(1/2)$ is presented [2] . It measures the strength of dependence between X_1 and X_2 : they are independent if and only if $\theta = 2$ and they are perfectly dependent if and only if $\theta = 1$, this is, the strength of dependence increases as θ decreases. Again, an alternative coefficient is given as output by the *R* package *evd*: it is called *Dependence* and it is defined as $2(1-A(1/2))$; note that $A(1/2) = A^*(1/2)$. The interpretation is similar: independence corresponds to $Dependence = 0$ and perfect dependence to $Dependence = 1$: in this case, the strength of dependence increases as *Dependence* increases.

We end this section by presenting a bivariate counterpart of the max-stability property:

$$G^n(x_1,x_2) = G(n^{-1}x_1, n^{-1}x_2), \quad \forall n \in \mathbb{N}, \quad x_1>0,\ x_2>0 \tag{3.13}$$

This property states that if a bivariate random vector has distribution function G, then the distribution function of the vector of componentwise maxima (for any sample size n) is also G, apart from a rescaling by n^{-1}. Analogously to the univariate case, it can be proved that the only distributions that have the max-stability property written above are the bivariate extreme value distributions. Expression (3.13) is obtained by means of the relationship

$$V(a^{-1}x_1, a^{-1}x_2) = aV(x_1,x_2), \quad \forall a>0, \quad x_1>0,\ x_2>0 \tag{3.14}$$

which is straightforward to verify from (3.6).

[2] In this chapter, θ stands for the dependence between two variables. It should not be confused with the *extremal index*, which was presented in the previous chapter (in Subsection 2.2.4) using the same letter.

3.2 Parametric Models

Unlike the univariate case, the class of bivariate extreme value distributions has no finite parameterisation. Resorting to convenient parametric sub-models is a very common way of dealing with this issue. Conceptually, it seems feasible that a parametric family for G can be obtained from a parametric family for H satisfying (3.7) by substitution into (3.6) and (3.5). However, it is not so straightforward in practice because the integral in (3.6) is frequently not tractable and, additionally, the mean of the resulting parametric family may happen to be parameter-dependent, which is also a problem. (Coles, 2001)

We will now present some of the most common bivariate parametric models. For each model, we will indicate the bivariate distribution function G and some of its most important characteristics. In the following expressions, for each $j \in \{1,2\}$, consider $\tilde{y}_j = 1/\tilde{x}_j$, with $\tilde{x}_j$ defined as (3.9). Note that $\tilde{x}_j$ is a function of x_j and, therefore, $\tilde{y}_j$ is also a function of x_j, for each $j \in \{1,2\}$.

1. *Logistic Model* (Gumbel, 1960) :

$$G(x_1,x_2) = \exp\left[-\left(\tilde{y}_1^{1/\alpha} + \tilde{y}_2^{1/\alpha}\right)^{\alpha}\right], \quad x_1 > 0,\, x_2 > 0 \tag{3.15}$$

where $0 < \alpha \leq 1$. X_1 and X_2 are independent if and only if $\alpha = 1$, and the strength of dependence increases as α decreases. The variables are perfectly dependent in the limit as α approaches zero. This model has the drawback that it is symmetric in the two variables, and it may not be appropriate in some situations.

2. *Asymmetric Logistic Model* (Tawn, 1988) :

$$G(x_1,x_2) = \exp\left\{-(1-\psi_1)\tilde{y}_1 - (1-\psi_2)\tilde{y}_2 - \left[(\psi_1\tilde{y}_1)^{1/\alpha} + (\psi_2\tilde{y}_2)^{1/\alpha}\right]^{\alpha}\right\}, x_1 > 0, x_2 > 0 \tag{3.16}$$

where $0 < \alpha \leq 1$ and $0 \leq \psi_1, \psi_2 \leq 1$. When $\psi_1 = \psi_2 = 1$, this model is equivalent to the *Logistic Model*. X_1 and X_2 are independent when $\alpha = 1$ or $\psi_1 = 0$ or $\psi_2 = 0$. When $\psi_1 = \psi_2 = 1$ and α approaches zero, the variables are perfectly dependent.

3. *Husler-Reiss Model* (Hüsler and Reiss, 1989):

$$G(x_1,x_2) = \exp\left(-\tilde{y}_1\Phi\left\{r^{-1} + 0.5r[\log(\tilde{y}_1/\tilde{y}_2)]\right\} - \tilde{y}_2\Phi\left\{r^{-1} + 0.5r[\log(\tilde{y}_2/\tilde{y}_1)]\right\}\right), \quad x_1 > 0,\, x_2 > 0 \tag{3.17}$$

where $r > 0$ and $\Phi(\cdot)$ is the standard normal distribution function. X_1 and X_2 are independent when r approaches zero and they are perfectly dependent when r tends to infinity. [3]

4. *Negative Logistic Model* (Joe, 1990) :

$$G(x_1,x_2) = \exp\left\{-\tilde{y}_1 - \tilde{y}_2 + \left[\tilde{y}_1^{-r} + \tilde{y}_2^{-r}\right]^{-1/r}\right\}, x_1 > 0,\, x_2 > 0 \tag{3.18}$$

where $r > 0$. The interpretation of r with respect to independence and perfect dependence is the same as in the *Husler-Reiss Model*.

5. *Asymmetric Negative Logistic Model* (Joe, 1990) :

$$G(x_1,x_2) = \exp\left\{-\tilde{y}_1 - \tilde{y}_2 + \left[(\psi_1\tilde{y}_1)^{-r} + (\psi_2\tilde{y}_2)^{-r}\right]^{-1/r}\right\}, x_1 > 0,\, x_2 > 0 \tag{3.19}$$

where $r > 0$ and $0 < \psi_1, \psi_2 \leq 1$. When $\psi_1 = \psi_2 = 1$, this model reduces to the *Negative Logistic Model*. X_1 and X_2 are independent when either r, ψ_1 or ψ_2 approaches 0. The variables are perfectly dependent when ψ_1 and ψ_2 approach 1 and r tends to infinity.

[3] This parameterization, which is used in the *R* package *evd*, is slightly different from the one presented in the paper.

6. *Bilogistic Model* (Smith, 1990) :

$$G(x_1,x_2) = \exp\left\{-\tilde{y}_1 q^{1-\alpha} - \tilde{y}_2(1-q)^{1-\beta}\right\}, x_1 > 0,\ x_2 > 0 \tag{3.20}$$

where $q = q(\tilde{y}_1,\tilde{y}_2;\alpha,\beta)$ is the root of the equation $(1-\alpha)\tilde{y}_1(1-q)^{\beta} - (1-\beta)\tilde{y}_2 q^{\alpha} = 0$. The two parameters ($\alpha$ and β) lie in the interval $(0,1)$. When $\alpha = \beta$, this model reduces to the *Logistic Model* with dependence parameter $\alpha = \beta$. X_1 and X_2 are independent in two situations: when $\alpha = \beta$ approaches 1 , and when either α or β is fixed and the other parameter approaches 1. The variables are perfectly dependent when $\alpha = \beta$ approaches 0.

7. *Negative Bilogistic Model* (Coles and Tawn, 1994):

$$G(x_1,x_2) = \exp\left\{-\tilde{y}_1 - \tilde{y}_2 + \tilde{y}_1 q^{1+\alpha} + \tilde{y}_2(1-q)^{1+\beta}\right\}, x_1 > 0,\ x_2 > 0 \tag{3.21}$$

where $q = q(\tilde{y}_1,\tilde{y}_2;\alpha,\beta)$ is the root of the equation $(1+\alpha)\tilde{y}_1 q^{\alpha} - (1+\beta)\tilde{y}_2(1-q)^{\beta} = 0$. The two parameters ($\alpha$ and β) are greater than 0. When $\alpha = \beta$, this model reduces to the *Negative Logistic Model* with dependence parameter $\dfrac{1}{\alpha} = \dfrac{1}{\beta}$. X_1 and X_2 are independent in two situations: when $\alpha = \beta$ tends to infinity , and when either α or β is fixed and the other parameter tends to infinity. The variables are perfectly dependent when $\alpha = \beta$ approaches 0.

8. *Coles-Tawn Model* [4] (Coles and Tawn, 1991) :

$$G(x_1,x_2) = \exp\left\{-\tilde{y}_1\left[1 - Be(q;\alpha+1,\beta)\right] - \tilde{y}_2 Be(q;\alpha,\beta+1)\right\}, x_1 > 0, x_2 > 0 \tag{3.22}$$

where $q = \dfrac{\alpha\tilde{y}_2}{\alpha\tilde{y}_2 + \beta\tilde{y}_1}$ and $Be(q;a,b)$ is the beta distribution function evaluated at q with parameters a and b. The two parameters of the *Coles-Tawn Model* (α and β) are larger than 0. X_1 and X_2 are independent in two situations: when $\alpha = \beta$ approaches 0 , and when either α or β is fixed and the other one approaches 0. The variables are perfectly dependent when $\alpha = \beta$ tends to infinity.

(see, e.g., Beirlant et al. (2004) and Coles (2001) for further details about bivariate models).

In Figure 3.1 it is possible to see the Pickands dependence function (see (3.12)) for the logistic model (3.15), which is the most common model in practice, with three different values of its parameter α, namely 0.9, 0.7 and 0.5. As it is clearly visible, as the value of α decreases, the extremal dependence between X_1 and X_2 increases and thus the corresponding Pickands dependence function is closer to $A(t) = \max(t, 1-t)$, $t \in [0,1]$ (the perfect dependence case) and farther away from $A(t) = 1$, $t \in [0,1]$ (the independence case).

3.3 Statistical Modelling of Componentwise Block Maxima

Although this statistical approach will not be put into practice in our study, we will include a brief explanation about how to model componentwise block maxima (extensive information about this topic can be found in Beirlant et al. (2004)) :

Let $(x_1,y_1),...,(x_n,y_n)$ be the original series and $(z_{1,1},z_{2,1}),...,(z_{1,m},z_{2,m})$ the sequence of componentwise block maxima, with m being the number of blocks used (as in the univariate case, the block size is usually taken as one year of observations). The series $\left\{z_{1,j}\right\}_{j\in\{1,...,m\}}$ and $\left\{z_{2,j}\right\}_{j\in\{1,...,m\}}$ are firstly treated separately, being modelled as in the univariate case. For each $i \in \{1,2\}$, $\left\{z_{i,j}\right\}_{j\in\{1,...,m\}}$ is consid-

[4] This model is also called *Dirichlet Model* because the standard Dirichlet family is used to construct it.

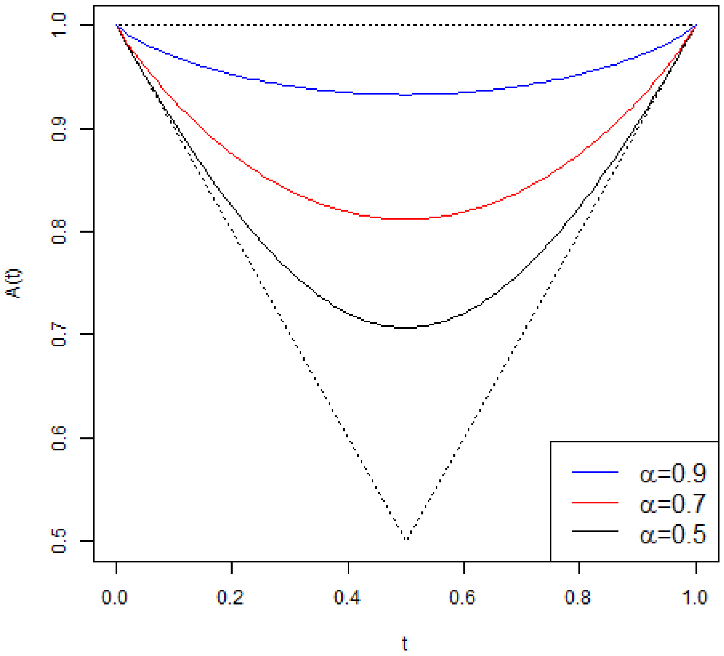


Figure 3.1: Pickands dependence function for the logistic model (3.15) with $\alpha = 0.9$ (solid blue line), $\alpha = 0.7$ (solid red line) and $\alpha = 0.5$ (solid black line). The dashed black lines refer to the functions $A(t) = 1,\ t \in [0,1]$ and $A(t) = \max(t, 1-t)\ ,\ t \in [0,1]$, which correspond to the independence and perfect dependence case, respectively. The plot was made using the *R* package *evd*.

ered a random sample of a variable $Z_i \frown GEV(\mu_i, \sigma_i, \xi_i)$. Then, by using the ML estimates $(\hat{\mu}_i, \hat{\sigma}_i, \hat{\xi}_i)$, for each $i \in \{1,2\}$, the transformed variable $\tilde{Z}_i = \left[1 + \hat{\xi}_i \left(\frac{Z_i - \hat{\mu}_i}{\hat{\sigma}_i}\right)\right]^{1/\hat{\xi}_i}$ follows approximately a standard Fréchet distribution. Applying this transformation to the pairs $(z_{1,j}, z_{2,j})$, the pairs $(\tilde{z}_{1,j}, \tilde{z}_{2,j})$ are obtained, which approximately constitute a random sample of a vector having distribution function of the form (3.5). We assume that G follows one of the parametric models that we presented in Section 3.2 . If g is the corresponding probability density function and $\theta = (\theta_1, \theta_2, ..., \theta_k)$ the parameter vector of the model, it is possible to maximize the corresponding log-likelihood $\ell(\theta) = \sum_{i=1}^{m} \log g(\tilde{z}_{1,i}, \tilde{z}_{2,i} | \theta)$, obtaining the ML estimates for $\theta_1, \theta_2, ..., \theta_k$. As usual, asymptotic confidence intervals for these parameters resulting from the approximation of the ML estimators to the normal distribution can be computed.

If we think of g as the density obtained from (3.10) rather than through (3.5), a joint likelihood is obtained, allowing marginal and dependence parameters to be estimated simultaneously. This one-step procedure improves statistical efficiency, but it is more demanding from the computational point of view. (Coles, 2001)

3.4 Excesses Over A Threshold

The componentwise block maxima approach that we introduced in the previous section has two important disadvantages: all data except the vector of maxima of each block are discarded and there is no guarantee that this vector has been observed. Therefore, in our study we will be considering the excesses over a threshold, which is a more flexible and efficient approach. The material contained in this section is fundamentally based on Coles (2001).

Bivariate Threshold Excess Model

Let $(x_{1,1},x_{1,2}),...,(x_{n,1},x_{n,2})$ be independent realizations of a random vector (X_1,X_2) with joint distribution function F. For each $j\in\{1,2\}$, it is considered that the observations above an appropriate threshold u_j follow a GPD (2.10). Letting N_{u_j} be the number of excesses over the threshold u_j, the following transformed variables are used:

$$\tilde{X}_j=-\left(\log\left\{1-\frac{N_{u_j}}{n}\left[1+\frac{\xi_j(X_j-u_j)}{\sigma_j}\right]^{-1/\xi_j}\right\}\right)^{-1},\quad j\in\{1,2\} \tag{3.23}$$

$\tilde{X}_j$ follows approximately a standard Fréchet distribution for $X_j>u_j$, for each $j\in\{1,2\}$. It is possible to use Theorem 3.1.2 considering the random vector $(\tilde{X}_1,\tilde{X}_2)$, with joint distribution function $\tilde{F}$. For each $j\in\{1,2\}$, let $\tilde{M}_j=\max\limits_{i\in\{1,...,n\}}\tilde{X}_{i,j}$, with $(\tilde{X}_{i1},\tilde{X}_{i2})$, $i\in\{1,...,n\}$ being i.i.d. to $(\tilde{X}_1,\tilde{X}_2)$. Therefore, the following relationship follows:

$$\tilde{F}^n(n\tilde{x}_1,n\tilde{x}_2)=P\left(\tilde{M}_1\le n\tilde{x}_1,\tilde{M}_2\le n\tilde{x}_2\right)=P\left(\frac{\tilde{M}_1}{n}\le\tilde{x}_1,\frac{\tilde{M}_2}{n}\le\tilde{x}_2\right)\approx G(\tilde{x}_1,\tilde{x}_2), \tag{3.24}$$

for $\tilde{x}_1,\tilde{x}_2>0$. Taking into account the expression above (3.24) and the property of max-stability (3.13), it follows that:

$$\begin{aligned}\tilde{F}(\tilde{x}_1,\tilde{x}_2)=\left\{\tilde{F}^n(\tilde{x}_1,\tilde{x}_2)\right\}^{1/n}=\left\{\tilde{F}^n\left(n\frac{\tilde{x}_1}{n},n\frac{\tilde{x}_2}{n}\right)\right\}^{1/n}\approx\left\{G\left(\frac{\tilde{x}_1}{n},\frac{\tilde{x}_2}{n}\right)\right\}^{1/n}=\\ =\left\{G^n(\tilde{x}_1,\tilde{x}_2)\right\}^{1/n}=G(\tilde{x}_1,\tilde{x}_2)\end{aligned} \tag{3.25}$$

Keeping in mind that $F(x_1,x_2)=\tilde{F}(\tilde{x}_1,\tilde{x}_2)$, it is possible to write:

$$F(x_1,x_2)\approx G(x_1,x_2),\quad x_1>u_1,\ x_2>u_2 \tag{3.26}$$

This is, for appropriate thresholds u_1 and u_2, an arbitrary distribution $F(x_1,x_2)$ can be approximated by a distribution of the form (3.5) within the region $x_1>u_1,\ x_2>u_2$.

Censored-likelihood method

It is difficult to make inference on the Bivariate Threshold Excess Model because it may happen that, for a given point (x_1,x_2), only one of its components exceeds the corresponding threshold. We will divide the plane into four regions:

$$\begin{aligned}R_{0,0}&=(-\infty,u_1)\times(-\infty,u_2) \qquad & R_{1,0}&=[u_1,\infty)\times(-\infty,u_2)\\ R_{0,1}&=(-\infty,u_1)\times[u_2,\infty) \qquad & R_{1,1}&=[u_1,\infty)\times[u_2,\infty)\end{aligned}$$

It is possible to apply model (3.26) to the points in $R_{1,1}$, so the likelihood contribution for that region can be obtained by directly using the density of F. However, F cannot be used for the other regions, so a censored-likelihood approach is performed. For instance, for a point $(x_1,x_2)\in R_{1,0}$, its likelihood contribution is as follows:

$$P(X_1 = x_1, X_2 \leq u_2) = \frac{\partial F}{\partial x_1}\Big|_{(x_1,u_2)} \tag{3.27}$$

The idea for this expression is that the only information regarding F corresponds to the x_1-component because $x_1 > u_1$ and $x_2 < u_2$. Applying the same reasoning to the other regions, the following likelihood function arises:

$$L(\theta;(x_{1,1},x_{1,2}),...,(x_{n,1},x_{n,2})) = \prod_{i=1}^{n} \psi(\theta;(x_{i,1},x_{i,2})), \tag{3.28}$$

where θ is the parameter vector of the model and

$$\psi(\theta;(x_1,x_2)) = \begin{cases} \frac{\partial^2 F}{\partial x_1 \partial x_2}\Big|_{(x_1,x_2)}, & (x_1,x_2) \in R_{1,1}, \\ \frac{\partial F}{\partial x_1}\Big|_{(x_1,u_2)}, & (x_1,x_2) \in R_{1,0}, \\ \frac{\partial F}{\partial x_2}\Big|_{(u_1,x_2)}, & (x_1,x_2) \in R_{0,1}, \\ F(u_1,u_2), & (x_1,x_2) \in R_{0,0}, \end{cases} \tag{3.29}$$

with F approximated by one of the parametric models G presented in Section 3.2 , in accordance with expression (3.26).

Using the likelihood function in (3.28), it is possible to obtain ML estimates and asymptotic confidence intervals for the parameters of the model. Analogously to what was explained in the case of the Componentwise Block Maxima approach, marginal and dependence parameters can be estimated in one step (joint estimation) or two steps (separate estimation). This is, in the one-step procedure the likelihood in (3.28) is a function of all parameters, while in the two-step procedure the marginal parameters (the parameters of each GPD) are estimated firstly (the transformations in (3.23) are undertaken afterwards), and therefore the likelihood in (3.28) is only a function of the dependence parameters.

Apart from this censored-likelihood method, there are other statistical approaches to deal with bivariate extremes, such as using point processes. However, we will not include point processes in this thesis because its use is not recommended: poor estimates can be obtained in practice (Ledford and Tawn, 1996) .

3.5 Asymptotic Independence and Asymptotic Dependence

Asymptotic independence is a situation under which models based on Theorem 3.1.2 are not appropriate because they usually lead to an overestimation of the dependence between X_1 and X_2. The explanations presented here can be basically found in Coles (2001). We will begin our explanation of this concept by assuming that X_1 and X_2 have the same distribution function F (this is the approach used in Sibuya (1960)). Let $\tau^+ = \sup\{\tau \in \mathbb{R} : F(\tau) < 1\}$, then (X_1,X_2) is said to be *asymptotically independent* if

$$\lim_{\tau \to \tau^+} P(X_2 > \tau \mid X_1 > \tau) = 0 \tag{3.30}$$

In contrast, they are *asymptotically dependent* if the previous limit is a constant different from zero.

In a general framework, considering F_1 and F_2 as the marginal distributions of X_1 and X_2 respectively, the following coefficient is defined:

$$\chi := \lim_{u \to 1} P(F_2(X_2) > u \mid F_1(X_1) > u) \tag{3.31}$$

χ takes values between 0 and 1 : when X_1 and X_2 are asymptotically independent, $\chi = 0$; and when they are asymptotically dependent, $0 < \chi \leq 1$. Regarding asymptotically dependent variables, the extremal dependence is stronger as χ increases.

Letting J be the joint distribution function of $(F_1(X_1), F_2(X_2))$ and taking into account that, for each $j \in \{1,2\}$, $F_j(X_j)$ follows a $U(0,1)$ distribution (Probability Integral Transformation), the coefficient χ defined in (3.31) may also be obtained as follows:

$$\chi = \lim_{u \to 1} \chi(u), \tag{3.32}$$

where

$$\chi(u) = 2 - \frac{\log P(F_1(X_1) \leq u, F_2(X_2) \leq u)}{\log P(F_1(X_1) \leq u)} = 2 - \frac{\log J(u,u)}{\log u}, \tag{3.33}$$

for $0 < u < 1$.
$\chi(u)$ is bounded by:

$$2 - \frac{\log\{\max(2u-1, 0)\}}{\log u} \leq \chi(u) \leq 1, \quad 0 < u < 1, \tag{3.34}$$

see, e.g. , Beirlant et al. (2004). The reasoning to obtain (3.32) from (3.31) is the following [5] :

$$\begin{aligned} P(F_2(X_2) > u \mid F_1(X_1) > u) &= \frac{P(F_1(X_1) > u, F_2(X_2) > u)}{P(F_1(X_1) > u)} = \\ &= \frac{1 - P(F_1(X_1) \leq u) - P(F_2(X_2) \leq u) + P(F_1(X_1) \leq u, F_2(X_2) \leq u)}{1 - P(F_1(X_1) \leq u)} = \frac{1 - 2u + J(u,u)}{1-u} = \\ &= 2 - \frac{1 - J(u,u)}{1-u} \sim 2 - \frac{\log J(u,u)}{\log u} \end{aligned} \tag{3.35}$$

as $u \to 1$.

In order to measure the strength of extremal dependence for asymptotically independent variables, there is an alternative coefficient $\bar{\chi}$, which is obtained as follows, by analogy with (3.32) and (3.33) :

$$\bar{\chi} = \lim_{u \to 1} \bar{\chi}(u), \tag{3.36}$$

where

$$\bar{\chi}(u) = \frac{2 \log P(F_1(X_1) > u)}{\log P(F_1(X_1) > u, F_2(X_2) > u)} - 1 = \frac{2 \log(1-u)}{\log P(F_1(X_1) > u, F_2(X_2) > u)} - 1, \tag{3.37}$$

for $0 < u < 1$.

[5] For the last step of this reasoning, note that $\lim_{u \to 1} \frac{\log u}{1-u} = -1$ and $\lim_{u \to 1} \frac{\log J(u,u)}{1 - J(u,u)} = -1$, so $1 - u \sim -\log u$ and $1 - J(u,u) \sim -\log J(u,u)$ as $u \to 1$. Therefore, $\frac{1 - J(u,u)}{1-u} \sim \frac{\log J(u,u)}{\log u}$ as $u \to 1$.

$\bar{\chi}(u)$ is bounded by:

$$\frac{2\log(1-u)}{\log\{\max(1-2u,0)\}} - 1 \leq \bar{\chi}(u) \leq 1, \quad 0 < u < 1. \tag{3.38}$$

see, e.g. , Beirlant et al. (2004). $\bar{\chi}$ takes values between -1 and 1 : when X_1 and X_2 are asymptotically dependent, $\bar{\chi} = 1$; and when they are asymptotically independent, $-1 \leq \bar{\chi} < 1$. For asymptotically independent variables, the extremal dependence is stronger as $\bar{\chi}$ increases.

Therefore, the pair $(\chi, \bar{\chi})$ is used as a summary of the extremal behaviour of (X_1,X_2), in the sense that:

- The variables are **asymptotically dependent** if $\bar{\chi} = 1$ and $0 < \chi \leq 1$, with χ quantifying how strong the dependence at extreme levels is.

- The variables are **asymptotically independent** if $-1 \leq \bar{\chi} < 1$ and $\chi = 0$, with $\bar{\chi}$ quantifying how strong the dependence at extreme levels is.

More detailed information on these coefficients can be found in Coles et al. (1999). In Ledford and Tawn (1996) another measure called *coefficient of tail dependence* (η) is presented: this coefficient satisfies $0 < \eta \leq 1$ and it can also be used for seeing if the variables are asymptotically dependent or asymptotically independent. However, in our study we prefer to use the pair $(\chi, \bar{\chi})$ due to the simplicity of the procedure.

R package *evd* enables to construct *chi plots* and *chi bar plots*, which are plots of $u \in (0,1)$ against empirical estimates of $\chi(u)$ and $\bar{\chi}(u)$ respectively, also containing approximate 95% confidence intervals computed via the delta method. These plots are very useful because observing the behaviour of the graphs as $u \to 1$ allows us to visualize if X_1 and X_2 are asymptotically dependent or asymptotically independent (and the strength of extremal dependence between them). This technique is more interesting for the Excesses Over A Threshold approach than for the Componentwise Block Maxima because in the latter there is frequently not enough data.

As it was said at the beginning of this section, the models that we presented in this thesis are not appropriate for asymptotically independent variables. The development of adequate models for that situation can be seen in Ledford and Tawn (1996, 1997, 1998) .

4 | Copulas

In the previous chapter of this thesis we have been focusing on the analysis of the dependence between two variables at extreme values. However, copulas provide a global picture of the dependence structure, being also very interesting from the extremal point of view. Therefore, they are useful for our study. In books such as Nelsen (2006) and Shemyakin and Kniazev (2017) it is possible to find extensive information on Copula Theory; in this chapter we will explain the most relevant concepts for our analysis.

Copulas have been extensively applied to environmental problems, in which the dependence structures between the variables are commonly non-linear and thus the traditional Gaussian bivariate model is frequently not the best option. This model is not suitable for modelling data which displays strong asymmetries or heavy tails, a situation that generally occurs when working with these kind of data.

In the literature, it is possible to find many studies using copulas to analyze the dependence structure of a pair of hydroclimatic variables, for instance: temperature and precipitation (Cong and Brady, 2012; Lazoglou and Anagnostopoulou, 2019), soil moisture and precipitation (AghaKouchak, 2015), drought duration and severity (Lee et al., 2013; Poonia et al., 2021), groundwater and precipitation (Reddy and Ganguli, 2012), etc. Moreover, copulas can also be used to study the dependence between observed and simulated data, for example in the case of wind speed (see André and de Zea Bermudez (2020)).

The dependence between two variables is usually measured by Pearson's linear correlation coefficient (ρ). Nevertheless, since ρ is based on the assumption of a linear association between the variables, there are other coefficients, like the concordance measures Kendall's tau (τ) and Spearman's rho (ρ_S), which are more useful in this framework because they can be used in the case of non-linear relationships. Moreover, ρ cannot be written as a function of a copula, while τ and ρ_S can; for extensive information about these measures see Embrechts et al. (2003).

In this chapter, we will consider that (X_1, X_2) is a pair of **continuous** random variables because this is the type of variables that we have in our study.

4.1 The Concept of Copula

Letting F_1 and F_2 be the marginal distribution functions of (X_1, X_2), a **copula** C is the joint distribution function of (U_1, U_2), where $U_i = F_i(X_i) \frown U(0,1)$ for $i \in \{1,2\}$. This is,

$$C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2), \quad u_1, u_2 \in (0,1). \tag{4.1}$$

The *copula density function* is defined as:

$$c(u_1, u_2) = \frac{\partial^2 C}{\partial u_1 \partial u_2} = \frac{\partial^2 C}{\partial u_2 \partial u_1}, \quad u_1, u_2 \in (0,1), \tag{4.2}$$

where it is assumed that $\dfrac{\partial^2 C}{\partial u_1 \partial u_2}$ and $\dfrac{\partial^2 C}{\partial u_2 \partial u_1}$ exist and are continuous.

According to **Sklar's Theorem** (Sklar (1959)), there exists a copula C such that the joint distribution function of (X_1, X_2), which we will denote by H, can be expressed as a function of C and the marginal distribution functions, this is:

$$H(x_1, x_2) = C(F_1(x_1), F_2(x_2)), \quad x_1, x_2 \in \mathbb{R} \tag{4.3}$$

Since we are regarding X_1 and X_2 as continuous random variables, the copula C is unique.

