Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Paulo Vitor da Costa Pereira

Modelando precipitação extrema no Brasil pela teoria dos valores extremos

Modeling daily rainfall in Brazil with extreme value theory

Maringá
2016

Paulo Vitor da Costa Pereira

Modelando precipitação extrema no Brasil pela teoria dos valores extremos

Dissertação apresentada à Universidade Estadual de Maringá, como parte das exigências do Programa de Pós-Graduação em Bioestatística, área de concentração em Estatística Aplicada, para a obtenção do título de Mestre em Bioestatística.

Orientadora: Isolde T. S. Previdelli
Coorientador: Anthony C. Davison[1]

26 de Agosto de 2016

[1] Chair of Statistics, École polytechnique fédérale de Lausanne

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Central - UEM, Maringá, PR, Brasil)

Pereira, Paulo Vitor da Costa
P436m Modelando precipitação extrema no Brasil pela teoria dos valores extremos / Paulo Vitor da Costa Pereira. – - Maringá, 2016.
69 f. : il., figs., gráf.

Orientadora: Prof.ª Dr.ª Isolde Terezinha Santos Previdelli.
Co-orientador: Prof. Dr. Anthony Christopher Davison.
Dissertação (mestrado)- Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2016.

1. Teoria estatística bayesiana. 2. Decisões estatísticas. 3. Modelo hierárquico bayesiana. 4. Precipitação pluviométrica – Análise. 5. Análise estatística. I. Previdelli, Isolde Terezinha Santos, orient. II. Davison, Anthony Christopher, co-orient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 22. ed.519.542
MGC - 001736

Acknowledgements

Agradeço principalmente à Isolde Previdelli pela orientação e companheirismo tanto dentro da universidade quanto fora. Agradeço também pela oportunidade que me foi dada em passar 6 meses fora do país, na cidade de Lausanne, na Suíça, com o professor Anthony Davison da École polytechnique fédérale de Lausanne. Sou extremamente grato ao Anthony por financiar minha estadia no exterior, e pela sua paciência e gentileza em ler e contribuir com este trabalho. A Isolde e o Anthony são fontes de inspiração para a minha vida pessoal e profissional, a experiência fora do país foi um divisor de águas nesses dois apectos da minha vida.

Agradeço aos professores e alunos do programa de pós-graduação em Bioestatística da Universidade Estadual de Maringá pela convivência agradável e construtiva. E também aos alunos com os quais tive contato na École polytechnique fédérale de Lausanne.

Agradeço ao suporte material ou financeiro da Universidade Estadual de Maringá, da École polytechnique fédérale de Lausanne, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Abstract

The accurate modeling of extreme events is growing in relevance, particularly in the environmental sciences in which such events can be seen as a result of climate change. In particular, measuring rainfall risk is also important for the design of hydraulic structures (dams, levees, drainage systems, bridges, etc.) and for flood mapping and zoning. The Brazilian regulatory agency, Agência Nacional de Águas (ANA), makes available rainfall series for 11,368 rain stations throughout Brazil, some of them dating from the 19th century. One of our goals was to produce, using the framework of extreme value theory, maps with reliable estimates of the 25-year return level of a extreme rainfall for each locality covered by ANA. Such dataset present many complex challenges: first, evaluating its quality; then, modeling spatial extremes over large random fields; modeling temporal nonstationarity of the extreme rainfall process due to natural climate seasonality and due to a possible trend owing to climate change; correcting biases resulting from misspecification of the model or from a small sample. In this study, we tackle all these issues. We perform a detailed quality control, and we make a deep discussion of biases resulting either from misspecification of the model or from a small sample, while providing important information regarding the modeling of rainfall extremes, and complementing recent previous studies. In particular, the shape parameter of the extreme-value model seems to have a mean asymptotic value of 0.06.

Key-words: Bayesian hierarchical model; penultimate bias; precipitation field; return level map; small sample bias.

Contents

Chapter 1

Introduction

In many regions of the world, changes in rainfall patterns are affecting quantity and quality of water resources. According to the "Summary for Policymakers" for the year of 2014 from the Intergovernmental Panel on Climate Change, "many human systems and some ecosystems reveal significant vulnerability and exposure to current climate variability." The amount of certainty about this statement was classified with "very high confidence."

In Brazil, where the main natural hazard comes from the lack or excess of rain (cyclones and earthquakes, for example, are rare and of relative low magnitude), when catastrophes happen, it is easy to be critical *ex post facto* of the absence of prevention. However, such catastrophic events are often seen as so unpredictable or implausible that even to the eyes of public managers they can be neglected. And prevention measures usually take place when it is too late. A good example is the city of São Paulo, which suffered severe floods and rainfall in 2010, and, shortly after, endured intense droughts that depleted its main reservoir system. So, not only we usually have optimism bias and underestimate true risks, but we also hare difficulty in making long-term policies and plans.