Conversely, for any distribution functions F_1 and F_2 and any copula C, it is possible to obtain a joint distribution function H by defining it in accordance with expression (4.3), being F_1 and F_2 the corresponding marginal distribution functions.

4.2 Copula Classes

In this thesis we will present and briefly review basic characteristics of some commonly used copulas. The copulas we will address belong to either the Elliptical or to the Archimedean families of copulas.

Among the **Elliptical copulas**, which do not have closed form expressions, the Gaussian and the Student-t are the most important examples. These two copulas are defined as follows:

- *Gaussian copula*:

  $$C(u_1, u_2; \rho) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \quad u_1, u_2 \in (0,1), \tag{4.4}$$

  where $\Phi_\rho(.,.)$ is the joint distribution function of a standard bivariate normal distribution with correlation ρ and $\Phi^{-1}(.)$ is the inverse of the distribution function of a standard (univariate) normal distribution.

- *Student-t copula*:

  $$C(u_1, u_2; \eta, \rho) = T_{\eta\rho}(T_\eta^{-1}(u_1), T_\eta^{-1}(u_2)), \quad u_1, u_2 \in (0,1), \tag{4.5}$$

  where $T_{\eta\rho}(.,.)$ is the joint distribution function of a Student-t distribution with η degrees of freedom and correlation ρ and $T_\eta^{-1}(.)$ is the inverse of the distribution function of the (univariate) Student-t distribution with η degrees of freedom.

For both the Gaussian and the Student-t copulas, Kendall's τ can be obtained from the correlation ρ by: $\tau = \dfrac{2}{\pi} \arcsin(\rho)$.

In contrast, the **Archimedean copulas** have the following general expression:

$$C(u_1, u_2) = \varphi^{[-1]}\left[\varphi(u_1) + \varphi(u_2)\right], \quad u_1, u_2 \in (0,1), \tag{4.6}$$

where $\varphi : [0,1] \to [0,\infty]$ is the *copula generator*, which is a strictly decreasing, continuous and convex function satisfying $\varphi(1) = 0$. The function $\varphi^{[-1]} : [0,\infty] \to [0,1]$, defined by:

$$\varphi^{[-1]}(t) := \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0) \\ 0, & \varphi(0) \le t \le \infty \end{cases} \tag{4.7}$$

is the *pseudo-inverse* of φ.

The Archimedean copulas have closed form expressions, and three commonly used members of this class are the Frank, Gumbel and Clayton copulas. The expressions of these three copulas are given as follows:

- *Frank copula*:

$$C(u_1,u_2;\alpha) = -\frac{1}{\alpha}\log\left(\frac{1-e^{-\alpha}-(1-e^{-\alpha u_1})(1-e^{-\alpha u_2})}{1-e^{-\alpha}}\right) \tag{4.8}$$

  where α is the association parameter, $\alpha \in \mathbb{R} \setminus \{0\}$, and Kendall's τ is given by $\tau = 1+\dfrac{4(D(\alpha)-1)}{\alpha}$, being $D(\alpha)$ Debye's integral, that is, $D(\alpha) = \dfrac{1}{\alpha}\displaystyle\int_0^{\alpha}\frac{t}{e^t-1}dt$.

- *Gumbel copula*:

$$C(u_1,u_2;\alpha) = \exp\left[-\{(-\log(u_1))^{\alpha}+(-\log(u_2))^{\alpha}\}^{1/\alpha}\right], \quad u_1,u_2 \in (0,1), \tag{4.9}$$

  where α is the association parameter, $\alpha \in [1,\infty)$, and Kendall's τ is given by $\tau = \dfrac{\alpha-1}{\alpha}$.

- *Clayton copula*:

$$C(u_1,u_2;\alpha) = \max\left\{\left(u_1^{-\alpha}+u_2^{-\alpha}-1\right)^{-1/\alpha},0\right\}, \quad u_1,u_2 \in (0,1), \tag{4.10}$$

  where α is the association parameter, $\alpha \in [-1,\infty)\setminus\{0\}$, and Kendall's τ is given by $\tau = \dfrac{\alpha}{\alpha+2}$.

In this thesis it is particularly important to study the *tail dependence*, that is, the dependence between X_1 and X_2 when both variables take high values (*upper tail dependence*) and when both of them take low values (*lower tail dependence*). There are two measures that provide us with this information (and if a copula has a closed form expression, it is possible to write them as a function of that copula) :

- The *coefficient of upper tail dependence*:

$$\lambda_U = \lim_{u\to 1^-} P\left(X_2 > F_2^{-1}(u)\,|\,X_1 > F_1^{-1}(u)\right) = \lim_{u\to 1^-}\frac{1-2u+C(u,u)}{1-u} \tag{4.11}$$

  as long as the limit $\lambda_U \in [0,1]$ exists.

- The *coefficient of lower tail dependence*:

$$\lambda_L = \lim_{u\to 0^+} P\left(X_2 \le F_2^{-1}(u)\,|\,X_1 \le F_1^{-1}(u)\right) = \lim_{u\to 0^+}\frac{C(u,u)}{u} \tag{4.12}$$

  as long as the limit $\lambda_L \in [0,1]$ exists.

Table 4.1 summarizes the tail dependence behaviour of the Elliptical and Archimedean copulas that we have just briefly reviewed.

Table 4.1: Coefficient of upper tail dependence (λ_U) and coefficient of lower tail dependence (λ_L) for some of the most used copulas.

Class	Copula	λ_U	λ_L
Elliptical	Gaussian	0	0
Elliptical	Student-*t*	$2T_{\eta+1}(s)$	$2T_{\eta+1}(s)$
Archimedean	Frank	0	0
Archimedean	Gumbel	$2-2^{-1/\alpha}$	0
Archimedean	Clayton	0	$2^{-1/\alpha}$

$T_{\eta+1}(s)$ is the distribution function of the (univariate) Student-*t* distribution with $\eta+1$ degrees of freedom evaluated at $s := -\sqrt{\eta+1}\sqrt{\frac{1-\rho}{1+\rho}}$.

4.3 Estimation

At this point, the issue consists of how to make inference on copulas when there is a sample $((x_{1,1},x_{1,2}),...,(x_{n,1},x_{n,2}))$ of (X_1,X_2), where X_1 and X_2 are continuous random variables. In the literature it is possible to find parametric, semi-parametric and non-parametric methods of estimation (Joe, 2014; Shemyakin and Kniazev, 2017; Genest and Favre, 2007).

In the parametric approach, parametric models are used for the marginal distributions. Therefore, apart from δ, which is the vector of the copula parameters, it is necessary to estimate $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$, which are the parameter vectors of the distribution of X_1 and X_2, respectively. In this framework, it is possible to use maximum likelihood estimation, which can be performed simultaneously for the marginal and the copula parameters (full ML estimation) or it can be carried out in two steps (*Inference Functions for Margins* (IFM)). In the IFM method, the ML estimates of the marginal parameters are firstly obtained, and then, they are used for computing the ML estimates of the copula parameters.

However, in this thesis we will focus on **semi-parametric** inference because by using this approach the marginal distribution functions are not specified parametrically (they are estimated by their empirical counterparts). In this manner, we avoid transferring to the copula fitting possible misspecification of the parametric models fitted to the marginals. Moreover, semi-parametric methods have the additional benefit of still permitting the use of maximum likelihood estimation. Letting $\widehat{F_1}$ and $\widehat{F_2}$ be the marginal empirical distribution functions of (X_1,X_2), the procedure can be described as follows:

1. For each $i \in \{1,2,...,n\}$, the pseudo-observation ($\hat{u}_i$,$\hat{v}_i$) is calculated, where

$$\hat{u}_i := \frac{n}{n+1}\widehat{F_1}(x_{i,1}) = \frac{1}{n+1}\sum_{j=1}^{n} I_{(x_{j,1}\leq x_{i,1})} \;,$$
$$\hat{v}_i := \frac{n}{n+1}\widehat{F_2}(x_{i,2}) = \frac{1}{n+1}\sum_{j=1}^{n} I_{(x_{j,2}\leq x_{i,2})} \;. \tag{4.13}$$

   Note that $I_A(.)$ is the indicator function, this is, it is a function that takes value 1 if A is satisfied and takes value 0 otherwise.

2. The estimator which results from using the semi-parametric method is called Maximum Pseudo-Likelihood Estimator (MPLE) and is obtained by:

$$\hat{\hat{\delta}}_{MPLE} = \operatorname{argmax} \sum_{i=1}^{n} \log(c(\hat{u}_i, \hat{v}_i)|\delta), \tag{4.14}$$

where δ is the vector that contains the parameters of the copula and $c(.,.)$ is the copula density function.

If the copula chosen is appropriate, then $\hat{\hat{\delta}}_{MPLE}$ is an asymptotically normal and consistent estimator for δ (Genest et al., 1995).

Regarding the non-parametric approach, one method consists on estimating the copula by the *empirical copula*, which is defined as:

$$C_n(u,v) = \frac{1}{n} \sum_{i=1}^{n} I_{(\hat{u}_i \leq u, \hat{v}_i \leq v)}, \quad u, v \in (0,1), \tag{4.15}$$

where $\{(\hat{u}_i, \hat{v}_i), i \in \{1, 2, ..., n\}\}$ are the pseudo-observations defined in (4.13) and $I_{(\hat{u}_i \leq u, \hat{v}_i \leq v)}$ is a function that takes value 1 if $\hat{u}_i \leq u$ and $\hat{v}_i \leq v$, and takes value 0 otherwise.
In Shemyakin and Kniazev (2017) it is possible to find further information about non-parametric copula inference.

4.4 Model Selection and Goodness-of-Fit

In order to compare the fitted copula models, the Akaike's Information Criterion (AIC) is generally used. According to this criteria, the best model is the one with the lowest AIC value (see Akaike, 1974). Using the semi-parametric approach, the AIC associated to a copula model M with k parameters is given by:

$$AIC(M) = 2k - 2\sum_{i=1}^{n} \log(c(\hat{u}_i, \hat{v}_i)|\hat{\hat{\delta}}_{MPLE}), \tag{4.16}$$

where $\{(\hat{u}_i, \hat{v}_i), i \in \{1, 2, ..., n\}\}$ are the pseudo-observations, $c(.,.)$ is the density function of the fitted copula and $\hat{\hat{\delta}}_{MPLE}$ is the MPLE of its parameter vector δ.

With regard to the goodness-of-fit tests for copulas, they are used to assess if the null hypothesis $H_0 : C \in C_0$ should be rejected or not at some level of significance α, where C_0 is a given family of copulas. According to Genest et al. (2009), the best tests for this purpose are the ones based on the $S_n^{(B)}$ and S_n statistics. S_n is a version of Cramér-von-Mises statistic that differs from $S_n^{(B)}$ by the fact that in $S_n^{(B)}$ a probability integral transformation described in Rosenblatt (1952) is used. These tests are based on the comparison of the fitted copula with the empirical one given in (4.15). Taking into account that the distributions followed by S_n and $S_n^{(B)}$ are unknown, bootstrap is required to obtain approximate p-values.

5 | Procedure and Results

5.1 Univariate analysis of Transported Moisture

In this section we will address the univariate analysis of the series of Transported Moisture (TM) from the GPLLJ source region to the jet domain, which was introduced in Chapter 1 . We will firstly present a brief exploratory analysis of the series and, afterwards, the POT analysis with declustering that was carried out.

5.1.1 A brief exploratory analysis

The series of TM for the summer periods (months of June, July and August) is expressed in **mm/day** and has 3496 observations. The data was recorded from 1980 to 2017. The plot of the series, a table of summary statistics as well as the histogram with the kernel density estimate are presented in Figure 5.1, Table 5.1 and Figure 5.2, respectively.

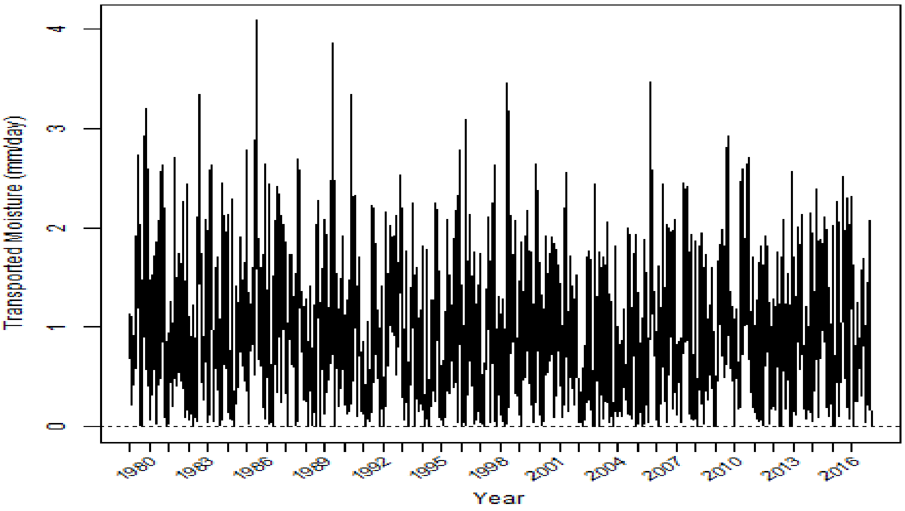


Figure 5.1: Plot of the series of TM from the GPLLJ source region to the jet domain. It comprises daily observations of the summer months (June, July, August) from 1980 to 2017 (38 summers: 3496 observations).

Table 5.1: Summary statistics for the TM series. n denotes the number of observations of the series; $x_{1:n}$ is the minimum and $x_{n:n}$ is the maximum of the values of the series; $\bar{x}$ refers to the mean and sd to the standard deviation. $Q_{0.25}$, $Q_{0.50}$ and $Q_{0.75}$ are the first quartile, the median and the third quartile of the data, respectively.

n	$x_{1:n}$	$Q_{0.25}$	$Q_{0.50}$	$\bar{x}$	$Q_{0.75}$	$x_{n:n}$	sd
3496	0.000	0.411	0.836	0.916	1.321	4.086	0.631

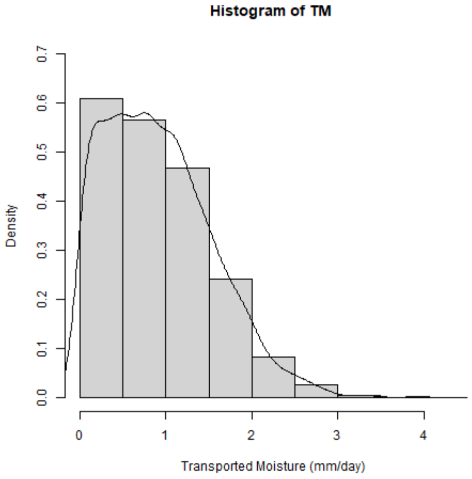


Figure 5.2: Histogram and kernel density estimate for the TM series.

TM is a variable that only takes positive values and its distribution has a right tail. The histogram also suggests positive skewness and positive excess kurtosis, as it was confirmed by the *R* package *moments* (skewness equal to 0.719 and Pearson's measure of kurtosis equal to 3.41).

The plot of the series suggests that it is reasonably stationary except for the largest values, for which a declining trend seems to exist. In the left-hand plot of Figure 5.3 we compare the values of TM which were observed in the first 19 years with the ones recorded during the latest 19 years; in the right-hand plot of that figure we analyze the TM's yearly evolution. We can see that the magnitude of the large values cleary decreases as time goes by.

5.1.2 Threshold Models Approach

We model the data by the POT methodology due to its advantages when compared to the traditional block maxima method. We used a declustering scheme for the exceedances over the chosen threshold in order to deal with the short-term temporal dependence existing between them. First, we will present the threshold selection procedure and, afterwards, the POT analysis with four different *run lengths*.

Threshold selection

The two methods presented in Subsection 2.2.2 were applied to the series under study.
Regarding the first method, the estimated mean excess function presented in Figure 5.4 leads to think that a value around 2 might be an appropriate threshold, as a linearity pattern is clearly visible to the right of that value (see the solid blue line in that figure).
As for the second method, looking at Figure 5.5 , the ML estimates for the shape and modified scale parameters are approximately constant above $u = 2$ (the solid blue line lies within all confidence intervals

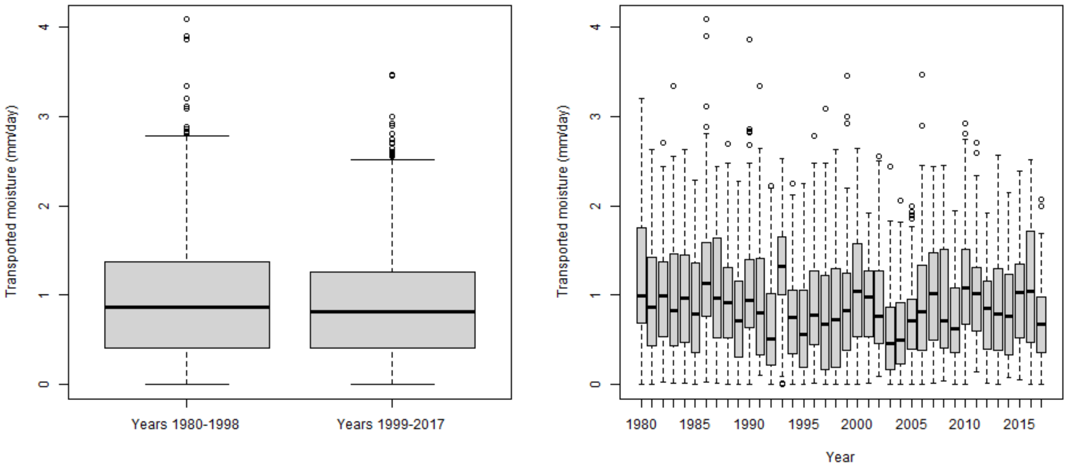


Figure 5.3: Boxplots for the observations of the first and second half of the TM series (Years 1980-1998 and Years 1999-2017), on the left; and boxplots for the observations of each summer from 1980 to 2017 for the TM series, on the right.

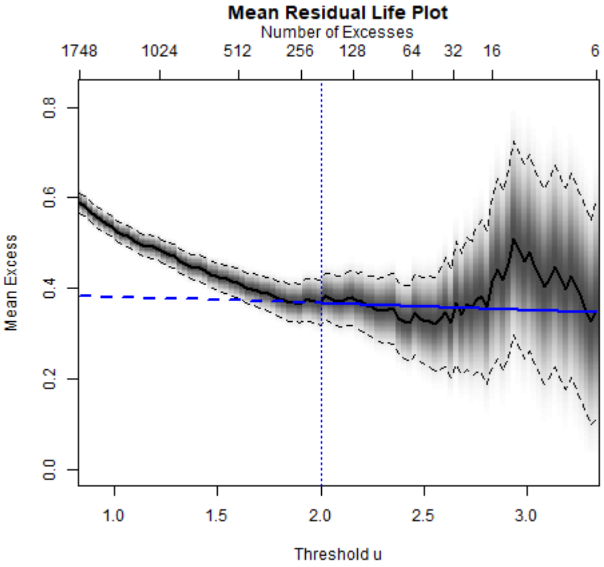


Figure 5.4: Estimated Mean Excess Function of Transported Moisture from the GPLLJ source region to the jet domain (solid black line), with 95% normal-approximation confidence intervals as black dashed lines and fitted solid blue line to the right of $u = 2$. This figure was constructed using *R* package *evmix*.

corresponding to thresholds greater than 2). Therefore, that figure also suggests that $u = 2$ looks as a reasonable choice.

Thus, the information presented in Figures 5.4 and 5.5 clearly supports the choice of $u = 2$. Several other alternative values for u were analyzed although we came to the conclusion that $u = 2$ was an appropriate threshold. The number of observations that exceed the threshold is $N_u = 201$, which approximately corresponds to the 5.75% largest observations of the series under study.

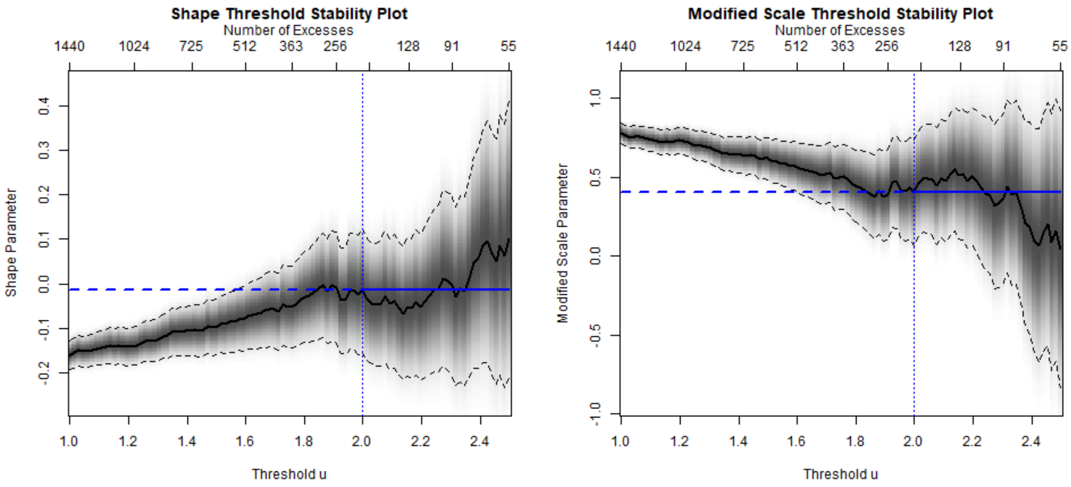


Figure 5.5: Maximum Likelihood estimates for the shape and modified scale parameters of the GPD models fitted to the Transported Moisture series, as a function of the chosen threshold u. The ML estimates are presented as a solid black line and the corresponding 95% normal-approximation confidence intervals as black dashed lines. A solid blue horizontal line is plotted to the right of $u = 2$, indicating the corresponding ML estimates for the shape and modified scale parameters of the GPD model fitted above $u = 2$, respectively. This figure was constructed using *R* package *evmix*.

POT analysis with declustering

In spite of the fact that we are only going to model by the GPD the excesses over the threshold $u = 2$, it is necessary to verify if there is some evidence of clusters of excesses. If so, it is necessary to apply a declustering method in order to remove the dependence between the excesses.

In Figure 5.6 the plot of the exceedances above the threshold $u = 2$ is presented. It is easy to see in that figure that there exists some temporal dependence between them (the exceedances are close to each other forming groups; see, for example, the exceedances corresponding to 1980 or 2010).

Thus, we used the *R* package *evd* to perform "runs-declustering" with *run length* (r) equal to 1, 2, 3 and 4. In Figure 5.7 the plots of the cluster maxima resulting from the declustering process with those values of r are shown. The problem of the dependence between the excesses is solved by applying the declustering process, since only the cluster maxima are used to fit the model and consequently they are more separated between each other. Therefore, the assumption of i.i.d excesses is reasonable for the declustered series. As expected, the effect of the declustering is more visible as the value of r increases.

Table 5.2 contains the results with regard to the number of clusters obtained (N_c), the estimate of the extremal index ($\hat{\theta} = \dfrac{N_c}{N_u}$) and the Maximum Likelihood estimates for the parameters of the GPD ($\hat{\xi}$, $\widehat{\sigma_{GPD}}$) and the Exponential model ($\widehat{\sigma_{EXP}}$), with their corresponding standard errors. Remembering that $\hat{\theta}$ is approximately the inverse of the mean cluster size, it can be said that for the values of r considered in this study, the mean cluster size goes from approximately 2 exceedances above $u = 2$ (for $r = 1$) to approximately 2.4 exceedances above that threshold (for $r = 4$). It is easy to see that the shape parameter of the GPD is very close to 0 for all the values of r considered and the scale parameters of both GPD and Exponential models are approximately equal to 0.4 for all values of r. The standard errors of $\hat{\xi}$, $\widehat{\sigma_{GPD}}$ and $\widehat{\sigma_{EXP}}$ are very similar for all the values of r considered.

Now the question is if we should consider the Exponential or the GPD model for the cluster maxima.

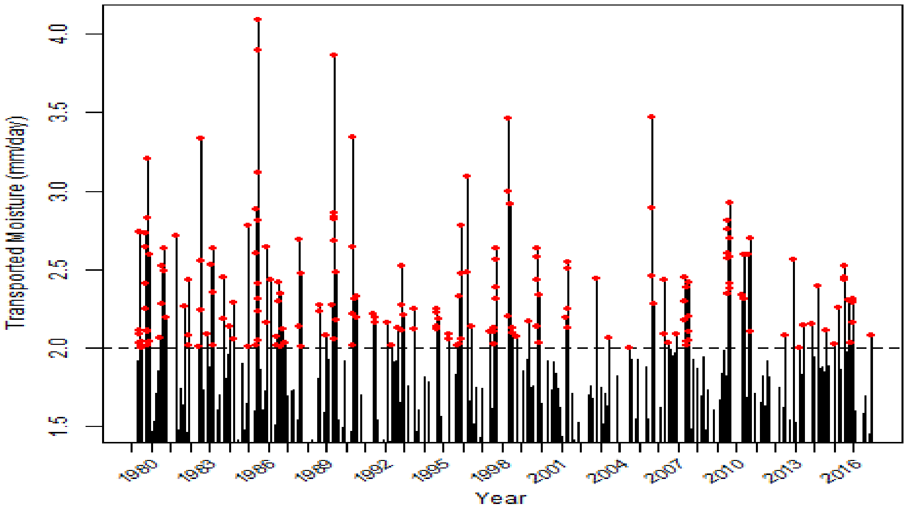


Figure 5.6: Exceedances of the TM series above the threshold $u = 2$.

Table 5.2: Results of the POT analysis with declustering for the TM series choosing threshold $u = 2$, with regard to: the number of clusters obtained (N_c), the estimate of the extremal index ($\hat{\theta} = \dfrac{N_c}{N_u}$) and the Maximum Likelihood estimates for the parameters of the GPD ($\hat{\xi}$, $\widehat{\sigma_{GPD}}$) and the Exponential model ($\widehat{\sigma_{EXP}}$), with their corresponding standard errors. The *run length* (r) is equal to 1, 2, 3 and 4, respectively for each column. Computations were performed using the *R* package *evd*.

	$\boldsymbol{r = 1}$	$\boldsymbol{r = 2}$	$\boldsymbol{r = 3}$	$\boldsymbol{r = 4}$
N_c	102	96	90	83
$\hat{\theta}$	0.507	0.478	0.448	0.413
$\hat{\xi}$ (Std.Err)	0.052 (0.114)	0.052 (0.118)	0.022 (0.118)	0.022 (0.123)
$\widehat{\sigma_{GPD}}$ (Std.Err)	0.367 (0.055)	0.373 (0.058)	0.401 (0.064)	0.408 (0.067)
$\widehat{\sigma_{EXP}}$ (Std.Err)	0.387 (0.038)	0.393 (0.040)	0.410 (0.043)	0.418 (0.046)

Remember that the Exponential model is a special case of the GPD, that is, the GPD model reduces to the Exponential one when the shape parameter ξ equals 0. For the threshold $u = 2$ and performing declustering with *run length* (r) equal to 4, it is possible to visualize the *profile log-likelihood* 95% confidence interval for ξ in Figure 5.8 (Recall expressions (2.7) and (2.8), which are written for the GEV distribution, but they can easily be adapted to the GPD with shape parameter ξ and scale parameter σ). The *profile log-likelihood* 95% confidence intervals for ξ are: $(-0.133, 0.322)$, $(-0.140, 0.334)$, $(-0.170, 0.307)$ and $(-0.178, 0.321)$, for $r = 1,2,3$ and 4, respectively. As we can see, in the four cases, the positive side of the intervals is wider than the negative part, although the value 0 belongs to all of them, so it is consistent with the hypothesis that $\xi = 0$, that is, the Exponential model is better than the GPD one to model the cluster maxima.

In Figure 5.9 it is possible to see the Exponential and the GPD QQ-Plots that were fitted to the cluster maxima, considering $u = 2$ and $r = 4$. The figure clearly shows that both the Exponential and the GPD models are appropriate for modelling the cluster maxima. This can be confirmed by the values of R^2, presented in Table 5.3 , which were obtained by fitting linear models to the theoretical and empirical quantiles (QQ-Plots).

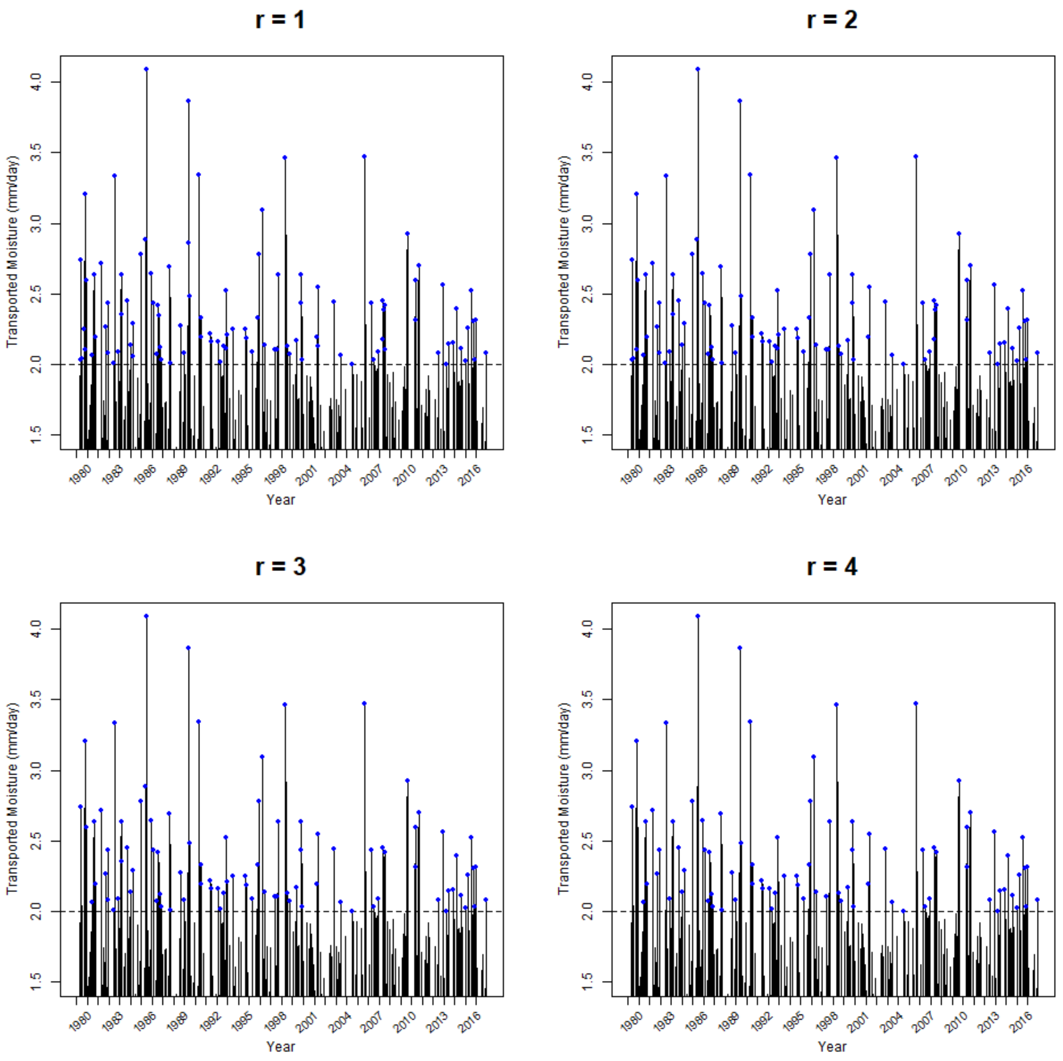


Figure 5.7: Cluster maxima of excesses of the TM series above $u = 2$, performing declustering with *run length* (r) equal to 1,2,3 and 4.

Table 5.3: R^2 of the linear model fitted to the Exponential and to the GPD QQ-Plots for the cluster maxima of excesses of the TM series, taking $u = 2$ and declustering *run length* of $r = 1$,2,3 and 4.

	***r*=1**	***r*=2**	***r*=3**	***r*=4**
Exponential QQ-Plot R^2	0.992	0.991	0.993	0.991
GPD QQ-Plot R^2	0.995	0.994	0.994	0.993

As expected, the values of R^2 for the linear models fitted to the Exponential and to the GPD QQ-Plots are very high and similar. In fact, R^2 is slightly higher for the GPD QQ-Plot than the Exponential one, for all the values of r considered. We will next verify by hypothesis testing if the GPD fits better to the excess data than the Exponential distribution.

Table 5.4 shows the results of the statistical tests that were carried out. Details of these tests can be found in Subsection 2.2.3. For the Cramér-von Mises and Anderson-Darling tests, the null hypothesis is that the cluster maxima resulting of the sample of excesses above $u = 2$ come from a GPD. For all values of r considered, the p-value obtained is larger than 0.5 and consequently the null hypothesis is not

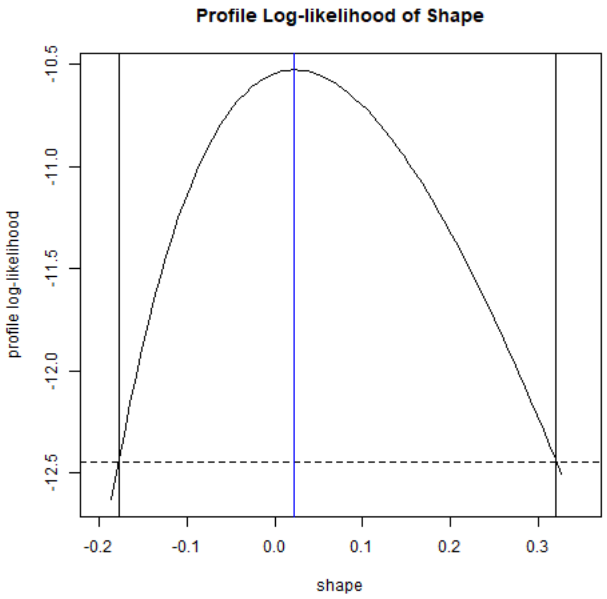


Figure 5.8: *Profile log-likelihood* function for the shape parameter (ξ) of the GPD model fitted to the cluster maxima of excesses of the TM series, considering $u = 2$ and $r = 4$. The black vertical lines are the limits of the 95% confidence interval for ξ, while the blue vertical line is the Maximum Likelihood Estimate. This figure was constructed using the *R* package *evd*.

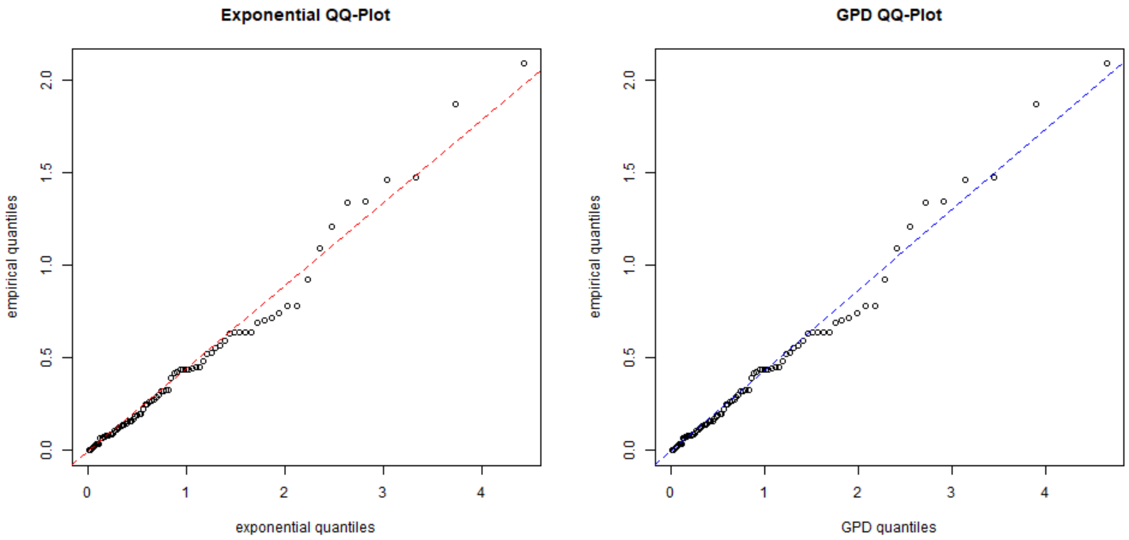


Figure 5.9: Exponential QQ-Plot (left-hand plot) and GPD QQ-Plot (right-hand plot) fitted for the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and *run length* $r = 4$.

rejected for the usual significance levels. Therefore, we can conclude that the GPD model fits well to the data.

The fact that the GPD model fits the declustered excess data does not necessarily mean that it is better than the Exponential model (a particular case of the GPD). Moreover, the analysis presented previously highly supports that possibility. In order to test whether the Exponential fits better to the data than the

GPD, a Likelihood Ratio Test (LRT) was performed, where the null hypothesis is that the GPD model can be reduced to the Exponential distribution ($\xi = 0$). The results of this LRT show that for all values of r considered, the p-value is much higher than the usual significance levels and therefore the null hypothesis is not rejected for those levels. Consequently, it is possible to conclude that the Exponential model is more appropriate to model the cluster maxima of the excesses above $u = 2$.

In order to assess the fit of the Exponential model itself, we used the Lilliefors-corrected Kolmogorov-Smirnov (LcKS) test with null hypothesis being that the cluster maxima of the excesses above $u = 2$ come from an Exponential distribution. For all values of r analyzed, the approximate p-value is much higher than the usual significance levels and therefore the null hypothesis is not rejected. Thus, the conclusion from this LcKS test is that the Exponential model fits well to the data.

Table 5.4: Observed value of the Cramér-von Mises (CvM), Anderson-Darling (AD), Likelihood Ratio Test (LRT) and Lilliefors-corrected Kolmogorov-Smirnov (LcKS) statistics and corresponding p-values for the TM series, with $u = 2$ and declustering *run length* (r) equal to 1, 2, 3 and 4.

	***r*=1**	***r*=2**	***r*=3**	***r*=4**
CvM statistic	0.035	0.036	0.033	0.032
approx. p-value	>0.5	>0.5	>0.5	>0.5
AD statistic	0.222	0.241	0.245	0.246
approx. p-value	>0.5	>0.5	>0.5	>0.5
LRT statistic	0.231	0.214	0.036	0.034
p-value	0.631	0.644	0.850	0.854
LcKS statistic	0.057	0.057	0.049	0.053
approx. p-value	0.755	0.782	0.940	0.921

As a consequence of the statistical tests performed, we will use the Exponential model for the cluster maxima of the excesses above the threshold $u = 2$ of the TM series. The question now is that, as it was observed in Figure 5.3 , there seems to exist a declining trend in the largest values of the series, reflecting non-stationarity as time evolves. In this framework, it is reasonable to allow the scale parameter of the Exponential distribution to vary according time. That corresponds to introduce the year of observation as a covariate. As the scale parameter is always positive, the *log* link function is used. Thus, the expression for the scale parameter of the Exponential model that we fitted, considering the non-stationarity features of the data that were previously mentioned, is as follows:

$$\sigma_t = \exp\{\phi_0 + \phi_1\, t\}\,, \tag{5.1}$$

where $t = Year - 1979$. The purpose of this location change is to enable time to vary between 1, 2, ..., 38.

The results shown in Table 5.5 indicate that, as suspected, the ML estimate of the parameter ϕ_1 is negative for all the values of r considered, what means that the estimate of the scale parameter of the Exponential model is lower in more recent years when compared to the initial period. This decrease in the estimate of the scale parameter with time seems to be more important as r increases. In fact, the estimate of ϕ_1 is "more negative" for those values.

Now the question is if it is worthwhile to use the non-stationary Exponential model compared to the stationary one. As usual in the case of nested models, a Likelihood Ratio Test (LRT) can be used: the null hypothesis of that test in this case is $\phi_1 = 0$ (that is, the stationary model is more appropiate) vs. an alternative, $\phi_1 \neq 0$.

Table 5.5: Maximum Likelihood estimates (with their corresponding standard errors) for the parameters of the non-stationary Exponential model (with scale parameter given by expression (5.1)) fitted to the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and performing declustering with *run length* equal to 1, 2, 3 and 4. Computations were performed using the *R* package *extRemes*.

	***r*=1**	***r*=2**	***r*=3**	***r*=4**
$\hat{\phi}_0$ (Std.Err)	-0.726 (0.180)	-0.705 (0.187)	-0.603 (0.195)	-0.538 (0.211)
$\hat{\phi}_1$ (Std.Err)	-0.014 (0.009)	-0.014 (0.009)	-0.018 (0.009)	-0.020 (0.010)

Table 5.6: Observed value of the LRT statistic and corresponding p-value. This LRT is used for comparing the stationary Exponential model and the non-stationary one (see expression (5.1)), fitted to the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and performing declustering with *run length* (r) equal to 1, 2, 3 and 4, respectively for each column.

$H_0 : \phi_1 = 0$	***r*=1**	***r*=2**	***r*=3**	***r*=4**
LRT statistic	2.375	2.306	3.442	3.926
p-value	0.123	0.129	0.064	0.048

Looking at Table 5.6 , we can see that for $r = 1$ and $r = 2$, the null hypothesis is not rejected at the usual levels of significance, that is, there is not statistical evidence that ϕ_1 is different from 0, so the conclusion is that for those values of r, the stationary Exponential model is more appropriate. For $r = 3$, the situation is borderline, in the sense that the null hypothesis is rejected for $\alpha = 0.10$ but not rejected for $\alpha = 0.05$. So, depending on the level of significance considered, the conclusion is different with respect to the most adequate model. For $r = 4$, the evidence that ϕ_1 is different from 0 is stronger, in the sense that for that value of r the null hypothesis is also rejected for $\alpha = 0.05$, concluding that the non-stationary Exponential model is more appropriate than the stationary one.

Estimating return levels

As it was said in Chapter 2 , estimating return levels is very interesting in applications. Thus, we will end this section by presenting and explaining the results related to this topic regarding the TM series.

In Figure 5.10 it is possible to see the return level plots for the stationary Exponential model for the cluster maxima of the TM series, for $u = 2$ and performing declustering with *run length* (r) equal to 1,2,3 and 4. We can clearly see that, irrespectively of the values of r considered, the results are very similar, in the sense that the empirical return levels match very well the return levels estimated by the model. As expected, the fit gets worse as the empirical return levels increase. Likewise, the width of the confidence intervals for the estimates of the m-observation return levels increases as m increases [1] . Moreover, the confidence intervals tend to become more asymmetric on the right side as m increases, reflecting a higher level of uncertainty associated to large values of TM.

If we take into account the non-stationarity of the excesses above $u = 2$, for each value of $t \in \{1,2,...,38\}$ the corresponding $\widehat{x_m}(t)$ is obtained by using $\widehat{\sigma_t} = \exp\left\{\hat{\phi}_0 + \hat{\phi}_1\, t\right\}$ in expression (2.26). With respect to the estimated return levels considering the non-stationary Exponential model for the cluster maxima of excesses, we will just show the plot corresponding to $r = 4$ (see Figure 5.11). The reason is that 4 is the only value of r (among the ones that we considered) for which the non-stationary model is significantly better than the stationary one (at level of significance $\alpha = 0.05$).

[1] Remember that the m-observation return level (x_m) satisfies $P(X > x_m) = p$, where $p = \dfrac{1}{m}$. That is, x_m is exceeded once in every m observations.

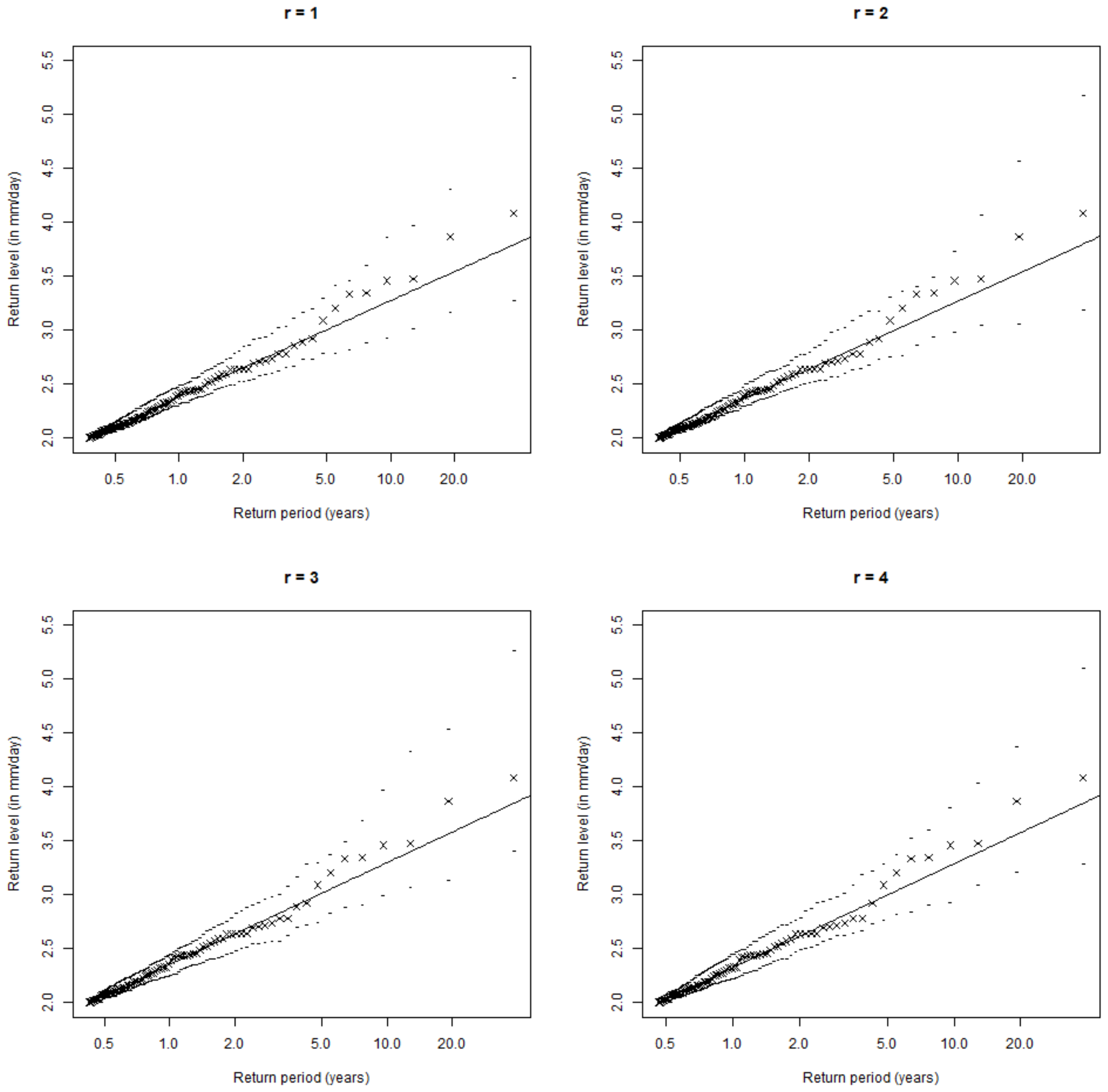


Figure 5.10: Return level plots for the stationary Exponential model for the cluster maxima of excesses of the TM series, choosing threshold $u = 2$ and *run length* (r) equal to 1,2,3 and 4. The solid black line refers to the estimates of the m-observation return level (in mm/day), calculated using expression (2.26) , with $m = 92x$, for x being the corresponding x-coordinate (remember that for each year, the series under study has 92 observations, which is the number of days of June, July and August). The dashed lines refer to the simulated 95% confidence intervals for the estimates of the return levels, while the crosses represent the empirical return levels. Computations were performed using the *R* package *evd*.

As we can see in Figure 5.11 , at the beginning of the period considered by the TM series (summer 1980), the estimated m-observation return levels are higher than at the end of that period (summer 2017), for $m = 92 \times 38$, $m = 92 \times 50$ and $m = 92 \times 100$ (remember that there are 92 observations per year in the TM series, corresponding to the daily observations of June, July and August). It is also interesting to highlight that the differences between the estimated return levels get smaller over time. We extracted from Figure 5.11 the values of the estimated return levels in the first summer of the TM series and in the last summer of that series (see Table 5.7).

It is possible to see in Table 5.7 that the ratio between the estimated 38-year return level for the last

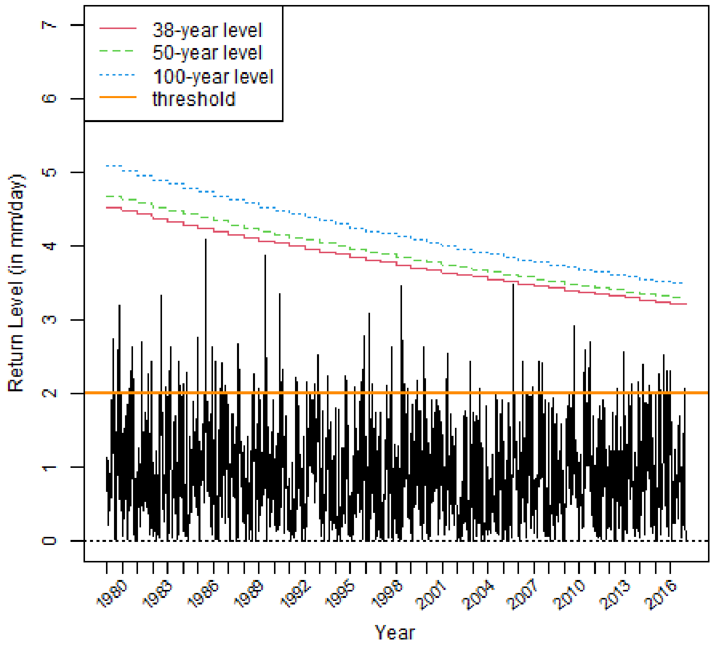


Figure 5.11: Estimated return levels for the TM series, considering a non-stationary Exponential model for the cluster maxima of excesses above the threshold $u = 2$ (orange line), having performed declustering with *run length* (r) equal to 4. The red line refers to the 38-year return level (38 years is the length of the period considered in the TM series: 1980-2017), while the green line corresponds to the 50-year return level and the blue line to the 100-year return level. "Year" should be understood as "summer", as the TM series has 92 observations per year, corresponding to the daily observations of June, July and August. Computations were performed using the *R* package *extRemes*.

Table 5.7: Estimated 38-year, 50-year and 100-year return levels for the TM series in the first and the last summer of the period considered (summers of 1980 and 2017, respectively) , using a non-stationary Exponential model for the cluster maxima of excesses above the threshold $u = 2$, having performed declustering with *run length* (r) equal to 4. The values in this table are expressed in mm/day. They are extracted from Figure 5.11 .

	38-year return level	50-year return level	100-year return level
Summer 1980	4.531	4.688	5.085
Summer 2017	3.228	3.304	3.497

summer of the TM series and the first summer of that series is approximately equal to 0.712 (representing a decrease of approximately 28.8% in the estimated 38-year return level from the beginning to the end of the series). In the case of the 50-year return level, the ratio mentioned before is approximately equal to 0.705 (decrease of approximately 29.5% in the estimated 50-year return level); and in the case of the 100-year return level, the ratio equals approximately 0.688 (decrease of approximately 31.2% in the estimated 100-year return level). Thus, as it is obvious, the interpretation of the results of this table is in line with the interpretation of Figure 5.11 . Moreover, the other comment we made on that figure can also be checked in Table 5.7, since the difference between the estimated 100-year return level and the 38-year return level is approximately 0.554 for the first summer of the TM series and approximately 0.269 for the last summer of that series. That is, over the period under study, the difference between those estimated return levels has approximately decreased 51.4% of the value corresponding to the first summer.

From these results, and provided the atmospheric conditions evolve in the current manner, it is possible to say that we expect to observe a persistent decrease in the extreme values of TM as time goes by (see Figure 5.11).

All in all, we can conclude that $r = 4$ is the most appropriate choice for the *run length* of the declustering scheme used for the excesses over the threshold $u = 2$ of the TM series. The reason is that, among the values of r considered, $r = 4$ is the one that best guarantees the independence between the excesses and, moreover, it is the only choice for which the non-stationarity of the cluster maxima is detected.

5.2 Bivariate analysis of Precipitation and "-omega"

In Chapter 1 , we introduced the series of **precipitation** (measured in mm/day) in the GPLLJ sink region and the series of tropospheric stability in that region (**omega**, measured in Pa/s). As it was already referred, these series consist of 3496 observations, corresponding to the daily observations of the summer months (June, July and August) of the period 1980-2017. Now, interest focuses on studying the extremal dependence between precipitation and "-omega" (the sign of "omega" is reversed because the meteorological interest lies on studying the joint behaviour of the upper tail of precipitation and the lower tail of "omega"). In fact, our study consists in analyzing the bivariate extremes of precipitation and "-omega" for two subsamples of the series: for the days when the transported moisture from the GPLLJ source region to the jet domain (TM series, analyzed in depth in Section 5.1) is high and when it is low. Thus, one subsample consists of the days with the 25% lowest values of TM, whereas the other one includes the 25% highest values of that variable (consequently, each subsample includes 874 observations). It is important to mention that the TM series was lagged 1 day with respect to the series of precipitation and "-omega", that is, for example, for an observed pair of (-omega,precipitation) occurring on 2 June 1980, the corresponding value of TM is the one that occurred on 1 June 1980. The reason for doing so is meteorological: precipitation and "-omega" are observed in the GPLLJ sink region, while the TM is computed on its way from the source region to the jet domain. Hence, the moisture arrives at the sink region (approximately) 1 day after it is observed, and that is why the adjustment that we carried out was necessary.

5.2.1 Preliminary analysis

Remember that in our bivariate analysis, we do not use all the observed sample, but two subsamples, one corresponding to the days with high TM and the other one to the lowest values of TM. However, we thought it would be interesting to first briefly analyze each of the variables as a whole. The plots of the complete series of "-omega" and precipitation, as well as some summary statistics, can be found in Figure 5.12 and Table 5.8, respectively. From this information, we can see that "-omega" is a variable that takes positive and negative values, being the mean and the median slightly positive and very similar. For the precipitation series, taking into account that it corresponds to the precipitation integrated in the whole moisture sink region, it is reasonable that there is no day with precipitation equal to 0. It is obvious that precipitation can only take non-negative values and, in this series, the mean is bigger than the median, as there are large values that "push" the mean to the right.

Now, let us focus on the subsamples of interest: the observed pairs (-omega,precipitation) for the days with low TM and for those corresponding to high TM.

As we can see in Figure 5.13 , the boxplots show that there are higher extreme values of

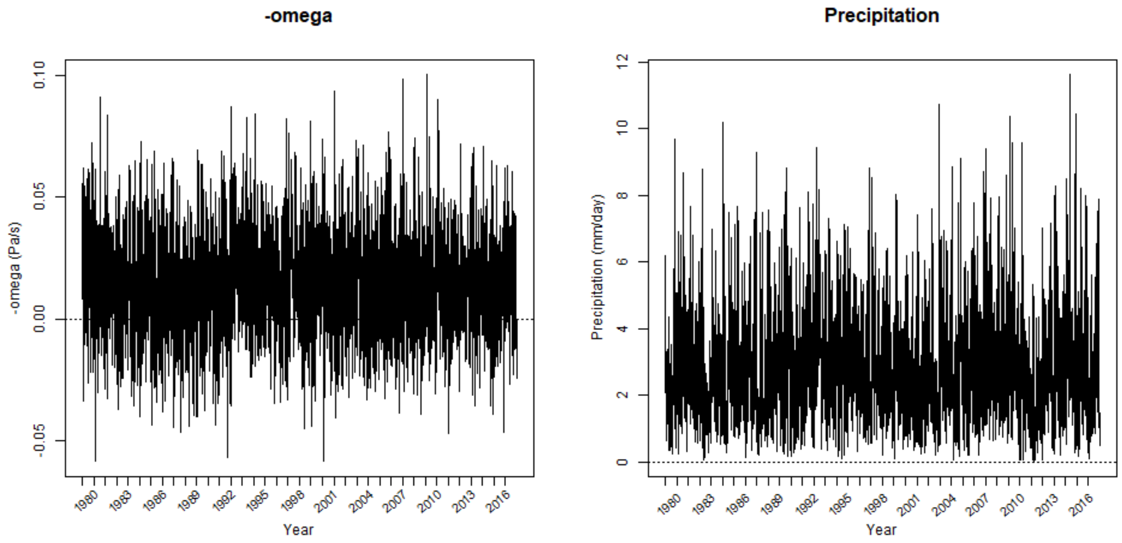


Figure 5.12: Plots of the series of "-omega" (on the left) and precipitation (on the right). They comprise daily observations of the summer months (June, July, August) from 1980 to 2017 (38 summers: 3496 observations).

Table 5.8: Summary statistics for the "-omega" and precipitation series. n denotes the number of observations of the series; $x_{1:n}$ is the minimum and $x_{n:n}$ is the maximum of the values of the series; $\bar{x}$ refers to the mean and sd to the standard deviation. $Q_{0.25}$, $Q_{0.50}$ and $Q_{0.75}$ are the first quartile, the median and the third quartile of the data, respectively.

	n	$x_{1:n}$	$Q_{0.25}$	$Q_{0.50}$	$\bar{x}$	$Q_{0.75}$	$x_{n:n}$	sd
-omega (Pa/s)	3496	-0.058	-0.002	0.013	0.014	0.029	0.100	0.022
prec. (mm/day)	3496	0.026	1.473	2.556	2.814	3.785	11.609	1.737

"-omega" when the TM is low than when it is high. In contrast, there are higher extreme values of precipitation when the TM is high than when it is low. Looking at the first quartile, the median and the third quartile, these quantities are also higher for high TM than low TM in the case of precipitation, and the same can be said for "-omega" (contrarily to what occurs in the extreme values of that variable).

In Figure 5.14 , the reasoning with respect to the extreme values is analogous to the one that we made when commenting Figure 5.13 . Moreover, we observe that the estimate of the density of "-omega" is more symmetrical than the one corresponding to precipitation in both cases (low and high TM). Additionally, as it can be confirmed in Table 5.9 , it is possible to see that "-omega" is slightly more positively skewed and has higher kurtosis when there is low TM than when the TM is high. With regard to precipitation, it is more positively skewed and has higher kurtosis when there is high TM than when the TM is low. Finally, with respect to the standard deviation, in the case of "-omega", there is not much difference between low TM and high TM. In the case of precipitation, the standard deviation is a little higher when the TM is high than when it is low.

5.2.2 Fitting the Bivariate Threshold Excess Model

In this thesis, we will fit the Bivariate Threshold Excess Model (see Section 3.4) for the bivariate extremes of "-omega" and precipitation in two cases: when the TM is low and when it is high. Before doing that, it is necessary to fit to each margin an univariate threshold model to the excesses over an

Table 5.9: Standard deviation (*sd*), skewness and Pearson's measure of kurtosis for the "-omega" and precipitation series in the cases of low TM and high TM. The *R* package *moments* was used for computing the values of skewness and kurtosis.

	-omega		**precipitation**	
	Low TM	High TM	Low TM	High TM
sd	0.023	0.021	1.595	1.781
skewness	0.385	0.112	0.892	1.028
kurtosis	3.358	2.689	3.674	4.306

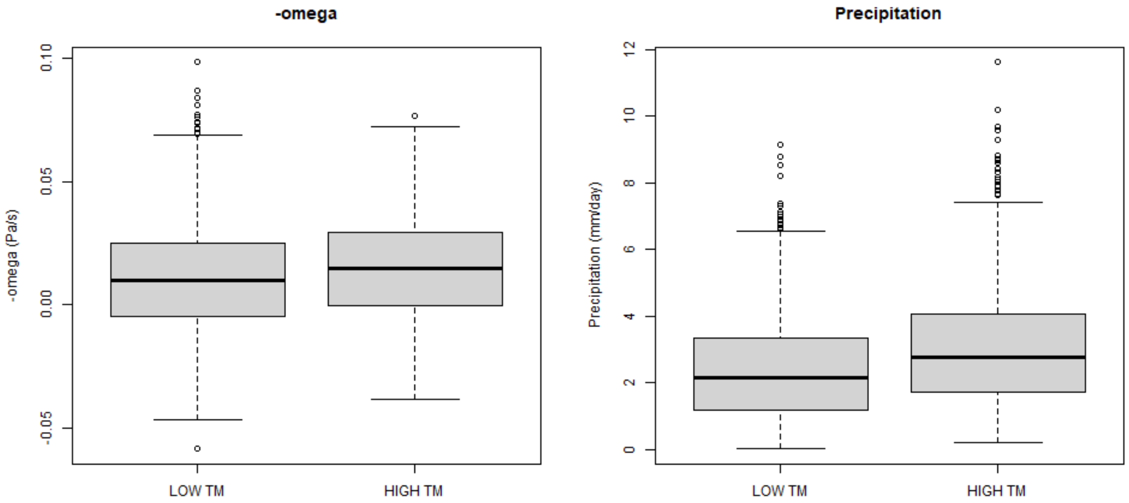


Figure 5.13: Boxplots of "-omega" (on the left) and precipitation (on the right) for low TM and high TM. For each variable, the boxplot corresponding to low TM is on the left and the one corresponding to high TM is on the right.

appropriate threshold.

Univariate threshold models for the margins

Applying the two methods presented in Subsection 2.2.2 , we came to the conclusion that $u_1 = 0.03$ is a suitable threshold for "-omega" and $u_2 = 5.2$ is an adequate threshold for the precipitation, for both low TM and high TM.

With respect to the selection of the threshold for "-omega" in the two situations considered, the plots that helped us to make a decision can be found in Figures 5.15 and 5.16 . In Figure 5.15 , it can be seen that the estimated mean excess function for both low TM and high TM is consistent with the choice of $u_1 = 0.03$ as a suitable threshold, since a linearity pattern to the right of that value is clearly visible (see the solid blue line in the two plots of that figure). Looking at Figure 5.16 , the ML estimates for the shape and modified scale parameters are approximately constant above $u_1 = 0.03$ for both low TM and high TM, so this figure also suggests that $u_1 = 0.03$ is an appropriate threshold.

Regarding the threshold selection for the precipitation, Figures 5.17 and 5.18 show , for the cases of low TM and high TM, the estimated mean excess functions and the ML estimates for the shape and modified scale parameters as a function of a set of candidate thresholds. Looking at those figures and reasoning as in the case of "-omega", it was decided that $u_2 = 5.2$ is a proper threshold for precipitation

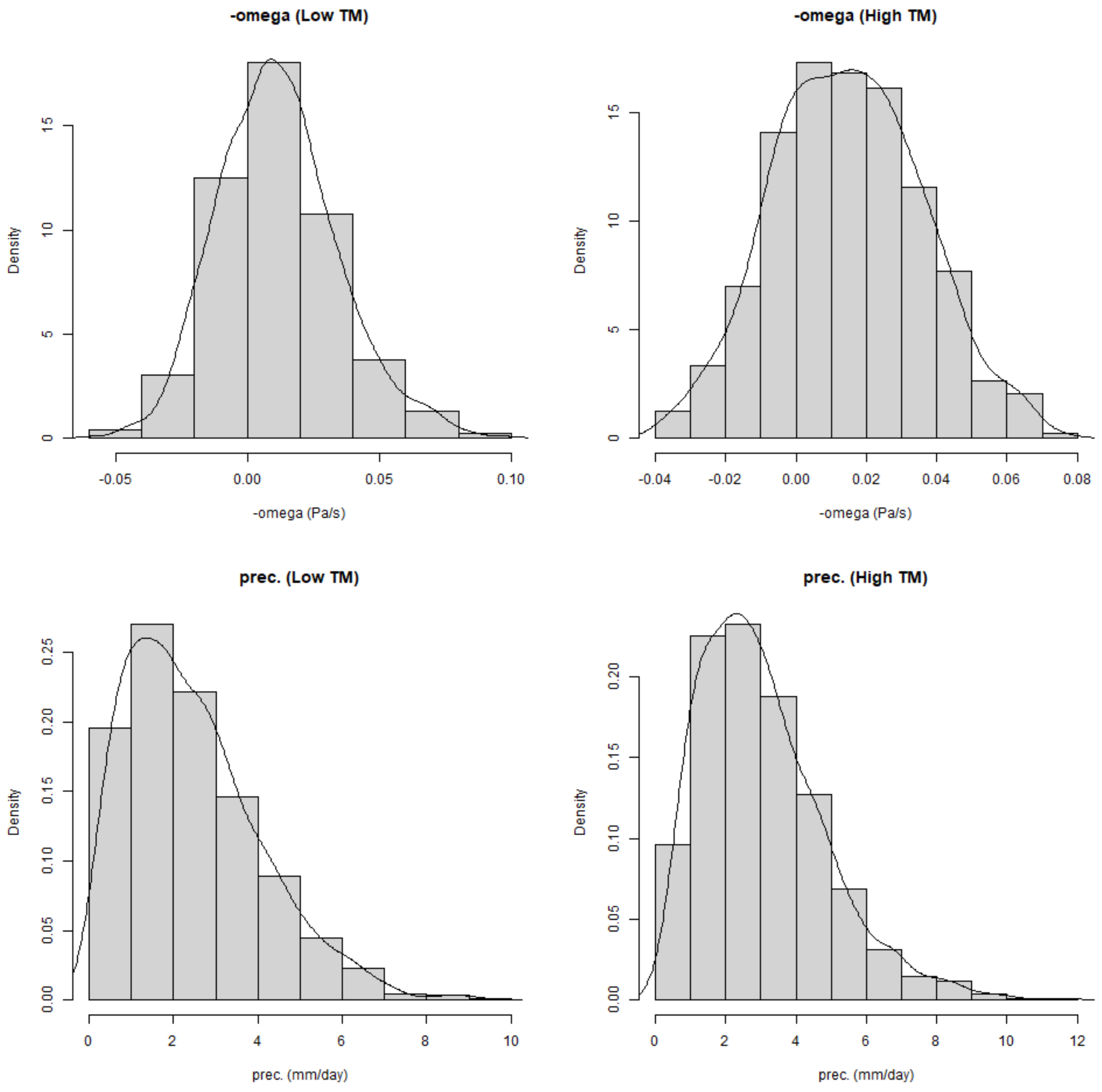


Figure 5.14: Histograms and kernel density estimates of "-omega" and precipitation for low TM and high TM. The first row of plots corresponds to the histograms of "-omega" and the second one to those of precipitation. For each row, the histogram on the left corresponds to the days with low TM and the one on the right corresponds to the days with high TM.

for both low and high TM.

Having chosen threshold $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation for low TM and high TM, the results of the POT analysis performed using the *R* package *evd* can be found in Table 5.10 . As we can see in that table, the ML estimates for the shape parameter of the GPD are all negative and larger than -0.5, which guarantees the asymptotic properties of ML estimation (de Zea Bermudez and Kotz, 2010a). It is also important to mention that for the four situations, the right endpoint of the corresponding variable (computed using (2.13)) is finite and greater than the sample maximum of each case, which is an indispensable condition for accepting the GPD model (note that this condition is only required when the estimate of the shape parameter is negative).

In Figure 5.19 it is possible to visualize the *profile log-likelihood* 95% confidence intervals of ξ for "-omega" and precipitation in the cases of low and high TM. As it can be seen in that figure, the value

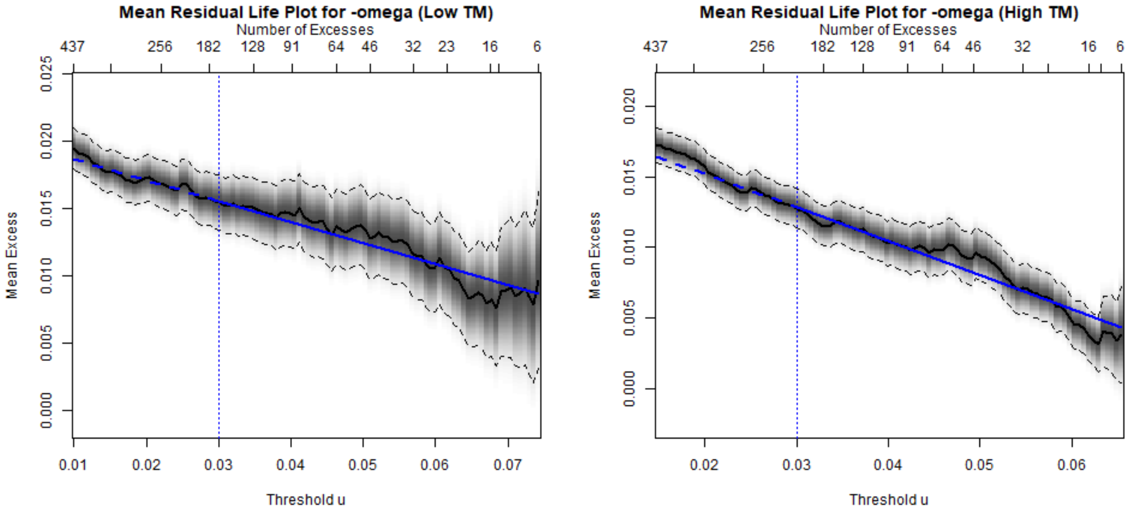


Figure 5.15: Estimated Mean Excess Function of "-omega" for low TM (left plot) and high TM (right plot). The estimated MEF is represented by a solid black line, with 95% normal-approximation confidence intervals as black dashed lines and fitted solid blue line to the right of $u_1 = 0.03$, in both the cases of low and high TM. These plots were constructed using the *R* package *evmix*.

Table 5.10: Results of the POT analysis for the "-omega" and precipitation series in the cases of low TM and high TM, considering thresholds $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation. Apart from the number and percentage of excesses, this table includes the ML estimates for the parameters of the GPD model fitted to the excesses above the corresponding threshold (as usual, ξ denotes the shape parameter and σ denotes the scale parameter), together with their standard errors. Moreover, the right endpoint of each variable is estimated in each case (it is denoted by $\widehat{x^F}$) and the sample maximum ($x_{n:n}$) is also shown. Computations were performed using the *R* package *evd*.

	-omega(low TM)	prec. (low TM)	-omega(high TM)	prec. (high TM)
Threshold	0.03	5.2	0.03	5.2
Number of excesses	170	55	211	98
Percentage of excesses	19.5 %	6.3 %	24.1 %	11.2 %
$\hat{\xi}$ (Std.Err)	-0.180 (0.072)	-0.160 (0.128)	-0.311 (0.059)	-0.163 (0.087)
$\hat{\sigma}$ (Std.Err)	0.018 (0.002)	1.185 (0.219)	0.017 (0.001)	1.676 (0.222)
$\widehat{x^F}$	0.132	12.583	0.084	15.499
$x_{n:n}$	0.098	9.134	0.077	11.609

0 is contained in the confidence interval of ξ for precipitation, which suggests that the one-parameter Exponential model would be more appropriate than the GPD, which has two parameters (see expression (2.10)). Using the *R* package *evd*, the 95% confidence intervals for ξ obtained in terms of the precipitation are $(-0.382, 0.153)$ when TM is low and $(-0.300, 0.055)$ when TM is high. Regarding the confidence intervals of ξ for "-omega", they only contain negative values, so thay are consistent with the fact that the GPD model is more appropriate than the Exponential one in those cases. The obtained 95% confidence interval of ξ for "-omega" are $(-0.293, -0.013)$ when TM is low and $(-0.381, -0.178)$ when TM is high.

In Figure 5.20 , looking at the Exponential and GPD QQ-Plots, it is possible to see that, with regard to "-omega" and for both cases of low and high TM, the GPD model seems to be more appropriate for

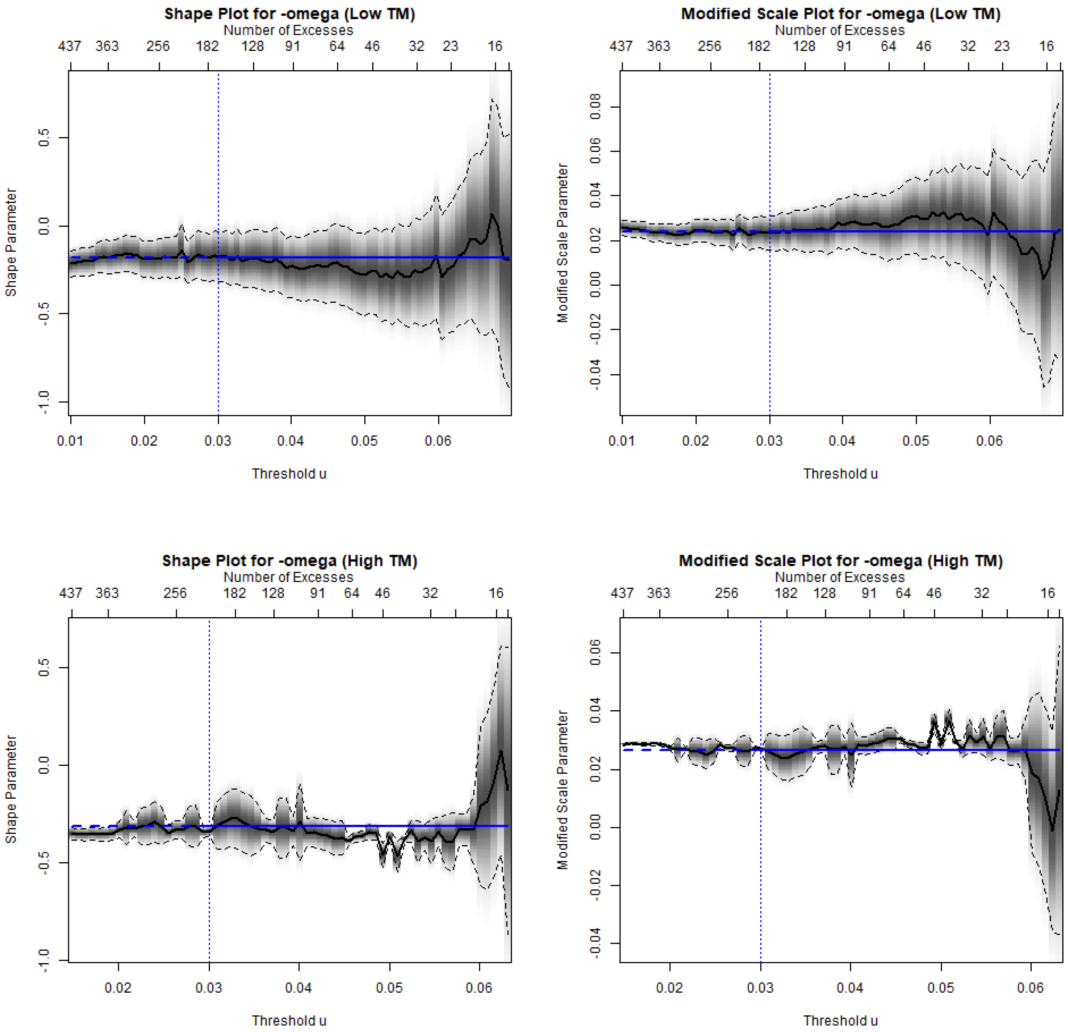


Figure 5.16: Maximum Likelihood estimates for the shape and modified scale parameters of the GPD models fitted to the "-omega" series (for the cases of low TM and high TM), as a function of the chosen threshold u. The first row of plots corresponds to the case of low TM, whereas the second row corresponds to the case of high TM. Within each row, the shape plot is on the left and the modified scale plot is on the right. The ML estimates are presented as a solid black line and the corresponding 95% normal-approximation confidence intervals as black dashed lines. For both the cases of low and high TM, a solid blue horizontal line is plotted to the right of $u_1 = 0.03$, indicating the corresponding ML estimates for the shape and modified scale parameters of the GPD model fitted when $u_1 = 0.03$, respectively. These plots were constructed using the *R* package *evmix*.

modelling the excesses above the threshold $u_1 = 0.03$. This can be confirmed by the R^2 of the linear models fitted to the pairs of empirical and theoretical quantiles (see Table 5.11) : looking at "-omega" for both the cases of low and high TM, the R^2 is slightly larger for the GPD than for the Exponential models, although both extremely close to 1.

In contrast, with respect to precipitation, it seems that the Exponential model is somewhat better than the GPD one in the low TM case (see Figure 5.21). It can also be seen in Table 5.11 that the R^2 of the linear model associated to the Exponential QQ-Plot is slightly greater than the one corresponding to the GPD QQ-Plot in that situation. For the high TM, the two models are very much alike and thus either could be chosen. However, we should always take into account the principle of parsimony,

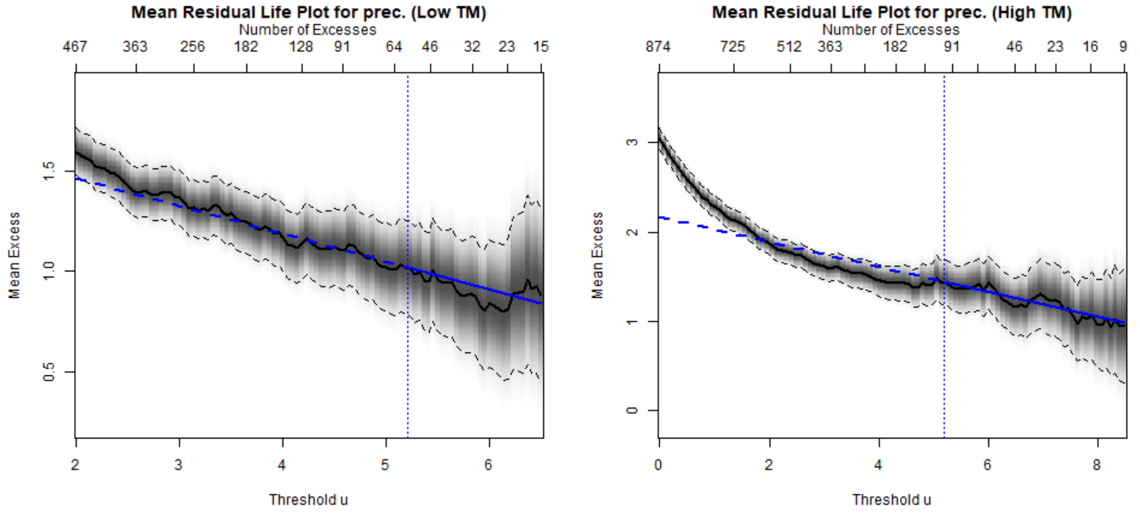


Figure 5.17: Estimated Mean Excess Function of precipitation for low TM (left plot) and high TM (right plot). The estimated MEF is represented by a solid black line, with 95% normal-approximation confidence intervals as black dashed lines and fitted solid blue line to the right of $u_2 = 5.2$, in both the cases of low and high TM. These plots were constructed using the *R* package *evmix*.

which points towards the Exponential model. With respect to the R^2 in the high TM case, the values of R^2 corresponding to the linear models fitted to the Exponential and GPD QQ-Plots practically coincide.

Table 5.11: R^2 of the linear models associated with the Exponential and GPD QQ-Plots for the excesses above the threshold $u_1 = 0.03$ for "-omega" and the excesses above the threshold $u_2 = 5.2$ for precipitation, for low TM and high TM.

	-omega(low TM)	prec. (low TM)	-omega (high TM)	prec. (high TM)
Exp. QQ-Plot R^2	0.991	0.993	0.975	0.995
GPD QQ-Plot R^2	0.998	0.989	0.998	0.996

In Table 5.12 it is possible to see the results of the statistical tests that were performed. The details about these tests can be found in Subsection 2.2.3 . For the Cramér-von Mises and Anderson-Darling tests, the null hypothesis is that the excesses above the chosen threshold come from a GPD. For "-omega" and precipitation in both cases of low and high TM, the p-value obtained is larger than 0.5. Therefore, the null hypothesis is not rejected for the usual significance levels and we can conclude that the GPD model fits well to the data in the situations considered.

As usual, a Likelihood Ratio Test is carried out to see if the GPD model is significantly better than the Exponential one, being $H_0 : \xi = 0$ and $H_1 : \xi \neq 0$. The results of this test show that for "-omega", both for low and high TM, the p-values are lower than 0.05. So, at the level of significance 0.05, the null hypothesis is rejected in both cases, which enables us to conclude that the GPD model is more appropriate than the Exponential one. For precipitation, the p-value is higher than the usual significance levels for both cases of low and high TM and consequently the null hypothesis is not rejected at those significance levels. Thus, it is possible to conclude that the Exponential model is more adequate than the GPD one.

Table 5.13 contains the information regarding the Exponential model that we fitted to the excess

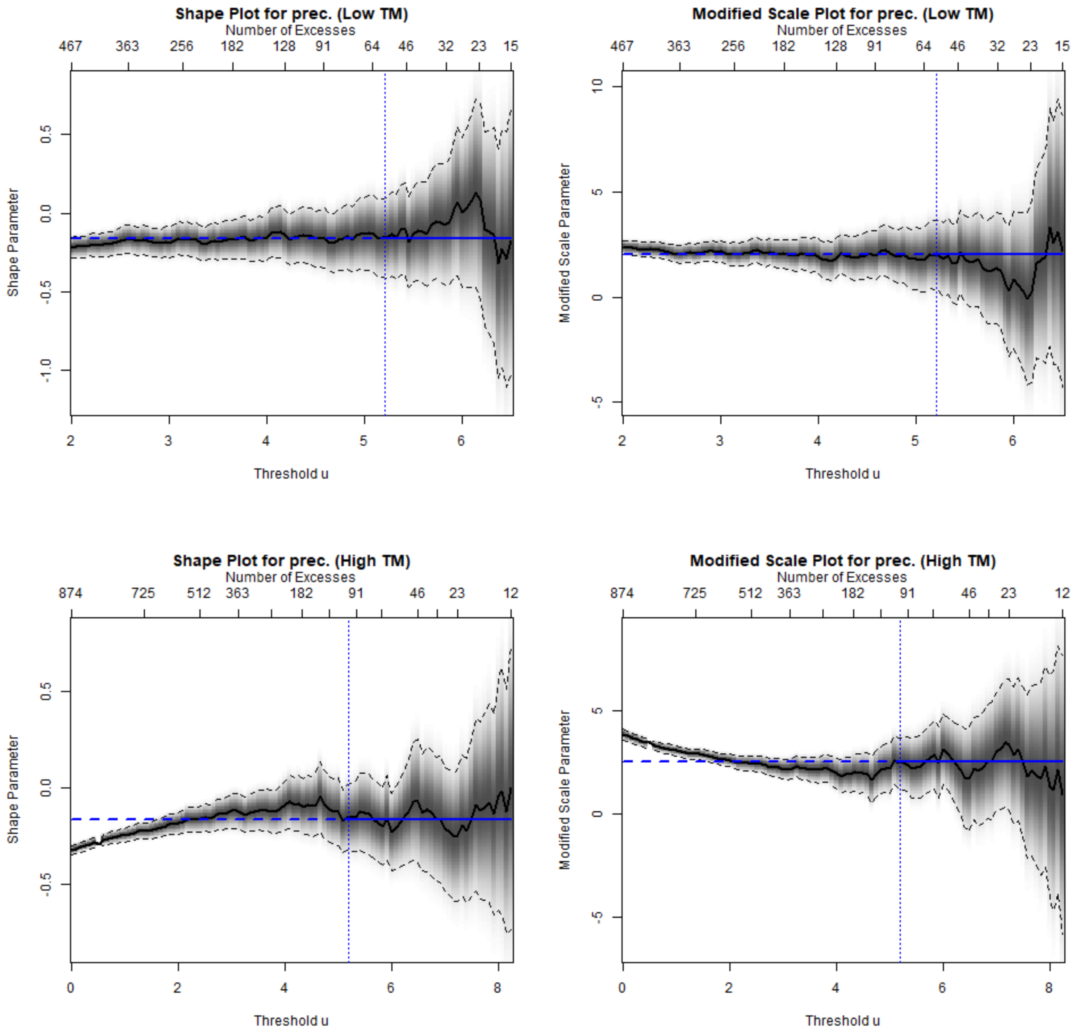


Figure 5.18: Maximum Likelihood estimates for the shape and modified scale parameters of the GPD models fitted to the precipitation series (for low and high TM), as a function of the chosen threshold u. The top row corresponds to the case of low TM, whereas the bottom row refers to the case of high TM. Within each row, the shape plot is on the left and the modified scale plot is on the right. The ML estimates are presented as a solid black line and the corresponding 95% normal-approximation confidence intervals as black dashed lines. For both cases of low and high TM, a solid blue horizontal line is plotted to the right of $u_2 = 5.2$, indicating the corresponding ML estimates for the shape and modified scale parameters of the GPD model fitted to the excess data, respectively. These plots were constructed using the *R* package *evmix*.

data in each of the cases. Apart from the corresponding ML estimate for the scale parameter (with its standard error), it is possible to see the observed value of the test statistic and the approximate p-value for the Lilliefors-corrected Kolmogorov-Smirnov (LcKS) test, which was used to assess if the Exponential distribution fits well to the data. As it can be seen in the referred table, in both cases of low and high TM, the approximate p-values are larger than the usual significance levels. This means that, in each case, the null hypothesis that the corresponding excesses come from an Exponential distribution is not rejected at the usual significance levels. Therefore, we can conclude that the Exponential model fits well to the excesses above $u_2 = 5.2$ in both cases of precipitation with low and high TM.

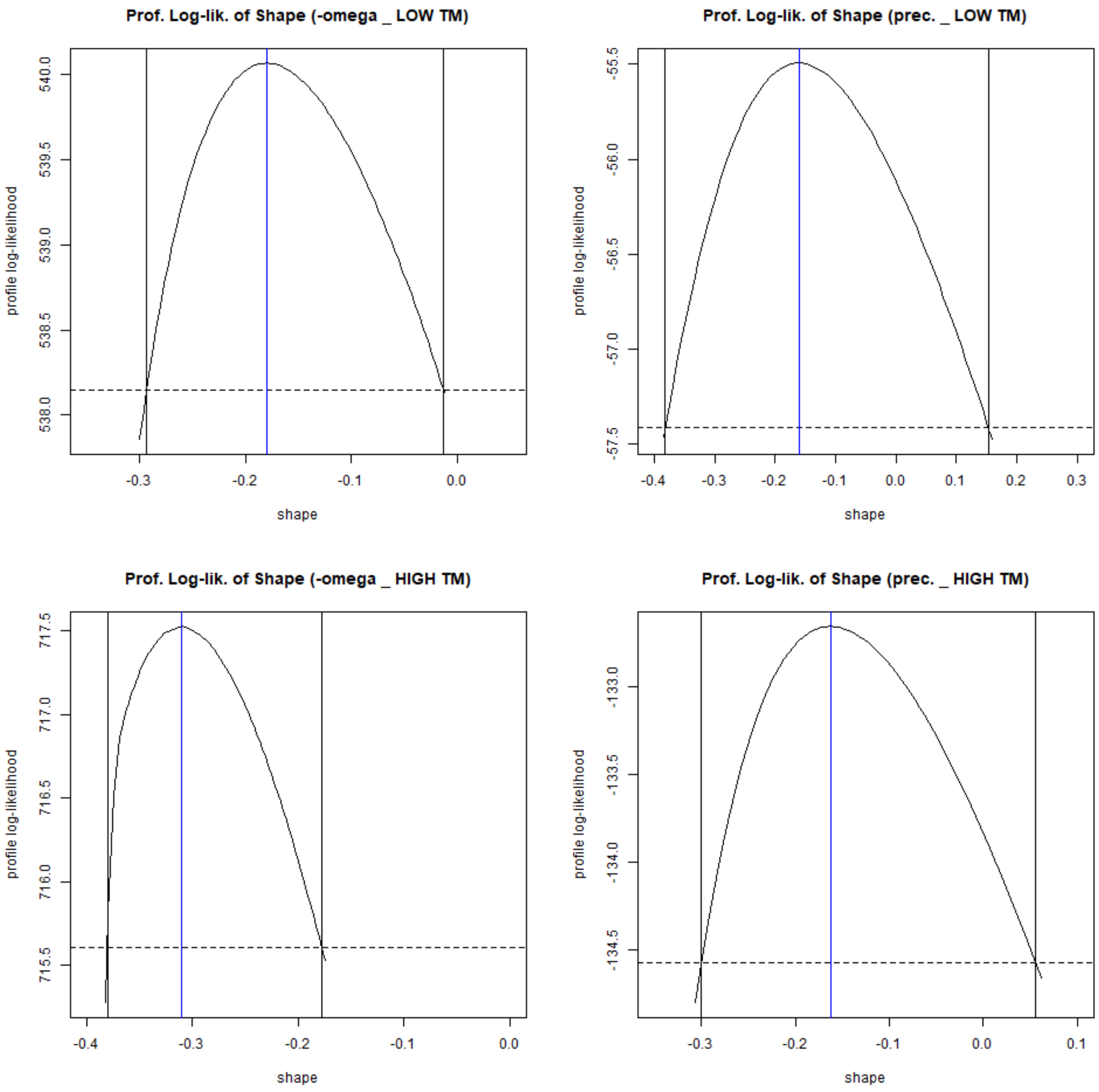


Figure 5.19: *Profile log-likelihood* functions for the shape parameter of the GPD model fitted to the "-omega" and precipitation series in the cases of low TM and high TM, considering $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation. The plots presented in the top row correspond to the low TM case, whereas the ones presented in the bottom row correspond to high TM. Within each row, the variable on the left is "-omega" and the one on the right is precipitation. The black vertical lines are the limits of the 95% confidence interval for ξ, while the blue vertical line is the Maximum Likelihood Estimate. This figure was constructed using the *R* package *evd*.

Bivariate analysis

We concluded that the GPD model is more appropriate for modelling the excesses over the threshold of "-omega" (both for low and high TM) and the Exponential model is better for the excess data of the variable precipitation (in both cases of low and high TM). Therefore, we consider these models for the marginal distributions in order to fit a Bivariate Threshold Excess Model to the pair of variables (-omega,precipitation). We used the *R* package *evd* for doing so.

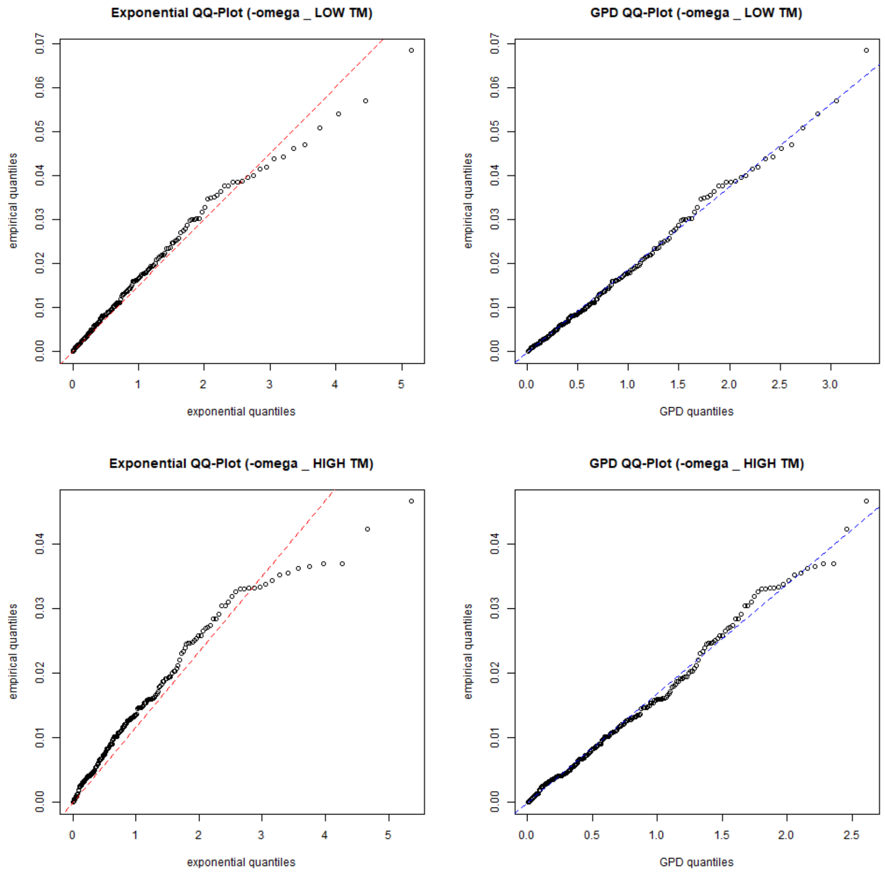


Figure 5.20: Exponential and GPD QQ-Plots fitted for the excesses above the threshold $u_1 = 0.03$ for "-omega" in both the cases of low TM and high TM. The top row corresponds to the case of low TM, whereas the bottom one corresponds to high TM. Within each row, the Exponential QQ-Plot is presented on the left and the GPD QQ-Plot on the right.

Table 5.12: Observed values of the Cramér-von Mises , Anderson-Darling and Likelihood Ratio Test statistics and corresponding p-values for the "-omega" and precipitation series, for the thresholds $u_1 = 0.03$ for "-omega" and $u_2 = 5.2$ for precipitation.

	-omega(low TM)	prec. (low TM)	-omega(high TM)	prec. (high TM)
CvM statistic	0.024	0.043	0.060	0.036
approx. p-value	>0.5	>0.5	>0.5	>0.5
AD statistic	0.173	0.306	0.427	0.219
approx. p-value	>0.5	>0.5	>0.5	>0.5
LRT statistic	4.344	1.24	15.503	2.369
p-value	0.037	0.265	≈ 0	0.124

In Figure 5.22 we represent the observed points of (-omega,precipitation), for high and low TM, along with two lines representing the thresholds ($u_1 = 0.03$, $u_2 = 5.2$). This enables the definition of three "extremal" quadrants as follows:

A - large values only in "-omega".

B - large values only in precipitation.

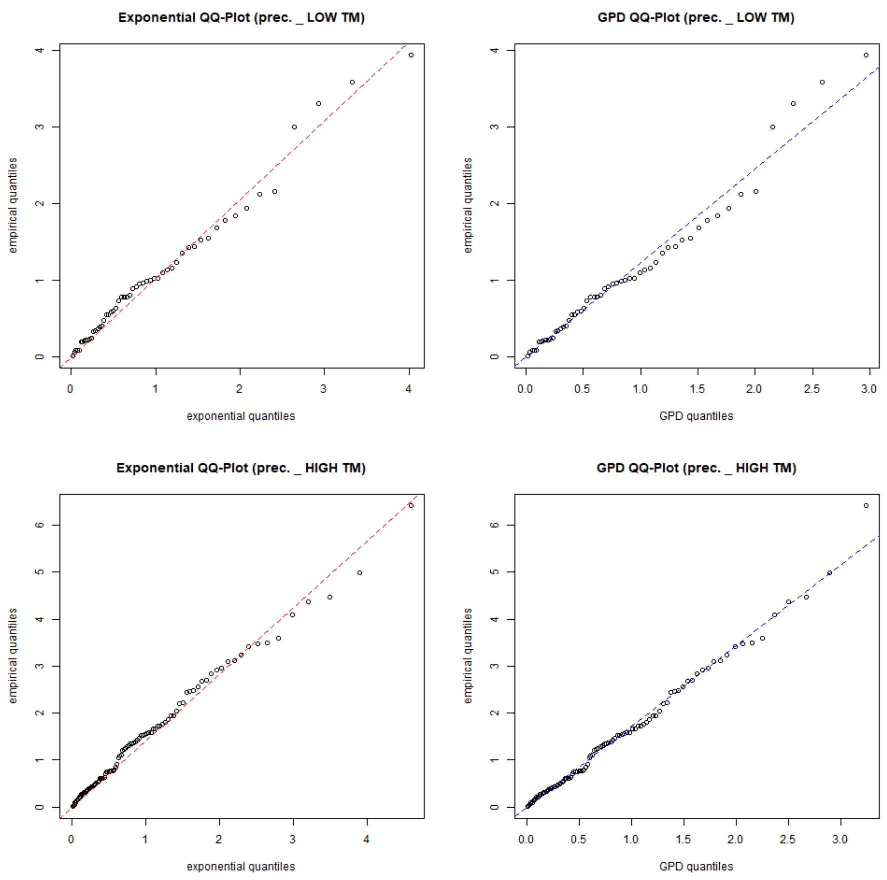


Figure 5.21: Exponential and GPD QQ-Plots fitted for the excesses above the threshold $u_2 = 5.2$ for precipitation in both the cases of low TM and high TM. The top row corresponds to the case of low TM, whereas the bottom one corresponds to high TM. Within each row, the Exponential QQ-Plot is presented on the left and the GPD QQ-Plot on the right.

Table 5.13: ML estimate for the scale parameter of the Exponential distribution ($\widehat{\sigma_{EXP}}$) and observed value of the LcKS statistic (with the approximate p-value of that test) for the precipitation series (in the cases of low and high TM), choosing threshold $u_2 = 5.2$ in both cases. In each case, the standard error of $\widehat{\sigma_{EXP}}$ is also shown.

	prec. (low TM)	prec. (high TM)
$\widehat{\sigma_{EXP}}$ (Std.Err)	1.020 (0.138)	1.442 (0.146)
LcKS statistic	0.099	0.076
approx. p-value	0.413	0.376

C - large values in both variables.

In the plots we also indicate the number of points belonging to each of the quadrants and the corresponding percentage in terms of the total sample size. It should be mentioned that the largest difference in the percentages is observed in quadrant C (which reflects the situation of extremes in both variables).

In Figure 5.23 it is possible to see the *chi plots* and *chi bar plots* for (-omega,precipitation) in the cases of low TM and high TM. As it can be seen in that figure, looking firstly at the *chi plots*, the empirical estimates of $\chi(u)$ are greater than 0 in both cases of low and high TM for the values of u close to 1, so it is consistent with $\chi > 0$ in both cases, and consequently with the fact that "-omega" and precipitation are **asymptotically dependent**. Therefore, we can assume that the models that we present in this thesis are appropriate for these variables. In the case of high TM, the idea that the variables under study are asymptotically dependent is reinforced by the *chi bar plots*, since the 95% confidence intervals for $\bar{\chi}(u)$ as u increases contain the value 1, so $\bar{\chi} = 1$ would be feasible. For low TM, the confidence intervals corresponding to the values of u close to 1 do not contain the value 1, but the empirical estimates of $\bar{\chi}(u)$ increase as $u \in (0,1)$ increases, as desired. The *chi plots* suggest that the extremal dependence is stronger in the case of high TM than when TM is low, since the empirical estimates of $\chi(u)$, as $u \to 1$, are larger for high TM. That is consistent with a larger value of χ in the case of high TM.

Taking into account (3.26), the joint distribution function of (X_1, X_2) = (-omega,precipitation) in the cases of low TM and high TM can be approximated by one of the parametric models of Section 3.2 within the region $x_1 > u_1,\ x_2 > u_2$, where $u_1 = 0.03$ and $u_2 = 5.2$ in both cases. In order to estimate the parameters of the models, the one-step censored-likelihood method was carried out using the *R* package *evd*. That is, the likelihood function in (3.28) is a function of all parameters (both the marginal and the dependence ones), so by maximizing that joint likelihood function it is possible to obtain simultaneously the ML estimates for the marginal and dependence parameters.

In Table 5.14 it is possible to see the results that we obtained when fitting the parametric models of Section 3.2 to the pair (-omega,precipitation) in the cases of low and high TM within the region $x_1 > 0.03,\ x_2 > 5.2$ in both cases.

The AIC value of a model with k parameters is given by $2k - 2\log(L)$, where L represents the maximized likelihood for the model. The AIC is used for comparing models. According to this criterium, the model which should be selected is the one that has the lowest AIC. The expression of AIC penalizes complexity, and as such, tends to point towards the model with the least number of parameters (in accordance with the parsimony principle); see Akaike (1974). In Table 5.14 , in each case, the model with the lowest AIC appears in bold: for low TM, the best model is the Bilogistic one (AIC=270.826) and for high TM, the Logistic one (AIC=311.341). It is important to mention that, within each case, the values of AIC are quite similar between them, which means that there is not a specific model that we can say that is much better than the others.

With respect to the coefficient *Dependence*, as it was explained in Section 3.1 , it is defined as $2(1 - A(1/2))$, where $A(.)$ is the corresponding Pickands dependence function, and should be interpreted as follows: independence corresponds to $Dependence = 0$ and perfect dependence to $Dependence = 1$; the strength of dependence increases as *Dependence* increases [2] . In our analysis, as it can be seen in Table 5.14 , the value of *Dependence* is larger for high TM than for low TM for all the parametric models considered, which means that the extremal dependence between "-omega" and precipitation is stronger in the case of high TM than when there is low TM.

Therefore, we choose the Bilogistic (3.20) as the parametric model for the joint distribution function of (-omega,precipitation) within the region $x_1 > 0.03,\ x_2 > 5.2$ in the case of low TM, and the Logistic model (3.15) in the case of high TM, within the same region. The ML estimates for the marginal and dependence parameters of those models, as well as the corresponding standard errors, can be found in Table 5.15 . Apart from those results, it is important to mention that a score test following the methodology in Tawn (1988) was also used in order to test $H_0 : \alpha = 1$ *vs.* $H_1 : \alpha < 1$, where α is the dependence parameter of the Logistic model. In the *R* package *evd*, this test is only implemented for the Logistic model, that is why it was not performed for the low TM case. Applying this test to the high TM case, a p-value smaller than 2.2×10^{-16} is obtained, which implies that, at all usual significance levels, the null hypothesis is rejected, so we can conclude that $\alpha < 1$. This means that, in the case of high TM, "-omega"

[2] It is important to refer one more time that the *Dependence* coefficient is interpreted differently from $\theta = 2A(1/2)$. According to θ, the intensity of the dependence increases as θ decreases.

Table 5.14: AIC and *Dependence* coefficient for the parametric models of Section 3.2 that we fitted to model the joint distribution function of (-omega,precipitation) in the cases of low and high TM within the region $x_1 > 0.03$, $x_2 > 5.2$ in both cases. The *Dependence* coefficient is defined as $2(1-A(1/2))$, where $A(.)$ is the corresponding Pickands dependence function. Computations were performed using the *R* package *evd.*

	Low TM		High TM	
Parametric Model	AIC	*Dependence* $2(1-A(1/2))$	AIC	*Dependence* $2(1-A(1/2))$
Logistic	272.498	0.232	**311.341**	**0.359**
Asymmetric Logistic	277.621	0.212	331.227	0.323
Husler-Reiss	271.328	0.227	311.952	0.352
Negative Logistic	271.351	0.229	311.366	0.357
Asymmetric Negative Logistic	273.722	0.205	315.197	0.348
Bilogistic	**270.826**	**0.207**	319.310	0.418
Negative Bilogistic	279.130	0.315	322.641	0.318
Coles-Tawn	274.363	0.264	312.850	0.353

and precipitation are not independent at extreme values.

Table 5.15: ML estimates (standard errors in brackets) for the marginal and dependence parameters of the bilogistic model for (-omega,precipitation) in the case of low TM; and the logistic model for (-omega,precipitation) in the case of high TM. The estimates were obtained using the one-step censored-likelihood method: $\widehat{\sigma_1}$ and $\hat{\xi}_1$ refer to the estimates of the scale and shape parameter of the GPD fitted to the excess data of "-omega", respectively ; $\widehat{\sigma_2}$ stands for the estimate of the scale parameter of the Exponential model fitted to the excess data of precipitation; $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the dependence parameters of the Bilogistic model (3.20) and $\hat{\alpha}$ is the estimate of the dependence parameter of the Logistic model (3.15). Computations were performed using the *R* package *evd.*

Bilogistic Model (Low TM)				
$\widehat{\sigma_1}$	$\hat{\xi}_1$	$\widehat{\sigma_2}$	$\hat{\alpha}$	$\hat{\beta}$
0.018 (0.002)	-0.154 (0.077)	1.025 (0.132)	0.911 (0.033)	0.662 (0.114)

Logistic Model (High TM)			
$\widehat{\sigma_1}$	$\hat{\xi}_1$	$\widehat{\sigma_2}$	$\hat{\alpha}$
0.016 (0.001)	-0.264 (0.067)	1.498 (0.145)	0.715 (0.034)

The Pickands dependence functions corresponding to the models presented in Table 5.15 can be found in Figure 5.24. In that figure it is possible to see that the Pickands dependence function corresponding to the Logistic model for high TM is closer to $A(t) = \max(t, 1-t)$, $t \in [0,1]$ (the perfect dependence case), which means that the extremal dependence between "-omega" and precipitation is stronger when there is high TM than when the TM is low, as we had also concluded before.

Finally, let us denote the quantile curve of a joint distribution function F at lower tail probability p as $Q(F,p)$, that is, $Q(F,p) := \{(x_1,x_2) : F(x_1,x_2) = p\}$. In Figure 5.25 it is possible to see the estimates of the quantile curves $Q(F_j, 0.95)$, $Q(F_j, 0.975)$ and $Q(F_j, 0.99)$ for each $j \in \{1,2\}$, where F_1 denotes the joint distribution function of (-omega,precipitation) in the case of low TM and F_2 in that of high TM. In order to construct those estimated curves, we used the models presented in Table 5.15 , fitted within the region $x_1 > 0.03$, $x_2 > 5.2$. As it can be seen in that figure, there are 10 days over the estimate of $Q(F_1, 0.95)$ and 13 days over the estimate of $Q(F_2, 0.95)$, which also shows that the dependence at extreme values is stronger in the case of high TM than in that of low TM. Looking at the estimates of $Q(F_1, 0.99)$ and $Q(F_2, 0.99)$, there are no days over the curve corresponding to low TM and there is one day over the one of high TM, so this is another visual evidence for our conclusion.

5.3 Copula analysis of Precipitation and "-omega"

In the previous section we analyzed the dependence between "-omega" and precipitation in the cases of low and high TM from an extremal point of view. However, copula models provide information not only about the tail dependence between the variables, but also about the global dependence structure. Therefore, we carried out a copula analysis in the same context of the previous section, as a complement to those results. We used the semi-parametric approach described in Section 4.3 because of the advantages explained there.

In Figure 5.26 it is possible to see the pseudo-observations for (-omega,precipitation) in the cases of low and high TM, calculated according to (4.13). As it is visible there, there is a slightly positive relationship between "-omega" and precipitation in both cases. With respect to the tail dependence, the main difference that we can see in the figure is that it seems that the upper tail dependence is stronger in the case of high TM, as there is a larger accumulation of points in the top-right corner. As it was explained in Section 4.3, the pseudo-observations are used for computing the Maximum Pseudo-Likelihood Estimator for the parameters of the copula models that were fitted.

Using the *R* package *VineCopula*, we found out that, for high TM, the two copula models with the lowest AIC value (among the ones presented in Section 4.2) are the Student-*t* and the Gumbel. In what regards low TM, the three best models are the Student-*t*, Gaussian and Gumbel copulas. Due to the fact that the values of AIC were very similar, for comparative purposes, we present the results regarding the Gumbel instead of the Gaussian (AIC=-162.0869 for the Gumbel and AIC=-168.8751 for the Gaussian). As such, we consider that the Student-*t* and the Gumbel are appropriate for this analysis.

Let δ be the vector that contains the parameters of the copula. In the case of the Student-*t* copula, $\delta = (\delta_1, \delta_2) = (\rho, \eta)$, where ρ is the correlation coefficient and η is the number of degrees of freedom. In the Gumbel copula, $\delta = \delta_1 = \alpha$, where α is the association parameter. In Table 5.16 it is possible to find the MPLEs for the copula parameters of the Student-*t* and the Gumbel copulas fitted to the pair (-omega,precipitation) in the cases of low and high TM. They were computed using the *R* package *VineCopula*. That table also includes, for each copula fitted, the estimate of Kendall's τ obtained by substituting the MPLE of ρ (Student-*t* copulas) and α (Gumbel copulas) into the functional relationships between τ and those parameters, which were shown in Section 4.2 . As it can be seen from the values of $\hat{\tau}$ obtained, they are larger in the case of high TM than low TM, for both types of copulas. This means that the global dependence between "-omega" and precipitation is stronger in the case of high TM than when TM is low.

In Table 5.16 the estimates of the upper and lower tail dependence coefficients (λ_U and λ_L, defined in (4.11) and (4.12), respectively) can also be found. The estimates of those coefficients are obtained by substituting the MPLEs of the copula parameters into the functional relationships presented in Table 4.1. Looking at the values of $\hat{\lambda}_L$ and $\hat{\lambda}_U$ of the fitted copulas, it can be said that, regarding the Student-*t* copulas, $\hat{\lambda}_L$ and $\hat{\lambda}_U$ are larger in the case of high TM than when there is low TM; although the values are quite small in both cases. With respect to the Gumbel copulas, the value of $\hat{\lambda}_U$ is also slightly larger in the case of high TM than when TM is low. Therefore, these results show that the upper tail dependence between "-omega" and precipitation is stronger in the case of high TM than when there is low TM, as we had also concluded in the previous section. Additionally, by the comparison of the fitted Student-*t* copulas, we can conclude that the lower tail dependence is slightly stronger in the case of high TM than when there is low TM.

The AIC values of the fitted copulas are obtained by (4.16) using the *R* package *VineCopula* and can also be found in Table 5.16 . By looking at those values we can see that the AIC of the fitted Student-*t* copula is lower than the one corresponding to the Gumbel copula in the case of low TM, and the opposite occurs in the case of high TM. Therefore, it is possible to say, in terms of AIC, that the Student-*t* copula is the most appropriate model for low TM and, in contrast, the Gumbel copula is the best choice in the case of high TM.

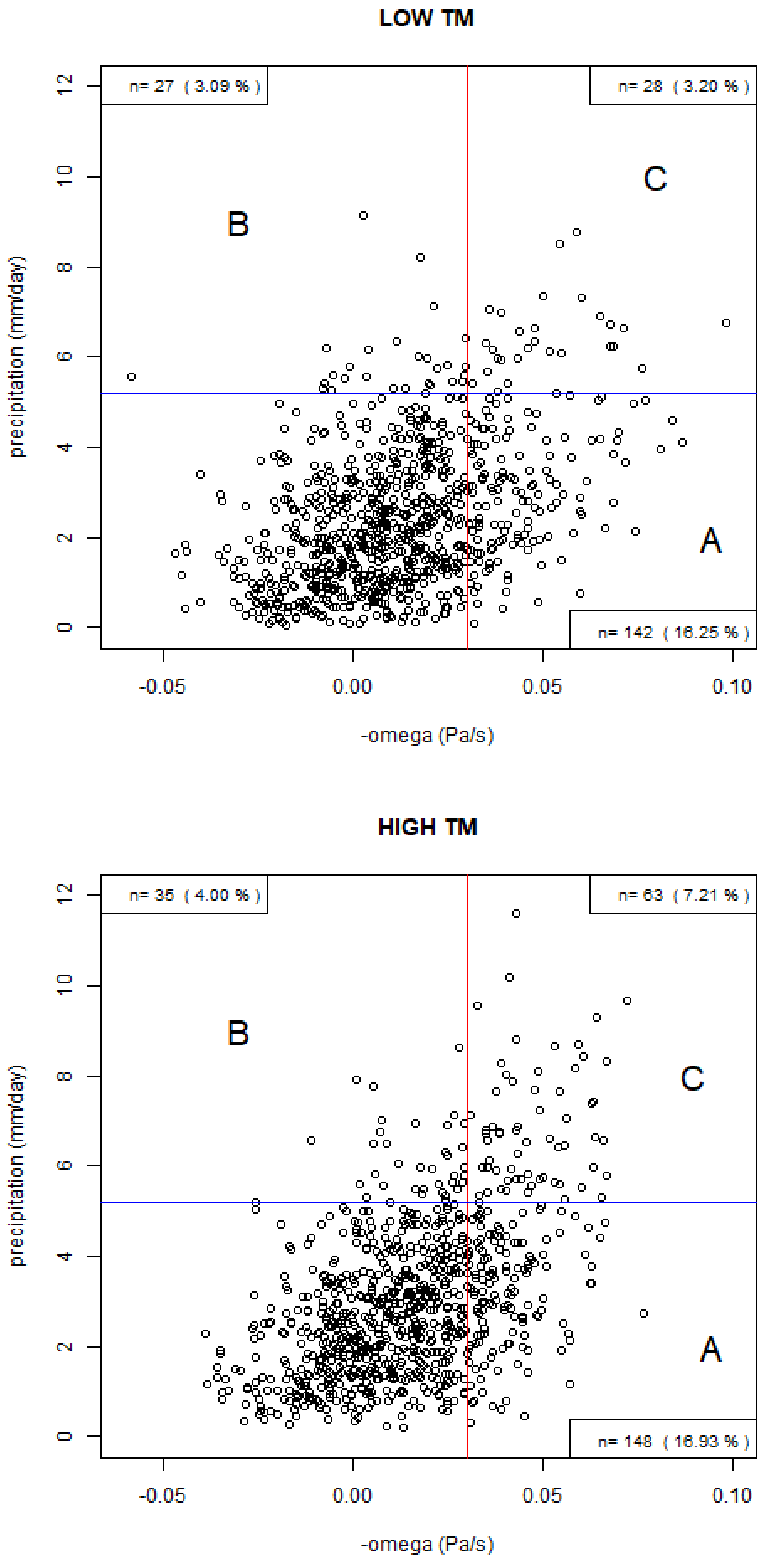
In terms of interpretation, it makes a lot of sense to fit a Gumbel copula (which only has upper tail dependence) to the pair (-omega, precipitation) when TM is high. In fact, as pointed out before when presenting Figure 5.26 , there seems to be a clear accumulation of points in the extreme upper corner of the plot. In the case of low TM, the association between the variables in the extreme upper corner of Figure 5.26 is less intense than in the high TM case.

Additionally, Table 5.16 includes the results of the goodness-of-fit tests that were performed using the *R* package *gofCopula*. As it was said in Section 4.4, the null hypothesis of these tests is $H_0 : C \in C_0$, where C_0 is a given family of copulas. The tests based on the S_n and $S_n^{(B)}$ statistics were carried out for each of the fitted copulas. In both cases, the approximate *p*-values obtained are larger than all the usual significance levels, which means that, at those levels, the null hypothesis is not rejected. This allows us to conclude that the Student-*t* and Gumbel copulas fit well to the data in both the cases of low and high TM.

Table 5.16: **Information about the Student-*t* and the Gumbel copulas fitted to the pair (-omega,precipitation) in the cases of low and high TM.** $\hat{\delta}_1$ refers to the MPLE for the correlation coefficient ρ in the case of the Student-*t* copulas and the association parameter α in the Gumbel copulas. $\hat{\delta}_2$ is the MPLE for the degrees of freedom (η) of the Student-*t* copulas. $\hat{\tau}$ is the estimate of Kendall's τ obtained by substituting the MPLE of ρ (Student-*t* copulas) and α (Gumbel copulas) into the functional relationships between τ and those parameters, which were shown in Section 4.2 . $\hat{\lambda}_U$ and $\hat{\lambda}_L$ are the estimates of the upper and lower tail dependence coefficients (respectively), obtained by substituting the MPLEs of the copula parameters into the functional relationships presented in Table 4.1 . The AIC of each copula model is obtained by (4.16) and the observed value of the S_n and $S_n^{(B)}$ statistics explained in Section 4.4 (with their corresponding approximate *p*-values) are also included. Computations were performed using the *R* packages *VineCopula* and *gofCopula*, the latter being used to carry out the goodness-of-fit tests.

	Low TM		**High TM**	
	Student-*t*	Gumbel	Student-*t*	**Gumbel**
$\hat{\delta}_1$	0.428	1.349	0.494	1.454
$\hat{\delta}_2$	23.392	–	13.820	–
$\hat{\tau}$	0.282	0.258	0.329	0.312
$\hat{\lambda}_L$	0.005	0	0.041	0
$\hat{\lambda}_U$	0.005	0.328	0.041	0.389
AIC	$-\mathbf{168.939}$	-162.087	-236.600	$-\mathbf{238.267}$
S_n statistic approx. *p*-value	0.015 ≈ 1	0.042 0.976	0.024 ≈ 1	0.027 ≈ 1
$S_n^{(B)}$ statistic approx. *p*-value	0.030 0.998	0.057 0.895	0.036 0.992	0.035 0.994

Finally, we will focus our attention on the Student-*t* copula (in the low TM case) and the Gumbel copula (in the high TM case). In Figure 5.27 we present the copula density function (see (4.2)) of the fitted Student-*t* copula for (-omega,precipitation) in the case of low TM, as well as the fitted Gumbel copula in the case of high TM. In Figure 5.28 it is possible to see, in each case of low and high TM, the pseudo-observations and simulated data from those fitted copula models. That is, using the *R* package *VineCopula*, we obtained 874 simulated observations from each of the models (Remember that 874 is the number of observations of (-omega,precipitation) in each case of low and high TM). Looking at that figure, the pattern that we can see in the pseudo-observations is very similar to the one that is observed in the simulated data, in each case of low and high TM. Therefore, this is a complementary visual evidence that the fitted Student-*t* copula is an appropriate model for (-omega,precipitation) in the case of low TM and the same can be said for the fitted Gumbel copula in the case of high TM.


LOW TM

HIGH TM

Figure 5.22: Scatterplots for (-omega,precipitation) in the cases of low TM (top plot) and high TM (bottom plot). The red line refers to the threshold for "-omega" (0.03 in both cases of low and high TM) and the blue line corresponds to the one for precipitation (5.2 in both cases). At the top-left, top-right and bottom-right corners of each plot, it is possible to see the number of points in the corresponding quadrant (and the percentage that they represent in terms of the total sample size). The letters "A", "B" and "C" are used for identifying the quadrant in which they are located.

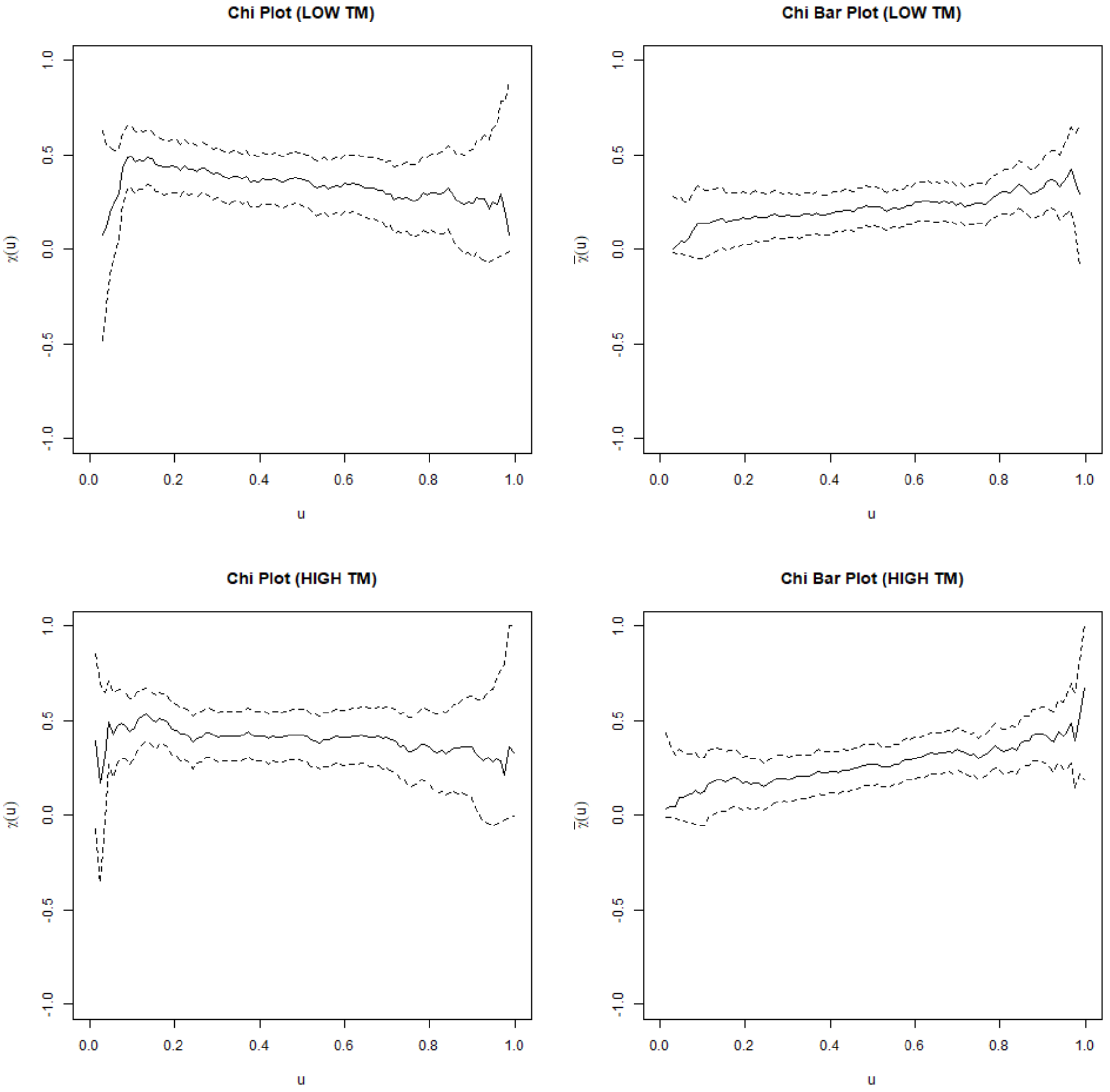


Figure 5.23: *Chi plots* and *chi bar plots* for (-omega,precipitation) in the cases of low TM and high TM. *Chi plots* and *chi bar plots* are plots of $u \in (0,1)$ against empirical estimates of $\chi(u)$ and $\bar{\chi}(u)$, respectively (remember Section 3.5). The dashed lines refer to the approximate 95% confidence intervals computed via the delta method. In this figure, the plots in the top row correspond to the case of low TM, while the ones in the bottom row refer to the case of high TM. Within each row, the *chi plot* is on the left and the *chi bar plot* is on the right. These plots were constructed using the *R* package *evd*.

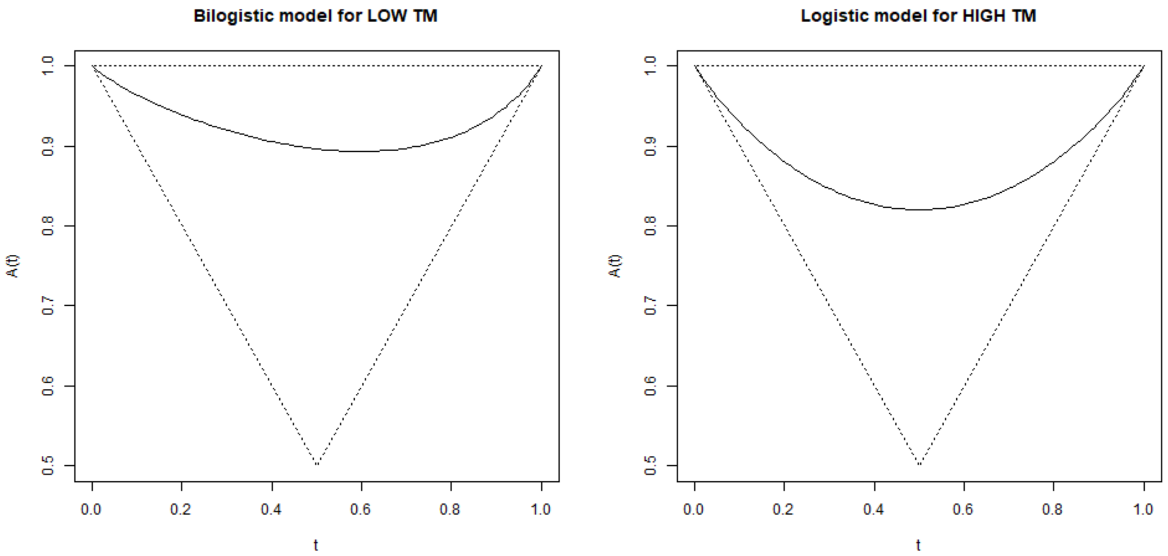


Figure 5.24: Pickands dependence functions corresponding to the fitted Bilogistic model for (-omega,precipitation) in the case of low TM (plot on the left) ; and corresponding to the fitted Logistic model for (-omega,precipitation) in the case of high TM (plot on the right). The information about the models whose Pickands dependence functions are represented here can be found in Table 5.15 . In these plots, the dashed black lines refer to the functions $A(t) = 1,\ t \in [0,1]$ and $A(t) = \max(t, 1-t)\ ,\ t \in [0,1]$, which correspond to the independence and perfect dependence case, respectively. This figure was constructed using the *R* package *evd*.

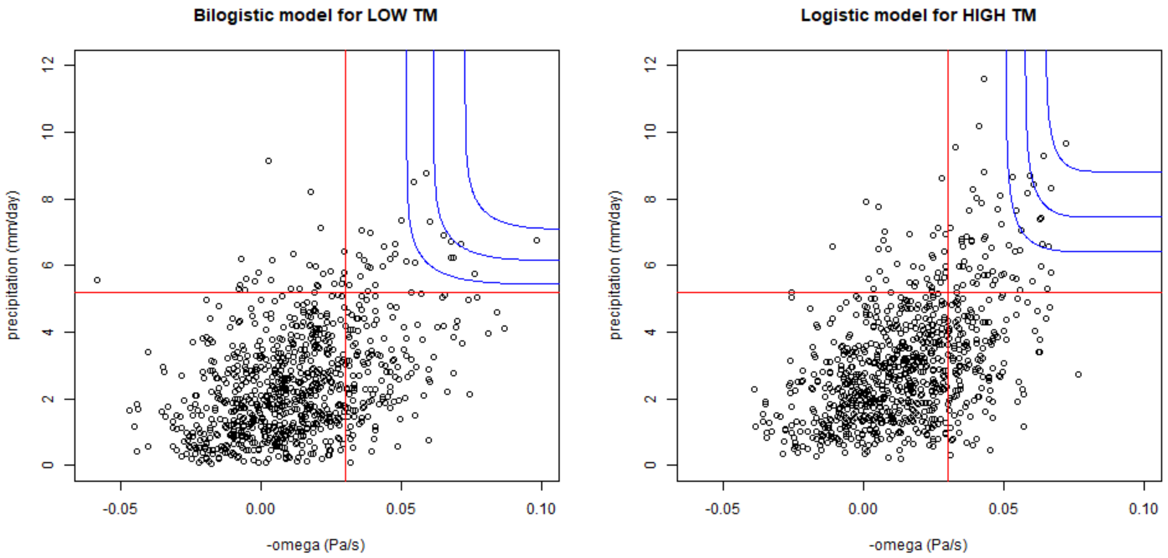


Figure 5.25: Estimated quantile curves at lower tail probabilities $p = 0.95, 0.975, 0.99$ for the joint distribution function of (-omega,precipitation) in the cases of low TM (left) and high TM (right), using the fitted Bilogistic model in the case of low TM and the Logistic model in the case of high TM. The models are fitted to the region $x_1 > 0.03,\ x_2 > 5.2$ in both cases of low and high TM. In these plots, the red vertical line refers to the threshold for "-omega" ($u_1 = 0.03$) ; and the red horizontal line refers to the threshold for precipitation ($u_2 = 5.2$). This figure was constructed using the *R* package *evd*.

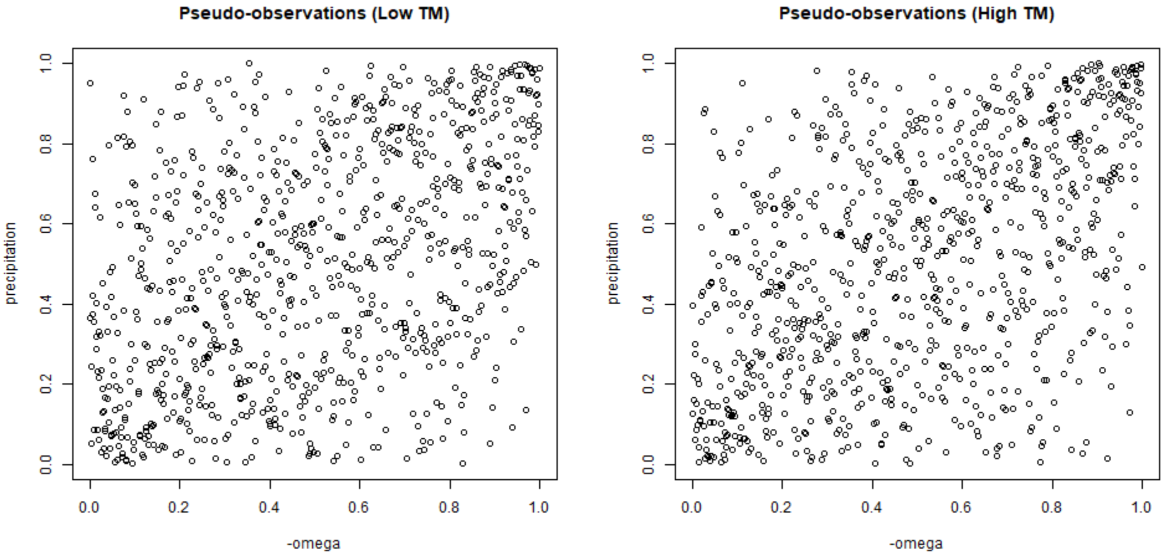


Figure 5.26: Pseudo-observations for (-omega,precipitation) in the cases of low TM (on the left) and high TM (on the right).

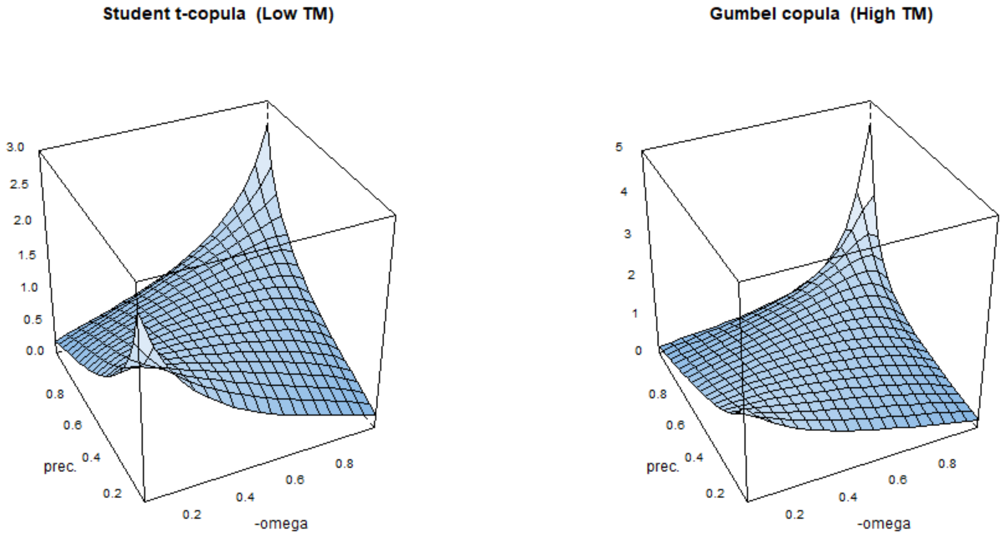


Figure 5.27: Copula density function for the most parsimonious fitted copula model for (-omega,precipitation) in each case of low and high TM. The plot on the left corresponds to the fitted Student-*t* copula (low TM case), whereas the one on the right is for the fitted Gumbel copula (high TM case).

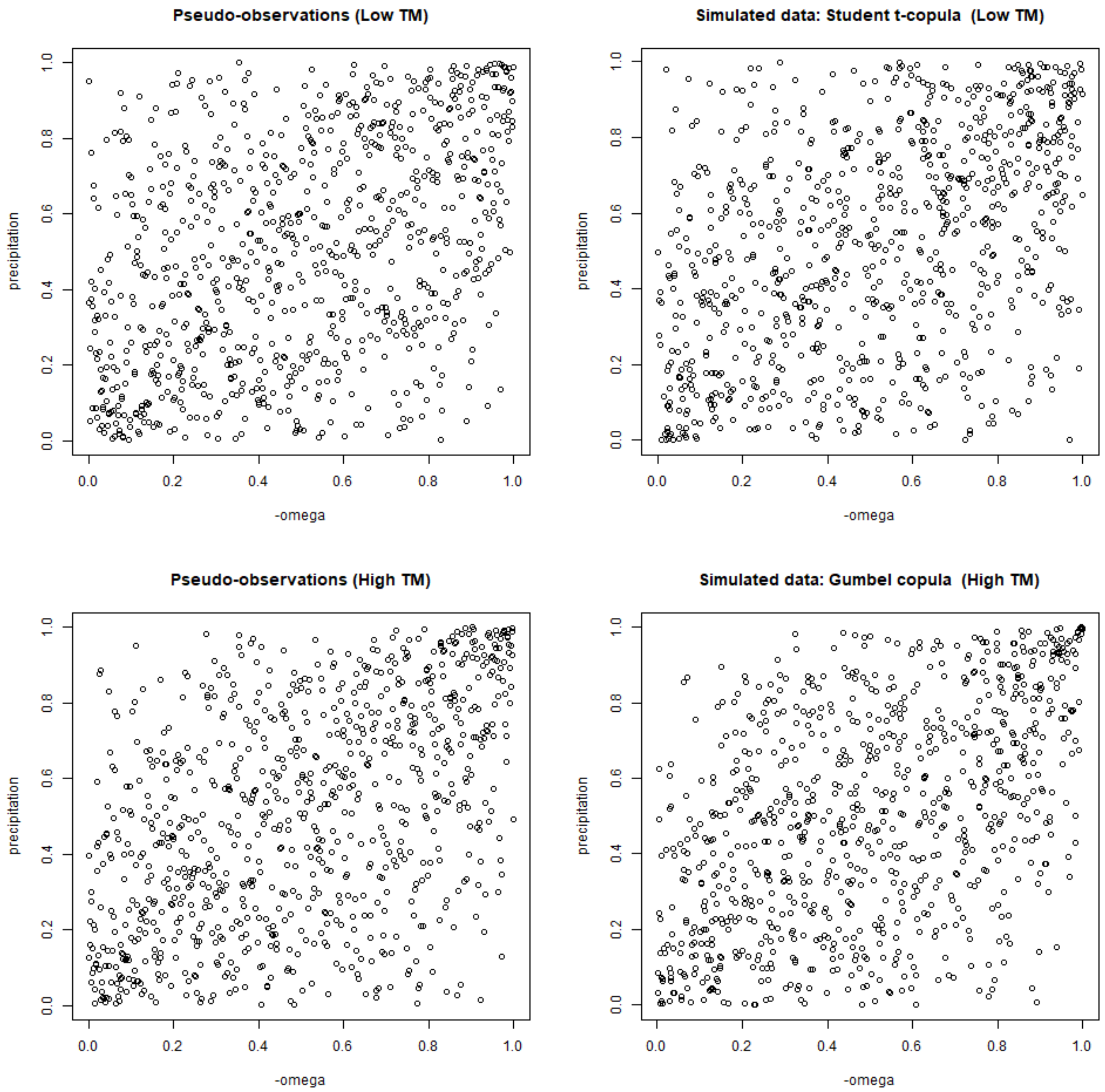


Figure 5.28: Pseudo-observations and simulated data from the most parsimonious fitted copula model for (-omega,precipitation) in each case of low and high TM. The first row of plots corresponds to the low TM case, while the second one is for high TM. Within each row, the plot on the left corresponds to the pseudo-observations and the one on the right to the simulated data from the fitted Student-*t* copula (in the low TM case) and the fitted Gumbel copula (in the high TM case). The number of simulated values is 874 in both the cases of low and high TM because this is the number of observations of (-omega,precipitation) in each case of low and high TM. Simulations were performed using the *R* package *VineCopula*.

6 | Comments, Conclusions and Future Work

In this thesis we placed ourselves in the context of the Great Plains Low-Level Jet (GPLLJ) system, which is a system of very strong winds in the lower troposphere that transports a huge amount of moisture from the Gulf of Mexico to the American Great Plains and is mainly active during the summer months, as it was explained in detail in Section 1.1 .

There were two main objectives in this work: first, to analyze the extremal behaviour of the Transported Moisture from the GPLLJ source region to the jet domain; second, in the cases of low and high TM, to study the global and extremal dependence between the upper tail of the precipitation in the GPLLJ sink region and the lower tail of the tropospheric stability in the GPLLJ sink region (omega). We should stress the fact the term “global” refers to the set of pair (-omega,precipitation) considering the entire sample of low values of TM (25% of the lowest values). The same term was applied for the case of high values of TM (25% of the highest values).

For this purpose, we used the series of daily observations of Transported Moisture, Precipitation and "omega" of all June-July-August periods from 1980 to 2017, that is, 3496 observations. This data was described in depth in Section 1.2 .

In Chapter 2 the fundamental concepts of univariate Extreme Value Theory were presented. It is relevant to emphasize the importance of Section 2.2 , since threshold models were essential for both objectives of the thesis: to perform the univariate extremal analysis of TM and as a necessary previous step before the study of the bivariate extremes.

In Chapter 3 it is possible to find some of the key topics of Bivariate Extreme Value Theory, which was used in order to tackle the extremal part of the second objective of this thesis. The first two sections of this chapter cover the probabilistic notions and some of the most important parametric models of bivariate extremes, while the next two deal with the statistical methodology that is commonly used to work with bivariate extremes. In our thesis, the contents of Section 3.4 were very important because in our analysis we used the censored-likelihood method to fit the Bivariate Threshold Excess Model, which is an approach that is carefully explained in that section. Finally, in Section 3.5 the notion of *Asymptotic independence* is addressed, which is a situation that should be analyzed when using the methodology presented in this thesis.

In order to obtain a global picture of the dependence structure, in the context of the second objective of this work, we resorted to copulas. In Chapter 4 we presented a summary of the most important aspects of Copula Theory, both from the probabilistic and the statistical points of view. Namely, we introduced the concept of *copula*, some of the most common copula models, different methods of estimation (focusing on the semi-parametric approach, the one used in our study), as well as some brief explanations about model selection and goodness-of-fit tests for copulas.

In Chapter 5 the procedure and the results of our study are presented. Section 5.1 addresses the univariate extremal analysis of TM. After carrying out a brief exploratory analysis in order to understand better the series under study, we used threshold models to study the behaviour of the extreme values of that series. By means of two of the most usual methods of threshold selection, we decided that $u = 2$ (mm/day) was an appropriate threshold for the TM series. As it was clearly visible that the excesses over the chosen threshold were not independent, a POT analysis with "runs-declustering" was used in order to remove that dependence as much as possible. The declustering was carried out considering four different values of *run length* (r), namely 1,2,3 and 4. Graphically and by means of some statistical tests, we came to the conclusion that the Exponential model was more appropriate than the GPD to model the cluster maxima of excesses over the chosen threshold, for all the values of r considered. Moreover, we came to the conclusion that in the case of $r = 4$, the non-stationary Exponential model was more adequate than the stationary one, in the sense that it was shown that the scale parameter of the Exponential model decreases with time. For this reason and because it is the value that guarantees the independence between the excesses the best, we concluded that $r = 4$ was the best choice among the ones we tried. We also computed the estimated 38-year, 50-year and 100-year return levels for the TM series using the non-stationary Exponential model for the cluster maxima of excesses. The results of those computations showed that the three estimated return levels decreased over time and that the difference between them became smaller. Therefore, it is possible to say that we expect to observe lower extreme values of TM in the future.

In Section 5.2 we analyzed the bivariate extremes of (-omega,precipitation) in the cases of low and high TM. Note that we changed the sign of "omega" because, meteorologically speaking, the interest lied on the study of the joint behaviour of the upper tail of precipitation and the lower tail of "omega". The series of precipitation and "-omega" were lagged 1 day with respect to the TM series due to the temporal nature of the GPLLJ system. After a preliminary analysis of the variables under study, we began the process of fitting a Bivariate Threshold Excess Model. First, it was necessary to fit univariate threshold models to the margins. In the cases of low and high TM, an appropriate threshold for "-omega" is $u_1 = 0.03$ (Pa/s) and, for precipitation, it is adequate to choose $u_2 = 5.2$ (mm/day). We also concluded that, selecting those thresholds, the GPD model was more adequate than the Exponential in the case of "-omega" and the opposite occurred for precipitation, in both the cases of low and high TM. Having chosen those distributions for the excesses over the respective threshold at each margin, we used the censored-likelihood method considering the eight different parametric models presented in Chapter 3 . Irrespectively of the model which was fitted to the bivariate data, the extremal dependence between "-omega" and precipitation was stronger in the case of high TM than when there is low TM. The AIC values corresponding to each of those models were also computed and the most parsimonious model in the case of low TM is the Bilogistic one, whereas in the case of high TM, it is the Logistic one. The most relevant information about the fitted Bilogistic model for low TM and the fitted Logistic model for high TM was shown, namely the ML estimates of their coefficients, the estimated Pickands dependence functions, as well as some estimated quantile curves. Moreover, by means of the *chi plots* and *chi bar plots*, we came to the conclusion that we can assume that the variables are asymptotically dependent, and therefore the models that were presented in this thesis are appropriate for this pair of variables, in both the cases of low and high TM.

Last but not least, in Section 5.3 it is possible to see the results corresponding to the copulas that were fitted for (-omega,precipitation) in both the cases of low and high TM. We came to the conclusion that the global dependence between "-omega" and precipitation is stronger in the case of high TM than when there is low TM. These conclusions were obtained through the fitted Student-*t* and Gumbel copulas.

Moreover, by means of the estimates of the tail dependence coefficients, we found out that the upper tail dependence between "-omega" and precipitation is stronger in the case of high TM than when there is low TM, as we already knew by the study of the bivariate extremes. Additionally, by using copulas it was shown that the same conclusion is true in what concerns the lower tail dependence.

Meteorological interpretation of the results obtained in this thesis

First objective of the thesis The TM from the Caribbean sources to the GPLLJ region is controlled by the position of the western ridge of the Azores Anticyclone (AA), a dominant feature of the climate at both sides of the Atlantic. The AA western ridge controls the intensity and location of the GPLLJ, modulating summertime moisture budgets and precipitation in the central and eastern United States (Nieto Ferreira and Rickenbach, 2020). There is an important interannual variability in the position of the AA western ridge (in part modulated by natural modes of variability) and a continuous observed shift eastward since 1978. This shift makes it difficult to exist a strong TM towards central United States and facilitates the transport towards southeastern United States. **This is consistent with the decrease in the estimated return levels in our TM series along the studied period.**

Second objective of the thesis Precipitation is a very complex process, but in essence it can be said to be regulated by two factors: the moisture content (water column) and the magnitude of the ascending movement (vertical velocity). When they occur simultaneously, conditions are highly favourable for extreme precipitation (see Kunkel et al., 2020). If instability (vertical ascent) occurs in the moisture sink region of the GPLLJ, precipitation is likely. However, it is even more likely if there is an abundant supply of moisture in the lower troposphere, transported by the GPLLJ from its Gulf of Mexico sources. **This is consistent with the fact that in the sink region the global and extremal dependence between precipitation and "-omega" is stronger in the case of high TM than when there is low TM.**

Future work

With respect to **future work** in this context, the large availability of spatial data regarding the GPLLJ system makes it possible to carry out extremal analyses accounting for the spatial dimension of the meteorological phenomena (see Davison et al., 2012). Moreover, in this thesis we have been using (bivariate) copulas, which are limited to the study of the dependence structure of a pair of variables. In order to study the dependence structure of three or more variables, *vine copulas* may be used (see Czado, 2019). Also, it should be mentioned that we could also address these problems in a Bayesian framework.

References

AghaKouchak, A. (2015). A multivariate approach for persistence-based drought prediction: Application to the 2010–2011 East Africa drought. *Journal of Hydrology*, 526:127–135.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Algarra, I., Eiras-Barca, J., Nieto, R., and Gimeno, L. (2019). Global climatology of nocturnal low-level jets and associated moisture sources and sinks. *Atmospheric Research*, 229:39–59.

André, L. and de Zea Bermudez, P. (2020). Modelling dependence between observed and simulated wind speed data using copulas. *Stochastic Environmental Research and Risk Assessment*, 34(11):1725–1753.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, pages 792–804.

Basara, J. B., Maybourn, J. N., Peirano, C. M., Tate, J. E., Brown, P. J., Hoey, J. D., and Smith, B. R. (2013). Drought and associated impacts in the Great Plains of the United States—a review. *International Journal of Geosciences*, 4(6B):72–81.

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

Burrows, D. A., Ferguson, C. R., Campbell, M. A., Xia, G., and Bosart, L. F. (2019). An objective classification and analysis of upper-level coupling to the Great Plains low-level jet over the twentieth century. *Journal of Climate*, 32(21):7127–7152.

Castillo, E. and Hadi, A. S. (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620.

Chen, C., Tao, W.-K., Lin, P.-L., Lai, G. S., Tseng, S., and Wang, T.-C. C. (1998). The intensification of the low-level jet during the development of mesoscale convective systems on a mei-yu front. *Monthly Weather Review*, 126(2):349–371.

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4):478–484.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):377–392.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):1–31.

Cong, R.-G. and Brady, M. (2012). The interdependence between rainfall and temperature: copula analyses. *The Scientific World Journal*, 2012.

Czado, C. (2019). *Analyzing dependent data with vine copulas*. Springer.

Davison, A. C., Padoan, S. A., Ribatet, M., et al. (2012). Statistical modeling of spatial extremes. *Statistical science*, 27(2):161–186.

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

de Zea Bermudez, P. and Kotz, S. (2010a). Parameter estimation of the generalized Pareto distribution—Part I. *Journal of Statistical Planning and Inference*, 140(6):1353–1373.

de Zea Bermudez, P. and Kotz, S. (2010b). Parameter estimation of the generalized Pareto distribution—Part II. *Journal of Statistical Planning and Inference*, 140(6):1374–1388.

Drumond, A., Stojanovic, M., Nieto, R., Vicente-Serrano, S. M., and Gimeno, L. (2019). Linking anomalous moisture transport and drought episodes in the IPCC reference regions. *Bulletin of the American Meteorological Society*, 100(8):1481–1498.

Embrechts, P., Lindskog, F., and McNeil, A. (2003). *Modelling dependence with copulas and applications to risk management, handbook of Heavy Tailed Distributions in Finance*. Elsevier/North-Holland, Amsterdam.

Fawcett, L. and Walshaw, D. (2008). Modelling environmental extremes. In *Short course for the 19th annual conference of the international environmetrics society*.

Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q. (2012). *Managing the risks of extreme events and disasters to advance Climate Change adaptation: special report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213.

Gimeno, L., Dominguez, F., Nieto, R., Trigo, R., Drumond, A., Reason, C. J., Taschetto, A. S., Ramos, A. M., Kumar, R., and Marengo, J. (2016). Major mechanisms of atmospheric moisture transport and their role in extreme precipitation events. *Annual Review of Environment and Resources*, 41:117–141.

Gimeno, L., Stohl, A., Trigo, R. M., Dominguez, F., Yoshimura, K., Yu, L., Drumond, A., Durán-Quesada, A. M., and Nieto, R. (2012). Oceanic and terrestrial sources of continental precipitation. *Reviews of Geophysics*, 50(4).

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, pages 423–453.

Gomes, M. I., Fraga Alves, M., and Neves, C. (2013). *Análise de Valores Extremos: Uma Introdução*. Edições SPE & INE.

Gumbel, E. J. (1960). Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

Helfand, H. M. and Schubert, S. D. (1995). Climatology of the simulated Great Plains low-level jet and its contribution to the continental moisture budget of the United States. *Journal of Climate*, 8(4):784–806.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hodges, D. and Pu, Z. (2019). Characteristics and variations of low-level jets in the contrasting warm season precipitation extremes of 2006 and 2007 over the Southern Great Plains. *Theoretical and Applied Climatology*, 136(1):753–771.

Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286.

Joe, H. (1990). Families of min-stable multivariate exponential and multivariate extreme value distributions. *Statistics & Probability letters*, 9(1):75–81.

Joe, H. (2014). *Dependence modeling with copulas*. CRC Press.

Kunkel, K. E., Stevens, S. E., Stevens, L. E., and Karl, T. R. (2020). Observed climatological relationships of extreme daily precipitation events with precipitable water and vertical velocity in the contiguous United States. *Geophysical Research Letters*, 47(12):e2019GL086721.

Lazoglou, G. and Anagnostopoulou, C. (2019). Joint distribution of temperature and precipitation in the Mediterranean, using the Copula method. *Theoretical and Applied Climatology*, 135(3):1399–1411.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):475–499.

Ledford, A. W. and Tawn, J. A. (1998). Concomitant tail behaviour for extremes. *Advances in Applied Probability*, pages 197–215.

Lee, T., Modarres, R., and Ouarda, T. B. (2013). Data-based analysis of bivariate copula tail dependence for drought duration and severity. *Hydrological Processes*, 27(10):1454–1463.

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

Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P. R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., et al. (2018). Global Warming of 1.5 C. *An IPCC Special Report on the impacts of Global Warming of 1.5 C*, 1:1–9.

Mo, K. C., Nogues-Paegle, J., and Paegle, J. (1995). Physical mechanisms of the 1993 summer floods. *Journal of Atmospheric Sciences*, 52(7):879–895.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science & Business Media.

Nieto Ferreira, R. and Rickenbach, T. M. (2020). Effects of the North Atlantic Subtropical High on summertime precipitation organization in the southeast United States. *International Journal of Climatology*, 40(14):5987–6001.

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

Pitchford, K. L. and London, J. (1962). The low-level jet as related to nocturnal thunderstorms over Midwest United States. *Journal of Applied Meteorology and Climatology*, 1(1):43–47.

Poonia, V., Jha, S., and Goyal, M. K. (2021). Copula based analysis of meteorological, hydrological and agricultural drought characteristics across Indian river basins. *International Journal of Climatology*, pages 1–16.

Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P., et al. (2014). Climate Change 2013: the physical science basis. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (eds TF Stocker et al.)*, pages 5–14.

Raymond, C., Horton, R. M., Zscheischler, J., Martius, O., AghaKouchak, A., Balch, J., Bowen, S. G., Camargo, S. J., Hess, J., Kornhuber, K., et al. (2020). Understanding and managing connected extreme events. *Nature Climate Change*, 10(7):611–621.

Reddy, M. J. and Ganguli, P. (2012). Bivariate flood frequency analysis of Upper Godavari River flows using Archimedean copulas. *Water Resources Management*, 26(14):3995–4018.

Ridder, N. N., Pitman, A. J., Westra, S., Ukkola, A., Do Hong, X., Bador, M., Hirsch, A. L., Evans, J. P., Di Luca, A., and Zscheischler, J. (2020). Global hotspots for the occurrence of compound events. *Nature Communications*, 11(1):1–10.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Schumacher, R. S. and Johnson, R. H. (2009). Quasi-stationary, extreme-rain-producing convective systems associated with midlevel cyclonic circulations. *Weather and Forecasting*, 24(2):555–574.

Shemyakin, A. and Kniazev, A. (2017). *Introduction to Bayesian estimation and copula models of dependence*. Wiley Online Library.

Sibuya, M. (1960). Bivariate Extreme Statistics, I. *Annals of the Institute of Statistical Mathematics*, 11(3):195–210.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Smith, R. L. (1990). Extreme Value Theory. *Handbook of Applicable Mathematics*, 7:437–471.

Squitieri, B. J. and Gallus, W. A. (2016). WRF forecasts of Great Plains nocturnal low-level jet-driven MCSs. Part I: Correlation between low-level jet forecast accuracy and MCS precipitation forecast skill. *Weather and Forecasting*, 31(4):1301–1323.

Squitieri, B. J. and Gallus Jr, W. A. (2016). WRF forecasts of Great Plains nocturnal low-level jet-driven MCSs. Part II: Differences between strongly and weakly forced low-level jet environments. *Weather and Forecasting*, 31(5):1491–1510.

Stensrud, D. J. (1996). Importance of low-level jets to climate: A review. *Journal of Climate*, 9(8):1698–1711.

Stephenson, A. (2018). Statistics of Multivariate Extremes. *Technical Report. CRAN.*

Stohl, A. and James, P. (2004). A Lagrangian analysis of the atmospheric branch of the global water cycle. Part I: Method description, validation, and demonstration for the August 2002 flooding in Central Europe. *Journal of Hydrometeorology*, 5(4):656–678.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, 56(1):79–107.

Tawn, J. A. (1988). Bivariate Extreme Value Theory: models and estimation. *Biometrika*, 75(3):397–415.

Walters, C. K. and Winkler, J. A. (2001). Airflow configurations of warm season southerly low-level wind maxima in the Great Plains. Part I: Spatial and temporal characteristics and relationship to convection. *Weather and Forecasting*, 16(5):513–530.

Weaver, S. J., Baxter, S., and Kumar, A. (2012). Climatic role of North American low-level jets on US regional tornado activity. *Journal of Climate*, 25(19):6666–6683.

Xie, P., Chen, M., and Shi, W. (2010). CPC unified gauge-based analysis of global daily precipitation. In *Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc*, volume 2.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., et al. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, 1(7):333–347.

Appendices

A Goodness-of-fit Tests Tables

Table A.1 and Table A.2 contain the simulated quantiles of asymptotic level p for W^2 and A^2 (in the case of both ξ and σ_u being unknown). This is, for each table entry z, $P(T \geq z) = p$, where T is W^2 in the case of the CvM test and A^2 in the case of the AD test. [1]

Table A.1: Simulated quantiles of asymptotic level p for the Cramér-von Mises Statistic (GPD with both parameters unknown). It is an adapted version of Table 2 of Choulakian and Stephens (2001).

$\xi \backslash p$	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
0.900	0.046	0.067	0.094	0.115	0.136	0.165	0.187	0.239
0.500	0.049	0.072	0.101	0.124	0.147	0.179	0.204	0.264
0.200	0.053	0.078	0.111	0.137	0.164	0.200	0.228	0.294
0.100	0.055	0.081	0.116	0.144	0.172	0.210	0.240	0.310
0.000	0.057	0.086	0.124	0.153	0.183	0.224	0.255	0.330
-0.100	0.059	0.089	0.129	0.160	0.192	0.236	0.270	0.351
-0.200	0.062	0.094	0.137	0.171	0.206	0.254	0.291	0.380
-0.300	0.065	0.100	0.147	0.184	0.223	0.276	0.317	0.415
-0.400	0.069	0.107	0.159	0.201	0.244	0.303	0.349	0.458
-0.500	0.074	0.116	0.174	0.222	0.271	0.338	0.390	0.513

Table A.2: Simulated quantiles of asymptotic level p for the Anderson-Darling Statistic (GPD with both parameters unknown). It is an adapted version of Table 2 of Choulakian and Stephens (2001).

$\xi \backslash p$	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
0.900	0.339	0.471	0.641	0.771	0.905	1.086	1.226	1.559
0.500	0.356	0.499	0.685	0.830	0.978	1.180	1.336	1.707
0.200	0.376	0.534	0.741	0.903	1.069	1.296	1.471	1.893
0.100	0.386	0.550	0.766	0.935	1.110	1.348	1.532	1.966
0.000	0.397	0.569	0.796	0.974	1.158	1.409	1.603	2.064
-0.100	0.410	0.591	0.831	1.020	1.215	1.481	1.687	2.176
-0.200	0.426	0.617	0.873	1.074	1.283	1.567	1.788	2.314
-0.300	0.445	0.649	0.924	1.140	1.365	1.672	1.909	2.475
-0.400	0.468	0.688	0.985	1.221	1.465	1.799	2.058	2.674
-0.500	0.496	0.735	1.061	1.321	1.590	1.958	2.243	2.922

[1] It should be mentioned that Choulakian and Stephens (2001) do not consider the same parameterization of the GPD that we are using in this work (see (2.10)). For these authors, the shape parameter of the GPD, represented by letter k, is equal to our $-\xi$. The parameterization that we consider in this work is also present in Coles (2001) and Beirlant et al. (2004), among many other texts. The one that Choulakian and Stephens (2001) utilize is the one used by Castillo and Hadi (1997), for example.

B | Scripts

In Section B.1 the *R* code that was written in the context of the univariate analysis of Transported Moisture (Section 5.1) is included.

Moreover, in Section B.2 the *R* code regarding bivariate extremes and copulas can be found. Subsection B.2.1 includes the code written to produce Figure 3.1 , whereas in Subsection B.2.2 it is possible to find the code used for carrying out the bivariate analysis of Precipitation and "-omega" that was presented in Section 5.2 , as well as the copula analysis of Section 5.3 .

B.1 Univariate extremes

```
library(readr)
library(evd)
library(moments)
library(KScorrect)
library(evmix)
library(KScorrect)
library(extRemes)



#TRANSPORTED MOISTURE SERIES



serie_hum <- read_table2("serie_p75_i3-5.txt")
datahum=as.data.frame(serie_hum)
JJA=which( datahum$mes==6 | datahum$mes==7 | datahum$mes==8 )
datahum_JJA=datahum[JJA,]
hum_JJA=datahum_JJA[,4]

length(hum_JJA)
```

```
#PLOT OF THE SERIES

n=length(hum_JJA)


plot(hum_JJA,pch=20,type="l",axes=F,ann=F,cex=0.8)
mtext("Year",side=1,line=2.5,font=1)
mtext("Transported Moisture (mm/day)",side=2,line=2.5,font=1)
box()
axis(2,at=axTicks(2))
abline(h=0,lty=3)
num<-92
pontos<-seq(1,n,by=num)
axis(1,at=pontos,labels=F)
lab=numeric()
lab[seq(1,38,by=3)]=seq(1980,2017,by=3)
lab[38]=NA

lab=as.character(lab)

text(-0.3,x=seq(1,n,by=num),labels=lab,srt=40,
pos=c(1,1),xpd=T,las=2,cex=0.8)



#########

#SUMMARY STATISTICS

summary(hum_JJA)
sd(hum_JJA)


#HISTOGRAM

hist(hum_JJA,prob=T,ylim=c(0,0.7),xlab="Transported Moisture (mm/day)",main="
   Histogram of TM ")
lines(density(hum_JJA),lwd=1)

##

skewness(hum_JJA)
kurtosis(hum_JJA)
```

```
#BOXPLOTS PER YEAR

year=datahum_JJA[,1]
boxplot(hum_JJA ~ year, xlab="Year", ylab="Transported moisture (mm/day)")


#BOXPLOTS FIRST HALF VS SECOND HALF

twogroups=c(rep("Years 1980-1998",1748),rep("Years 1999-2017",1748))
boxplot(hum_JJA ~ twogroups, xlab="", ylab="Transported moisture (mm/day)")



#THRESHOLD SELECTION


#MEAN EXCESS FUNCTION

evmix::mrlplot(hum_JJA,try.thresh=2,legend.loc=NULL)


#SHAPE PLOT

evmix::tshapeplot(hum_JJA,try.thresh=2,tlim=c(1,2.5),legend.loc=NULL)


#MODIFIED SCALE PLOT

evmix::tscaleplot(hum_JJA,try.thresh=2,tlim=c(1,2.5),legend.loc=NULL)



#CHOSEN THRESHOLD
humthres=2

#Number of EXCEEDANCES
length(hum_JJA[hum_JJA>humthres])

#Percentage of EXCEEDANCES
(length(hum_JJA[hum_JJA>humthres])/length(hum_JJA))*100
```

```
#PLOT OF EXCEEDANCES

n=length(hum_JJA)
w=which(hum_JJA>humthres)

plot(hum_JJA,type='h', axes=F ,ann=F, ylim=c(1.5,max(hum_JJA)))
abline(h=humthres, lty=2)
lines(w,hum_JJA[w], type="p",col="red",pch=20)


mtext("Year",side=1,line=2.5,font=1)
mtext("Transported Moisture (mm/day)",side=2,line=2.5,font=1)
box()
axis(2,at=axTicks(2))
abline(h=0,lty=3)
num<-92
pontos<-seq(1,n,by=num)
axis(1,at=pontos,labels=F)
lab=numeric()
lab[seq(1,38,by=3)]=seq(1980,2017,by=3)
lab[38]=NA

lab=as.character(lab)

text(1.3,x=seq(1,n,by=num),labels=lab,srt=40,
pos=c(1,1),xpd=T,las=2,cex=0.8)


####


# We are going to try  r= 1,2,3,4


for(j in 1:4){

        # List containing the clusters of exceedences
        cllisthum=clusters(hum_JJA,u=humthres,r=j)

        #Extremal index
        humtheta=exi(hum_JJA, u=humthres, r=j)


        #Vector containing the cluster maxima
```

```
clmaxhum=clusters(hum_JJA,u=humthres,r=j,cmax=T)







#PLOT OF CLUSTER MAXIMA

indclmaxhum=as.numeric(names(clmaxhum))

n=length(hum_JJA)
plot(hum_JJA,type=’h’,axes=F ,ann=F, ylim=c(1.5,max(hum_JJA)))
abline(h=humthres, lty=2)
lines(indclmaxhum, hum_JJA[indclmaxhum], type="p",col="blue",pch=20)


mtext("Year",side=1,line=2.5,font=1)
mtext("Transported Moisture (mm/day)",side=2,line=2.5,font=1)
mtext(paste("r","=",j,sep=" "),side=3,line=1.5,font=2,cex=1.8)

box()
axis(2,at=axTicks(2))
abline(h=0,lty=3)
num<-92
pontos<-seq(1,n,by=num)
axis(1,at=pontos,labels=F)
lab=numeric()
lab[seq(1,38,by=3)]=seq(1980,2017,by=3)
lab[38]=NA

lab=as.character(lab)

text(1.3,x=seq(1,n,by=num),labels=lab,srt=40,
pos=c(1,1),xpd=T,las=2,cex=0.8)



######

#GPD model

fitpothum=fpot(hum_JJA, threshold=humthres, model="gpd", cmax=TRUE, r=j,
npp=92)

    #scale GPD
    sigma=fitpothum$estimate[[1]]
    #Std.err. scale GPD
    fitpothum$std.err[[1]]

    #shape GPD
    xi=fitpothum$estimate[[2]]
    #Std.err. shape GPD
    fitpothum$std.err[[2]]


    #Deviance GPD
    devgpd=fitpothum$deviance


    #PROFILE LOG-LIKELIHOOD CI (shape parameter)


    prof=profile(fitpothum,conf=0.95,which="shape", mesh=0.008)
    CI=confint(prof)


    plot(prof,xlim=c(-0.2,0.35))

    abline(v=c(CI[1],CI[2]))
    abline(v=fitpothum$estimate[2], col="blue")


    ######

    #Cluster maxima (excesses)
    excesses_clusmax=clmaxhum-humthres

    #Number of clusters (Nc)
    Nc=length(excesses_clusmax)

    NC<-1:Nc



    #Goodness of fit (GPD)

    #Cramer-von Mises
    CVM<-sum((evd::pgpd(sort(excesses_clusmax),scale=sigma,shape=xi)-(2*NC
-1)/(2*Nc))^2)+1/(12*Nc)


    #Anderson-Darling
    AD<--Nc-(1/Nc)*sum((2*NC-1)*log(evd::pgpd(sort(excesses_clusmax),scale =
sigma,shape=xi))+(2*Nc+1-2*NC)*log(1-evd::pgpd(sort(excesses_clusmax),scale=
sigma, shape=xi)))


    ##EXPONENTIAL QQ-PLOT - SAMPLE OF CLUSTER MAXIMA

    pppot1<-(1:Nc)/(Nc+1)
    qqpot1<--log(1-pppot1)

    fitexppot1<-lm(sort(excesses_clusmax)~qqpot1-1)
    summary(fitexppot1)

    summary(fitexppot1)$r.squared


    plot(qqpot1,sort(excesses_clusmax),ylab='empirical quantiles',
    xlab='exponential quantiles', main='Exponential QQ-Plot')

    abline(fitexppot1, lty=2,col='red', cex=3)



    ##GPD QQ-PLOT - SAMPLE OF CLUSTER MAXIMA


    QQgpd<-evd::qgpd(pppot1,shape=xi)

    fitgpd<-lm(sort(excesses_clusmax) ~ QQgpd-1)
    summary(fitgpd)

    summary(fitgpd)$r.squared




    plot(QQgpd,sort(excesses_clusmax), ylab='empirical quantiles',
    xlab='GPD quantiles',main='GPD QQ-Plot')
    abline(fitgpd, lty=2,col='blue', cex=3)
```

```
#########


#Exponential model
fitpothum_exp=fpot(hum_JJA, threshold=humthres, model="gpd", cmax=TRUE,
r=j ,npp=92, shape=0)

#scale Exponential
sigmaexp=fitpothum_exp$estimate

#Std.err scale Exponential
SE_sigmaexp=fitpothum_exp$std.err


#Deviance Exponential
devexp=fitpothum_exp$deviance


#LIKELIHOOD RATIO TEST: EXPONENTIAL MODEL VS GPD MODEL

#Difference of Deviances
difdev=devexp-devgpd

#p-value
pvalue=pchisq(difdev,df=1,lower.tail=F)


###########




#RETURN LEVEL PLOT (STATIONARY EXPONENTIAL MODEL)
plot(fitpothum_exp,which=4,main=paste("r","=",j,sep=" "),xlab="Return
period (years)",ylab="Return level (in mm/day)",ylim=c(2,5.5))


#LcKS test (Exponential model)

out=LcKS(excesses_clusmax, cdf="pexp", nreps = 5000)

#Observed value test statistic
out$D.obs
```

```
    #p-value
    out$p.value



    ####

    ######  USING PACKAGE extRemes TO WORK WITH NON-STATIONARITY


    #STATIONARY EXPONENTIAL MODEL (with PACKAGE extRemes)
    declus <- decluster(hum_JJA, threshold = 2, method = "runs",r=j,groups=
rep(1:38, each=92))
    expfit=fevd(declus, threshold = 2, type = "Exponential", time.units="92/
year")

    #Scale Stationary Exponential (PACKAGE extRemes)
    expfit_scale=summary(expfit)$par

    #Std.err. Scale Stationary Exponential (PACKAGE extRemes)
    expfit_stderr=summary(expfit)$se.theta


    print("Checking that the estimate and std.err. of the scale of the Stat.
 Expo. model are approx. equal in PACKAGES evd and extRemes :")
    print(round(sigmaexp,3)==round(expfit_scale,3))
    print(round(expfit_stderr,3)==round(SE_sigmaexp,3))
    print("##########")


    #Neg Log-Likelihood Value
    nloglikexp=expfit$results$value

    #Deviance Stationary Exponential
    devexp=2*nloglikexp



    #Non-stationary Exponential model
    tempo=datahum_JJA[,1]-1979
    expfit_nonstat=fevd(declus, threshold = 2, method = "MLE",  type = "
Exponential", time.units="92/year", scale.fun= ~ tempo, use.phi=T)


    #phi_0
    phi0=summary(expfit_nonstat)$par[1]
    #Std.err. phi_0
    se_phi0=summary(expfit_nonstat)$se.theta[1]

    #phi_1
    phi1=summary(expfit_nonstat)$par[2]
    #Std.err. phi_1
    se_phi1=summary(expfit_nonstat)$se.theta[2]



    #Negative Log-Likelihood Value
    nloglikexp_nonstat=expfit_nonstat$results$value


    #Deviance Non-stationary Exponential
    devexp_nonstat=2*nloglikexp_nonstat


    #Difference of Deviances
    difdev=devexp-devexp_nonstat

    #p-value
    pvalue=pchisq(difdev,df=1,lower.tail=F)


    ####  Estimated 38, 50 and 100 year-return levels (Non-stationary case)

    exp_nonstat_lev=plot(expfit_nonstat, "rl",rperiods=c(38,50,100))$level

    #Summer 1980
    head(exp_nonstat_lev)

    #Summer 2017
    tail(exp_nonstat_lev)


    ####  Return-level plot: Non-stationary case
```

```
        plot(expfit_nonstat, "rl",rperiods=c(38,50,100),pch=20,axes=F,ann=F,cex
    =0.8, ylim=c(0,7))

        mtext("Year",side=1,line=2.5,font=1)
        mtext("Return Level (in mm/day)",side=2,line=2.5,font=1)
        box()
        axis(2,at=axTicks(2))
        abline(h=0,lty=3)
        num<-92
        pontos<-seq(1,length(hum_JJA),by=num)
        axis(1,at=pontos,labels=F)
        lab=numeric()
        lab[seq(1,38,by=3)]=seq(1980,2017,by=3)
        lab[38]=NA

        lab=as.character(lab)

        text(-0.5,x=seq(1,length(hum_JJA),by=num),labels=lab,srt=40,
        pos=c(1,1),xpd=T,las=2,cex=0.8)


        #######




}


##########
```

B.2 Bivariate extremes and copulas

B.2.1 Theoretical example of Pickands Dependence Functions

```
library(evd)
```

```
abvevd(dep = 0.5,  model = "log", plot = TRUE, rev=T)
abvevd(dep = 0.7,  model = "log", add = TRUE, rev=T, col="red")
abvevd(dep = 0.9,  model = "log", add = TRUE, rev=T, col="blue")

legend("bottomright", legend=c(expression(paste(alpha,"=0.9")), expression(paste
    (alpha,"=0.7")), expression(paste(alpha,"=0.5")) ),col=c("blue","red","black
    "), lty=c(1,1,1), cex=1.3)
```

B.2.2 Practical part

```
library(readr)
library(VineCopula)
library(evd)
library(KScorrect)
library(gofCopula)
library(evmix)
library(moments)



#OMEGA (SUMMER)


serie_w_p75 <- read.delim("serie_w_p75.txt")
datawp75 = as.data.frame(serie_w_p75)
JJA=which((datawp75$mes==6 | datawp75$mes==7) | datawp75$mes==8)
datawp75_JJA=datawp75[JJA,]
w_JJA= datawp75_JJA[,4]


#PLOT OF THE SERIES
n=length(-w_JJA)


plot(-w_JJA,type=’l’, axes=F ,ann=F)

mtext("Year",side=1,line=2.5,font=1)
mtext("-omega (Pa/s)",side=2,line=2.5,font=1)
mtext("-omega",side=3,line=2,font=2,cex=1.3)
```

```
box()
axis(2,at=axTicks(2))
abline(h=0,lty=3)
num<-92
pontos<-seq(1,n,by=num)
axis(1,at=pontos,labels=F)
lab=numeric()
lab[seq(1,38,by=3)]=seq(1980,2017,by=3)
lab[38]=NA

lab=as.character(lab)

text(-0.07,x=seq(1,n,by=num),labels=lab,srt=40,
pos=c(1,1),xpd=T,las=2,cex=0.8)


#####



summ_menosomega=summary(-w_JJA)
sd(-w_JJA)

summ_menosomega=c(summary(-w_JJA),sd(-w_JJA))
names(summ_menosomega)[7]="Std.Dev."


# PRECIPITATION (SUMMER)


serie_pcp = read_table2("serie_pcp_p75.txt")
dataprec = as.data.frame(serie_pcp)
JJA=which((dataprec$mes==6 | dataprec$mes==7) | dataprec$mes==8)
dataprec_JJA=dataprec[JJA,]
prec_JJA= dataprec_JJA[,4]
fechas=dataprec_JJA[,1:3]



#PLOT OF THE SERIES
n=length(prec_JJA)

plot(prec_JJA,type='l', axes=F ,ann=F)
```

```
mtext("Year",side=1,line=2.5,font=1)
mtext("Precipitation (mm/day)",side=2,line=2.5,font=1)
mtext("Precipitation",side=3,line=2,font=2,cex=1.3)

box()
axis(2,at=axTicks(2))
abline(h=0,lty=3)
num<-92
pontos<-seq(1,n,by=num)
axis(1,at=pontos,labels=F)
lab=numeric()
lab[seq(1,38,by=3)]=seq(1980,2017,by=3)
lab[38]=NA

lab=as.character(lab)

text(-0.92,x=seq(1,n,by=num),labels=lab,srt=40,
pos=c(1,1),xpd=T,las=2,cex=0.8)


####



summ_prec=summary(prec_JJA)

sd(prec_JJA)

summ_prec=c(summary(prec_JJA),sd(prec_JJA))
names(summ_prec)[7]="Std.Dev."



######## DIVIDING THE SAMPLE BY MOISTURE


serie_hum <- read_table2("serie_p75_i3-5.txt")
datahum=as.data.frame(serie_hum)
JJA=which( datahum$mes==6 | datahum$mes==7 | datahum$mes==8 )
datahum_JJA=datahum[JJA,]
hum_JJA=datahum_JJA[,4]
```

```
lag=1


if(lag>0){
        preclag=prec_JJA[-c(1:lag)]
        wlag=w_JJA[-c(1:lag)]
        fechaslag=fechas[-c(1:lag),]
        humlag=hum_JJA[-c((length(hum_JJA)-lag+1):(length(hum_JJA)))]
}else{
        preclag=prec_JJA
        wlag=w_JJA
        fechaslag=fechas
        humlag=hum_JJA
}


wprec_hum_dados=cbind(-wlag,preclag,humlag,fechaslag)

baja=which(humlag<quantile(humlag,prob=0.25))
alta=which(humlag>quantile(humlag,prob=0.75))


wprec_humbaja_fechas=wprec_hum_dados[baja,]
wprec_humalta_fechas=wprec_hum_dados[alta,]

wprec_humbaja=wprec_humbaja_fechas[,1:2]
wprec_humalta=wprec_humalta_fechas[,1:2]



#### A BRIEF EXPLORATORY ANALYSIS

#BOXPLOTS

boxplot(wprec_humbaja[,1], wprec_humalta[,1], main="-omega",ylab="-omega (Pa/s)
   ", names=c("LOW TM","HIGH TM"))
boxplot(wprec_humbaja[,2], wprec_humalta[,2], main="Precipitation", ylab="
   Precipitation (mm/day)", names=c("LOW TM","HIGH TM") )


#HISTROGRAMS

hist(wprec_humbaja[,1],prob=T, main="-omega (Low TM)",xlab="-omega (Pa/s)")
lines(density(wprec_humbaja[,1]),lwd=1)
```

```
hist(wprec_humbaja[,2],prob=T, main="prec. (Low TM)",xlab="prec. (mm/day)")
lines(density(wprec_humbaja[,2]),lwd=1)



hist(wprec_humalta[,1],prob=T, main="-omega (High TM)",xlab="-omega (Pa/s)")
lines(density(wprec_humalta[,1]),lwd=1)



hist(wprec_humalta[,2],prob=T, main="prec. (High TM)",xlab="prec. (mm/day)")
lines(density(wprec_humalta[,2]),lwd=1)


###


####stand.dev.
stddev=numeric()
#-omega
stddev[1]=sd(wprec_humbaja[,1])
stddev[2]=sd(wprec_humalta[,1])

#prec.
stddev[3]=sd(wprec_humbaja[,2])
stddev[4]=sd(wprec_humalta[,2])


####skewness
sk=numeric()
#-omega
sk[1]=skewness(wprec_humbaja[,1])
sk[2]=skewness(wprec_humalta[,1])

#prec.
sk[3]=skewness(wprec_humbaja[,2])
sk[4]=skewness(wprec_humalta[,2])



####kurtosis
kur=numeric()
```

```
#-omega

kur[1]=kurtosis(wprec_humbaja[,1])
kur[2]=kurtosis(wprec_humalta[,1])

#prec.

kur[3]=kurtosis(wprec_humbaja[,2])
kur[4]=kurtosis(wprec_humalta[,2])

sd_sk_kur=rbind(stddev,sk,kur)
colnames(sd_sk_kur)=c("-omega_low","-omega_high", "prec_low","prec_high")



######


####  LOW MOISTURE


#MARGINAL THRESHOLD MODELS


#VARIABLE X1
#-omega


#THRESHOLD SELECTION

#MEAN EXCESS FUNCTION
evmix::mrlplot(wprec_humbaja[,1],try.thresh=0.03,legend.loc=NULL,main="Mean
   Residual Life Plot for -omega (Low TM)")

#SHAPE PLOT
evmix::tshapeplot(wprec_humbaja[,1],try.thresh=0.03,nt=80,legend.loc=NULL, main
   ="Shape Plot for -omega (Low TM)")

#MODIFIED SCALE PLOT
evmix::tscaleplot(wprec_humbaja[,1],try.thresh=0.03,nt=80,legend.loc=NULL, main
   ="Modified Scale Plot for -omega (Low TM)")


#THRESHOLD FOR -OMEGA _ LOW MOISTURE
uw_humbaja=0.03
```

```
fitpot=fpot(wprec_humbaja[,1], threshold=uw_humbaja, method="Nelder-Mead")


#PROFILE CONF.INT. (shape parameter) -OMEGA _ LOW MOISTURE


prof=profile(fitpot,conf=0.95,which="shape", mesh=0.008)

plot(prof,xlim=c(-0.35,0.05),main="Prof. Log-lik. of Shape (-omega _ LOW TM)")
CI=confint(prof)
abline(v=c(CI[1],CI[2]))
abline(v=fitpot$estimate[2], col="blue")



#####

excesses1=wprec_humbaja[,1][wprec_humbaja[,1]>uw_humbaja]-uw_humbaja

#Number of excesses
Nu1=length(excesses1)

#Percentage of excesses
(Nu1/length(wprec_humbaja[,1]))*100


NU1<-1:Nu1




#scale
sigma<-fitpot$estimate[1]

#Std.err. scale
fitpot$std.err[1]


#shape
xi<-fitpot$estimate[2]

#Std.err. shape
fitpot$std.err[2]
```

```
#CHECKING GPD CONDITION

max(wprec_humbaja[,1])
uw_humbaja-sigma/xi



#GPD GOODNESS-OF-FIT

#Cramer-von Mises
CVM<-sum((evd::pgpd(sort(excesses1),scale=sigma, shape=xi)-
(2*NU1-1)/(2*Nu1))^2)+1/(12*Nu1)
#Anderson-Darling
AD<--Nu1-(1/Nu1)*sum((2*NU1-1)*log(evd::pgpd(sort(excesses1),scale =sigma,shape
    =xi))+(2*Nu1+1-2*NU1)*log(1-evd::pgpd(sort(excesses1),scale=sigma, shape=xi)
    ))


#EXPONENTIAL QQPLOT - EXCESSES


pppot1<-(1:Nu1)/(Nu1+1)

qqpot1<--log(1-pppot1)

fitexppot1<-lm(sort(excesses1)~qqpot1-1)

summary(fitexppot1)


summary(fitexppot1)$r.squared




plot(qqpot1,sort(excesses1),ylab='empirical quantiles',
xlab='exponential quantiles', main='Exponential QQ-Plot (-omega _ LOW TM)')


abline(fitexppot1, lty=2,col='red', cex=3)
```

```
#GPD QQPLOT - EXCESSES


QQgpd<-evd::qgpd(pppot1,shape=xi)

fitgpd<-lm(sort(excesses1) ~ QQgpd-1)
summary(fitgpd)

summary(fitgpd)$r.squared



plot(QQgpd,sort(excesses1), ylab='empirical quantiles',
xlab='GPD quantiles',main='GPD QQ-Plot (-omega _ LOW TM)')

abline(fitgpd, lty=2,col='blue', cex=3)




## LIKELIHOOD RATIO TEST: EXPONENTIAL VS GPD


devgpd=fitpot$deviance
fitpot_exp=fpot(wprec_humbaja[,1], threshold=uw_humbaja, model="gpd",shape=0)
devexp=fitpot_exp$deviance


difdev=devexp-devgpd

pvalue=pchisq(difdev,df=1,lower.tail=F)




## VARIABLE X2

#PRECIPITATION

#THRESHOLD SELECTION

#MEAN EXCESS FUNCTION
evmix::mrlplot(wprec_humbaja[,2],try.thresh=5.2,tlim=c(2,6.5),legend.loc=NULL,
   main="Mean Residual Life Plot for prec. (Low TM)")

#SHAPE PLOT
evmix::tshapeplot(wprec_humbaja[,2],try.thresh=5.2,tlim=c(2,6.5),legend.loc=NULL
   ,main="Shape Plot for prec. (Low TM)")

#MODIFIED SCALE PLOT
evmix::tscaleplot(wprec_humbaja[,2],try.thresh=5.2,tlim=c(2,6.5),legend.loc=NULL
   ,main="Modified Scale Plot for prec. (Low TM)")




#THRESHOLD FOR PRECIPITATION _ LOW MOISTURE
uprec_humbaja=5.2

fitpot2=fpot(wprec_humbaja[,2], threshold=uprec_humbaja, method="Nelder-Mead")




#PROFILE CONF.INT. (shape parameter)  PRECIPITATION _ LOW MOISTURE



prof2=profile(fitpot2,conf=0.95,which="shape", mesh=0.008)


plot(prof2,xlim=c(-0.4,0.3),main="Prof. Log-lik. of Shape (prec. _ LOW TM)")
CI_2=confint(prof2)
abline(v=c(CI_2[1],CI_2[2]))
abline(v=fitpot2$estimate[2], col="blue")




#####



excesses2=wprec_humbaja[,2][wprec_humbaja[,2]>uprec_humbaja]-uprec_humbaja

#Number of excesses
Nu2=length(excesses2)
```

```
#Percentage of excesses
(Nu2/length(wprec_humbaja[,2]))*100



NU2<-1:Nu2

#scale
sigma2<-fitpot2$estimate[1]

#Std.err. scale
fitpot2$std.err[1]


#shape
xi2<-fitpot2$estimate[2]


#Std.err. shape
fitpot2$std.err[2]



#CHECKING GPD CONDITION

max(wprec_humbaja[,2])
uprec_humbaja-sigma2/xi2




#GPD GOODNESS-OF-FIT

#Cramer-von Mises
CVM2<-sum((evd::pgpd(sort(excesses2),scale=sigma2, shape=xi2)-
(2*NU2-1)/(2*Nu2))^2)+1/(12*Nu2)
#Anderson-Darling
AD2<--Nu2-(1/Nu2)*sum((2*NU2-1)*log(evd::pgpd(sort(excesses2),scale=sigma2,shape
    =xi2))+(2*Nu2+1-2*NU2)*log(1-evd::pgpd(sort(excesses2),scale=sigma2, shape=
    xi2)))


#EXPONENTIAL QQPLOT - EXCESSES
```

```
pppot2<-(1:Nu2)/(Nu2+1)
qqpot2<--log(1-pppot2)
fitexppot2<-lm(sort(excesses2)~qqpot2-1)

summary(fitexppot2)

summary(fitexppot2)$r.squared



plot(qqpot2,sort(excesses2),ylab='empirical quantiles',
xlab='exponential quantiles', main='Exponential QQ-Plot (prec. _ LOW TM)')
abline(fitexppot2, lty=2,col='red', cex=3)




#GPD QQPLOT - EXCESSES


QQgpd<-evd::qgpd(pppot2,shape=xi2)
fitgpd2<-lm(sort(excesses2) ~ QQgpd-1)
summary(fitgpd2)


summary(fitgpd2)$r.squared



plot(QQgpd,sort(excesses2), ylab='empirical quantiles',
xlab='GPD quantiles',main='GPD QQ-Plot (prec. _ LOW TM)')
abline(fitgpd2, lty=2,col='blue', cex=3)





## LIKELIHOOD RATIO TEST: EXPONENTIAL VS GPD


devgpd2=fitpot2$deviance
fitpot2_exp=fpot(wprec_humbaja[,2], threshold=uprec_humbaja, model="gpd",shape
    =0)
```

```
sigmaexp_prec_humbaja=fitpot2_exp$estimate[1]
sigmaexp_prec_humbaja_stderr=fitpot2_exp$std.err[1]



devexp2=fitpot2_exp$deviance

difdev2=devexp2-devgpd2


pvalue2=pchisq(difdev2,df=1,lower.tail=F)




#LcKS Exponential for prec. _ LOW TM


excesses2=wprec_humbaja[,2][wprec_humbaja[,2]>uprec_humbaja]-uprec_humbaja
out=LcKS(excesses2, cdf="pexp", nreps = 5000)

out$D.obs
out$p.value





#BIVARIATES EXTREMES


## SCATTERPLOT


both_humbaja=length(which((wprec_humbaja[,1] > uw_humbaja) & (wprec_humbaja[,2]
    > uprec_humbaja)))
onlymenosomega_humbaja=length(which((wprec_humbaja[,1] > uw_humbaja) & (
    wprec_humbaja[,2] <= uprec_humbaja)))
onlyprec_humbaja=length(which((wprec_humbaja[,1] <= uw_humbaja) & (wprec_humbaja
    [,2] > uprec_humbaja)))
none_humbaja=length(which((wprec_humbaja[,1] <= uw_humbaja) & (wprec_humbaja[,2]
     <= uprec_humbaja)))


both_humbaja_perc=(both_humbaja/nrow(wprec_humbaja))*100
```

```
onlymenosomega_humbaja_perc=(onlymenosomega_humbaja/nrow(wprec_humbaja))*100
onlyprec_humbaja_perc=(onlyprec_humbaja/nrow(wprec_humbaja))*100



plot(wprec_humbaja[,1],wprec_humbaja[,2], main="LOW TM", xlab="-omega (Pa/s)",
    ylab="precipitation (mm/day)",xlim=c(-0.06,0.10),ylim=c(0,12))
abline(v=uw_humbaja,col="red")
abline(h=uprec_humbaja,col="blue")
legend("topleft", ncol=1, legend=c(paste("n=",onlyprec_humbaja," (",sprintf("%.2
    f",round(onlyprec_humbaja_perc,2)),"% )")),cex=0.8)
legend("topright", ncol=1, legend=c(paste("n=",both_humbaja, " (",sprintf("%.2f
    ",round(both_humbaja_perc,2)),"% )")),cex=0.8)
legend("bottomright", ncol=1, legend=c(paste("n=",onlymenosomega_humbaja, " (",
    sprintf("%.2f",round(onlymenosomega_humbaja_perc,2)),"% )" )),cex=0.8)
text(0.095,2,"A",cex=1.75)
text(-0.03,9,"B",cex=1.75)
text(0.08,10,"C",cex=1.75)




## CHI-PLOT AND CHI BAR-PLOT

chiplot(wprec_humbaja,which=1, main1="Chi Plot (LOW TM)",xlab="u",ylab1=bquote(
    chi(u)))
chiplot(wprec_humbaja,which=2, main2="Chi Bar Plot (LOW TM)",xlab="u",ylab2=
    bquote(bar(chi)(u)))



#PARAMETRIC MODELS


#1) Logistic model
fit1_log=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),model="
    log",shape2=0,likelihood=’censored’, method="Nelder-Mead")
AIC(fit1_log)

fit1_log$dep.summary[[1]]



#2) Asymmetric logistic model
fit1_alog=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),model
   ="alog",shape2=0,likelihood='censored', method="CG",std.err = FALSE)
AIC(fit1_alog)

fit1_alog$dep.summary[[1]]


#3)Husler-Reiss model

fit1_hr=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),model="
   hr",shape2=0,likelihood='censored', method="Nelder-Mead")
AIC(fit1_hr)

fit1_hr$dep.summary[[1]]


#4)Negative logistic model
fit1_neglog=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),
   model="neglog",shape2=0,likelihood='censored', method="Nelder-Mead")
AIC(fit1_neglog)

fit1_neglog$dep.summary[[1]]


#5)Asymmetric negative logistic model
fit1_aneglog=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),
   model="aneglog",shape2=0,likelihood='censored',std.err = FALSE)
AIC(fit1_aneglog)

fit1_aneglog$dep.summary[[1]]


#6)Bilogistic model
fit1_bilog=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),model
   ="bilog",shape2=0,likelihood='censored',method="Nelder-Mead")
AIC(fit1_bilog)

fit1_bilog$dep.summary[[1]]


alpha_bilog=fit1_bilog$estimate[4]
beta_bilog=fit1_bilog$estimate[5]

alpha_bilog_std_err=fit1_bilog$std.err[4]
beta_bilog_std_err=fit1_bilog$std.err[5]
```

```
estimates_bilog=fit1_bilog$estimate
stderr_bilog=fit1_bilog$std.err


#PICKANDS DEPENDENCE FUNCTION
plot(fit1_bilog,which=2,main="Bilogistic model for LOW TM")


#QUANTILE CURVES OF THE JOINT DISTRIBUTION FUNCTION
plot(fit1_bilog, which = 3, p = c(0.95,0.975,0.99), tlty = 0, col="blue",xlab="-
    omega (Pa/s)", ylab="precipitation (mm/day)",xlim=c(-0.06,0.10),ylim=c(0,12)
    ,main="Bilogistic model for LOW TM")
abline(v=uw_humbaja,col="red")
abline(h=uprec_humbaja,col="red")

###


#7)Negative bilogistic model
fit1_negbilog=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),
    model="negbilog",shape2=0,likelihood='censored', method="L-BFGS-B",std.err =
     FALSE)
AIC(fit1_negbilog)

fit1_negbilog$dep.summary[[1]]


#8)Coles-Tawn model
fit1_ct=evd::fbvpot(wprec_humbaja,threshold=c(uw_humbaja,uprec_humbaja),model="
    ct",shape2=0,likelihood='censored', method="CG",std.err = F)
AIC(fit1_ct)

fit1_ct$dep.summary[[1]]




####COPULAS


n=nrow(wprec_humbaja)

uni= apply(wprec_humbaja, 2, rank)/(n + 1)
```

```
#PSEUDO-OBSERVATIONS (LOW TM)
plot(uni[,1], uni[,2], xlab="-omega", ylab="precipitation",main="Pseudo-observations (Low TM)")


#TRYING THE FIVE COPULAS
BiCopSelect(uni[,1],uni[,2],familyset=1)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=2)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=3)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=4)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=5)$AIC


#THE BEST COPULA MODEL: STUDENT-T

bicopsel=BiCopSelect(uni[,1],uni[,2],familyset=1:5)
summary(bicopsel)

summary(bicopsel)$par
summary(bicopsel)$par2
summary(bicopsel)$tau
summary(bicopsel)$taildep$lower
summary(bicopsel)$taildep$upper
summary(bicopsel)$AIC


#GUMBEL COPULA

bicopsel2=BiCopSelect(uni[,1],uni[,2],familyset=4)
summary(bicopsel2)

summary(bicopsel2)$par
summary(bicopsel2)$tau
summary(bicopsel2)$taildep$lower
summary(bicopsel2)$taildep$upper
summary(bicopsel2)$AIC


#Goodness of fit


gofSn(copula="t",x=as.matrix(wprec_humbaja),processes=6)
gofSn(copula="gumbel",x=as.matrix(wprec_humbaja),processes=6)

gofRosenblattSnB(copula="t",x=as.matrix(wprec_humbaja), processes=7)
gofRosenblattSnB(copula="gumbel",x=as.matrix(wprec_humbaja), processes=7)



#COPULA DENSITY: STUDENT-T

plot(bicopsel, type = "surface",main="Student t-copula  (Low TM)",xlab="-omega",
    ylab="prec.")


#Simulated copula data: STUDENT-T

simdata <- BiCopSim(nrow(wprec_humbaja), bicopsel)


plot(simdata[,1],simdata[,2], xlab="-omega", ylab="precipitation", main="
    Simulated data: Student t-copula  (Low TM)")




###########


### HIGH MOISTURE



#MARGINAL THRESHOLD MODELS


#VARIABLE X1
#-omega

#THRESHOLD SELECTION


#MEAN EXCESS FUNCTION
evmix::mrlplot(wprec_humalta[,1],try.thresh=0.03,legend.loc=NULL, main="Mean
    Residual Life Plot for -omega (High TM)")


#SHAPE PLOT
evmix::tshapeplot(wprec_humalta[,1],nt=64,legend.loc=NULL, try.thresh=0.03, main
    ="Shape Plot for -omega (High TM)")
```

```
#MODIFIED SCALE PLOT
evmix::tscaleplot(wprec_humalta[,1],nt=64,legend.loc=NULL, try.thresh=0.03, main
   ="Modified Scale Plot for -omega (High TM)")



#THRESHOLD FOR -OMEGA _ HIGH MOISTURE
uw_humalta=0.03

fitpot=fpot(wprec_humalta[,1], threshold=uw_humalta, method="Nelder-Mead")


#PROFILE CONF.INT. (shape parameter)  -OMEGA _ HIGH MOISTURE


prof=profile(fitpot,conf=0.95,which="shape", mesh=0.008)


plot(prof,xlim=c(-0.4,0),main="Prof. Log-lik. of Shape (-omega _ HIGH TM)")
CI=confint(prof)
abline(v=c(CI[1],CI[2]))
abline(v=fitpot$estimate[2], col="blue")





#####

excesses1=wprec_humalta[,1][wprec_humalta[,1]>uw_humalta]-uw_humalta

#Number of excesses
Nu1=length(excesses1)


#Percentage of excesses
(Nu1/length(wprec_humalta[,1]))*100



NU1<-1:Nu1
```

```
#scale
sigma<-fitpot$estimate[1]

#Std.err. scale
fitpot$std.err[1]


#shape
xi<-fitpot$estimate[2]

#Std.err. shape
fitpot$std.err[2]




#CHECKING GPD CONDITION

max(wprec_humalta[,1])
uw_humalta-sigma/xi



#GPD GOODNESS-OF-FIT

#Cramer-von Mises
CVM<-sum((evd::pgpd(sort(excesses1),scale=sigma, shape=xi)-
(2*NU1-1)/(2*Nu1))^2)+1/(12*Nu1)
#Anderson-Darling
AD<--Nu1-(1/Nu1)*sum((2*NU1-1)*log(evd::pgpd(sort(excesses1),scale =sigma,shape
    =xi))+(2*Nu1+1-2*NU1)*log(1-evd::pgpd(sort(excesses1),scale=sigma, shape=xi)
    ))



#EXPONENTIAL QQPLOT - EXCESSES


pppot1<-(1:Nu1)/(Nu1+1)

qqpot1<--log(1-pppot1)

fitexppot1<-lm(sort(excesses1)~qqpot1-1)

summary(fitexppot1)
summary(fitexppot1)$r.squared


plot(qqpot1,sort(excesses1),ylab='empirical quantiles',
xlab='exponential quantiles', main='Exponential QQ-Plot (-omega _ HIGH TM)')


abline(fitexppot1, lty=2,col='red', cex=3)




#GPD QQPLOT - EXCESSES


QQgpd<-evd::qgpd(pppot1,shape=xi)

fitgpd<-lm(sort(excesses1) ~ QQgpd-1)
summary(fitgpd)

summary(fitgpd)$r.squared



plot(QQgpd,sort(excesses1), ylab='empirical quantiles',
xlab='GPD quantiles',main='GPD QQ-Plot (-omega _ HIGH TM)')

abline(fitgpd, lty=2,col='blue', cex=3)



## LIKELIHOOD RATIO TEST: EXPONENTIAL VS GPD


devgpd=fitpot$deviance
fitpot_exp=fpot(wprec_humalta[,1], threshold=uw_humalta, model="gpd",shape=0)
devexp=fitpot_exp$deviance

difdev=devexp-devgpd


pvalue=pchisq(difdev,df=1,lower.tail=F)
```

```
## VARIABLE X2

#PRECIPITATION

#THRESHOLD SELECTION


#MEAN EXCESS FUNCTION
evmix::mrlplot(wprec_humalta[,2],try.thresh=5.2,tlim=c(0,8.5),legend.loc=NULL,
   main="Mean Residual Life Plot for prec. (High TM)")


#SHAPE PLOT
evmix::tshapeplot(wprec_humalta[,2],try.thresh=5.2,tlim=c(0,8.25),legend.loc=
   NULL, main="Shape Plot for prec. (High TM)")


#MODIFIED SCALE PLOT
evmix::tscaleplot(wprec_humalta[,2],try.thresh=5.2,tlim=c(0,8.25),legend.loc=
   NULL, main="Modified Scale Plot for prec. (High TM)")



#THRESHOLD FOR PRECIPITATION _ HIGH MOISTURE
uprec_humalta=5.2

fitpot2=fpot(wprec_humalta[,2], threshold=uprec_humalta, method="Nelder-Mead")




#PROFILE CONF.INT. (shape parameter)  PRECIPITATION _ HIGH MOISTURE



prof2=profile(fitpot2,conf=0.95,which="shape", mesh=0.008)

plot(prof2,xlim=c(-0.35,0.1),main="Prof. Log-lik. of Shape (prec. _ HIGH TM)")
CI_2=confint(prof2)
abline(v=c(CI_2[1],CI_2[2]))
abline(v=fitpot2$estimate[2], col="blue")
```

```
#####


excesses2=wprec_humalta[,2][wprec_humalta[,2]>uprec_humalta]-uprec_humalta

#Number of excesses
Nu2=length(excesses2)


#Percentage of excesses
(Nu2/length(wprec_humalta[,2]))*100


NU2<-1:Nu2


#scale
sigma2<-fitpot2$estimate[1]


#Std.err. scale
fitpot2$std.err[1]


#shape
xi2<-fitpot2$estimate[2]


#Std.err. shape
fitpot2$std.err[2]




#CHECKING GPD CONDITION

max(wprec_humalta[,2])
uprec_humalta-sigma2/xi2






#GPD GOODNESS-OF-FIT
```

```
#Cramer-von Mises
CVM2<-sum((evd::pgpd(sort(excesses2),scale=sigma2, shape=xi2)-
(2*NU2-1)/(2*Nu2))^2)+1/(12*Nu2)
#Anderson-Darling
AD2<--Nu2-(1/Nu2)*sum((2*NU2-1)*log(evd::pgpd(sort(excesses2),scale=sigma2,shape
   =xi2))+(2*Nu2+1-2*NU2)*log(1-evd::pgpd(sort(excesses2),scale=sigma2, shape=
   xi2)))

#EXPONENTIAL QQPLOT - EXCESSES

pppot2<-(1:Nu2)/(Nu2+1)
qqpot2<--log(1-pppot2)
fitexppot2<-lm(sort(excesses2)~qqpot2-1)

summary(fitexppot2)
summary(fitexppot2)$r.squared


plot(qqpot2,sort(excesses2),ylab='empirical quantiles',
xlab='exponential quantiles', main='Exponential QQ-Plot (prec. _ HIGH TM)')
abline(fitexppot2, lty=2,col='red', cex=3)



#GPD QQPLOT - EXCESSES


QQgpd<-evd::qgpd(pppot2,shape=xi2)
fitgpd2<-lm(sort(excesses2) ~ QQgpd-1)
summary(fitgpd2)

summary(fitgpd2)$r.squared


plot(QQgpd,sort(excesses2), ylab='empirical quantiles',
xlab='GPD quantiles',main='GPD QQ-Plot (prec. _ HIGH TM)')
abline(fitgpd2, lty=2,col='blue', cex=3)



## LIKELIHOOD RATIO TEST: EXPONENTIAL VS GPD


devgpd2=fitpot2$deviance
```

```
fitpot2_exp=fpot(wprec_humalta[,2], threshold=uprec_humalta, model="gpd",shape
    =0)

sigmaexp_prec_humalta=fitpot2_exp$estimate[1]
sigmaexp_prec_humalta_stderr=fitpot2_exp$std.err[1]




devexp2=fitpot2_exp$deviance

difdev2=devexp2-devgpd2


pvalue2=pchisq(difdev2,df=1,lower.tail=F)




#LcKS Exponential for prec. _ HIGH TM

excesses2=wprec_humalta[,2][wprec_humalta[,2]>uprec_humalta]-uprec_humalta
out=LcKS(excesses2, cdf="pexp", nreps = 5000)

out$D.obs
out$p.value





#BIVARIATES EXTREMES


#SCATTERPLOT


both_humalta=length(which((wprec_humalta[,1] > uw_humalta) & (wprec_humalta[,2]
    > uprec_humalta)))
onlymenosomega_humalta=length(which((wprec_humalta[,1] > uw_humalta) & (
    wprec_humalta[,2] <= uprec_humalta)))
onlyprec_humalta=length(which((wprec_humalta[,1] <= uw_humalta) & (wprec_humalta
    [,2] > uprec_humalta)))
none_humalta=length(which((wprec_humalta[,1] <= uw_humalta) & (wprec_humalta[,2]
     <= uprec_humalta)))
```

```
both_humalta_perc=(both_humalta/nrow(wprec_humalta))*100
onlymenosomega_humalta_perc=(onlymenosomega_humalta/nrow(wprec_humalta))*100
onlyprec_humalta_perc=(onlyprec_humalta/nrow(wprec_humalta))*100



plot(wprec_humalta[,1],wprec_humalta[,2],main="HIGH TM",xlab="-omega (Pa/s)",
    ylab="precipitation (mm/day)",xlim=c(-0.06,0.10),ylim=c(0,12))
abline(v=uw_humalta,col="red")
abline(h=uprec_humalta,col="blue")

legend("topleft", ncol=1, legend=c(paste("n=",onlyprec_humalta," (",sprintf("%.2
    f",round(onlyprec_humalta_perc,2)),"% )")),cex=0.8)
legend("topright", ncol=1, legend=c(paste("n=",both_humalta," (",sprintf("%.2f",
    round(both_humalta_perc,2)),"% )")),cex=0.8)
legend("bottomright", ncol=1, legend=c(paste("n=",onlymenosomega_humalta," (",
    sprintf("%.2f",round(onlymenosomega_humalta_perc,2)),"% )")),cex=0.8)
text(0.095,2,"A",cex=1.75)
text(-0.03,9,"B",cex=1.75)
text(0.09,8,"C",cex=1.75)



#CHI-PLOT AND CHI BAR-PLOT


chiplot(wprec_humalta,which=1,main1="Chi Plot (HIGH TM)",xlab="u",ylab1=bquote(
    chi(u)))
chiplot(wprec_humalta,which=2,main2="Chi Bar Plot (HIGH TM)",xlab="u",ylab2=
    bquote(bar(chi)(u)))



#PARAMETRIC MODELS

#1) Logistic model
fit2_log=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),model="
    log",shape2=0,likelihood='censored', method="Nelder-Mead")
AIC(fit2_log)
fit2_log$dep.summary[[1]]

alpha_humalta=fit2_log$estimate[4]
std_err_alpha_humalta=fit2_log$std.err[4]
```

```
pval_log_humalta=evind.test(wprec_humalta, method="score")$p.value

estimates_log=fit2_log$estimate
stderr_log=fit2_log$std.err


#PICKANDS DEPENDENCE FUNCTION
plot(fit2_log,which=2, main="Logistic model for HIGH TM")

#QUANTILE CURVES OF THE JOINT DISTRIBUTION FUNCTION
plot(fit2_log, which = 3, p = c(0.95,0.975,0.99), tlty = 0, col="blue",xlab="-
    omega (Pa/s)", ylab="precipitation (mm/day)",xlim=c(-0.06,0.10),ylim=c(0,12)
    ,main="Logistic model for HIGH TM")
abline(v=uw_humalta,col="red")
abline(h=uprec_humalta,col="red")




#2) Asymmetric logistic model
fit2_alog=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),model
    ="alog",shape2=0, likelihood=’censored’, method="CG")
AIC(fit2_alog)
fit2_alog$dep.summary[[1]]

#3)Husler-Reiss model

fit2_hr=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),model="
    hr",shape2=0,likelihood=’censored’, method="Nelder-Mead")
AIC(fit2_hr)
fit2_hr$dep.summary[[1]]


#4)Negative logistic model
fit2_neglog=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),
    model="neglog",shape2=0,likelihood=’censored’)
AIC(fit2_neglog)
fit2_neglog$dep.summary[[1]]

#5)Asymmetric negative logistic model
fit2_aneglog=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),
    model="aneglog",shape2=0,likelihood=’censored’)
AIC(fit2_aneglog)
fit2_aneglog$dep.summary[[1]]
```

```
#6)Bilogistic model
fit2_bilog=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),model
    ="bilog",shape2=0,likelihood='censored', method="CG")
AIC(fit2_bilog)
fit2_bilog$dep.summary[[1]]

#7)Negative bilogistic model
fit2_negbilog=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),
    model="negbilog",shape2=0,likelihood='censored', method="CG", std.err =
    FALSE)
AIC(fit2_negbilog)
fit2_negbilog$dep.summary[[1]]


#8)Coles-Tawn model
fit2_ct=evd::fbvpot(wprec_humalta,threshold=c(uw_humalta,uprec_humalta),model="
    ct",shape2=0,likelihood='censored')
AIC(fit2_ct)
fit2_ct$dep.summary[[1]]




####COPULAS

n=nrow(wprec_humalta)


uni= apply(wprec_humalta, 2, rank)/(n + 1)


#PSEUDO-OBSERVATIONS (LOW TM)
plot(uni[,1], uni[,2], xlab="-omega", ylab="precipitation",main="Pseudo-
    observations (High TM)")


#TRYING THE FIVE COPULAS
BiCopSelect(uni[,1],uni[,2],familyset=1)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=2)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=3)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=4)$AIC
BiCopSelect(uni[,1],uni[,2],familyset=5)$AIC


#THE BEST COPULA MODEL: GUMBEL
```

```
bicopsel=BiCopSelect(uni[,1],uni[,2],familyset=1:5)
summary(bicopsel)


summary(bicopsel)$par
summary(bicopsel)$tau
summary(bicopsel)$taildep$lower
summary(bicopsel)$taildep$upper
summary(bicopsel)$AIC


#STUDENT-T COPULA

bicopsel2=BiCopSelect(uni[,1],uni[,2],familyset=2)
summary(bicopsel2)


summary(bicopsel2)$par
summary(bicopsel2)$par2
summary(bicopsel2)$tau
summary(bicopsel2)$taildep$lower
summary(bicopsel2)$taildep$upper
summary(bicopsel2)$AIC


#Goodness of fit


gofSn(copula="t",x=as.matrix(wprec_humalta),processes=6)
gofSn(copula="gumbel",x=as.matrix(wprec_humalta),processes=6)

gofRosenblattSnB(copula="t",x=as.matrix(wprec_humalta),processes=7)
gofRosenblattSnB(copula="gumbel",x=as.matrix(wprec_humalta),processes=7)



#COPULA DENSITY: GUMBEL
plot(bicopsel, type = "surface",zlim=c(0,5),main="Gumbel copula  (High TM)",xlab
    ="-omega",ylab="prec.")

#Simulated copula data: GUMBEL
```

```
simdata <- BiCopSim(nrow(wprec_humalta), bicopsel)


plot(simdata[,1],simdata[,2], xlab="-omega", ylab="precipitation",main="
   Simulated data: Gumbel copula  (High TM)")



#########################
```