According to Parmesan, Root e Willig [2000], changes in extremes of temperature are more responsible for changes in the nature than changes in mean temperature. Similarly, in finance, "financial solubility of an investment is likely to be determined by extreme changes in market conditions rather than typical changes" [COLES, 2001, p. 11]. As well put by Davison e Huser [2015], "in an evolving climate, changes in the sizes and frequencies of rare events, rather than changes in the averages, may be what lead to the most devastating losses of life and the greatest damage to infrastructure."

According to Parmesan, Root e Willig [2000], changes in extremes of temperature are more responsible for changes in the nature than changes in mean temperature. Similarly, in finance, "financial solubility of an investment is likely to be determined by extreme changes in market conditions rather than typical changes" [COLES, 2001, p. 11]. As well put by Davison

e Huser [2015], "in an evolving climate, changes in the sizes and frequencies of rare events, rather than changes in the averages, may be what lead to the most devastating losses of life and the greatest damage to infrastructure."

Extreme events are characterized as being of low frequency and having large time periods. In practice, very often it is required to estimate probabilities of events that have never yet been observed. In the absence of empirical or logical arguments to formulate an extrapolation rule, we are left with asymptotic arguments. This is the basis of extreme value theory, which provides techniques to estimate future extreme levels from a data-generating process [COLES, 2001].

However, due to the complex stochastic structure of the environment and of financial markets, naïve application of extreme value theory can provide illusory results and lead to a false sense of security. For example, in Venezuela, according to Coles e Pericchi [2003], "prior to 1999, simple extreme value techniques were used to assess likely future levels of extreme rainfall, and these gave no particular cause for concern. In December 1999, a daily precipitation event of more than 40 cm, almost three times the magnitude of the previously recorded maximum, caused devastation and an estimated 30,000 deaths." A more recent study of this catastrophic rainfall event was made by Süveges e Davison [2012].

In South America, floods and landslides are very frequent during the summer and occur after heavy and continuous rainfall. In Brazil, the most prominent and recent case occurred in the mountainous region of Rio de Janeiro State, between 11 and 12 January 2011, and led to 947 deaths. This is considered to be the worst natural disaster in Brazil's history. The accumulated rainfall in 24 hours was 241.8 mm, with a peak 61.8 mm in an hour [DOURADO; ARRAES; SILVA, 2012]. On the other hand, extreme droughts have severely affected eastern Brazil since 2012, damaging the country's agricultural and electrical production. According to Getirana [2016], this extreme drought is mostly related to lower-than-usual precipitation rates, and its impacts have been exacerbated by ineffective energy development and water management policies. At the same time, record-breaking rainfall and floods happened in Amazonia during the austral summer and fall of 2012 [MARENGO et al., 2013].

Of course the characterization of an extreme event depends on where it takes place. For example, in March 2015, the city of Antofagasta in the Atacama Desert (northern Chile), which usually receives about 1 to 3 mm of rain in a year, registered 24 mm of rain in just one day [LIBERTO, 2015]. The Atacama Desert is probably the driest and oldest desert in the world. Some weather stations there have never received rain. Although 24 mm seems small, because of the rock-hard ground, which does not absorb water, and the lack of vegetation, which leads to rapid erosion, dry river beds become rushing torrents of water capable of great destruction.

Extreme quantile estimation of rainfall is interesting for flood mapping and zoning, but also for the design of hydraulic structures (dams, levees, drainage systems, bridges, etc.)

since hydrological risk is highly dependent on rainfall risk [MULLER et al., 2009]. So, our main objective was to model extreme rainfall in Brazil and provide reliable estimates of extreme quantiles. There are two sources of error that we need to control: the error inherently present when using asymptotic models, and the bias induced by small samples. We discuss these errors, and complement the studies of Papalexiou e Koutsoyiannis [2013] and Serinaldi e Kilsby [2014]. We also investigated the relation between extreme rainfall and events such as El Niño and La Niña, as well as other possible covariates like deforestation and carbon dioxide levels.

The idea of including deforestation as a covariate came from the recent extreme drought events in eastern Brazil. The low precipitation in this region "seems unconnected to ocean temperatures or other large-scale weather phenomena" and is probably related to climate change [ESCOBAR, 2015]. Another probable cause is deforestation in Amazonia, since a large portion of the moisture produced in the Amazon basin is exported (through low-level jets, also called aerial rivers[1][ARRAUT et al., 2012]) to distant basins in southeast South America [NAZARENO; LAURANCE, 2015]. Getirana [2016] recommends future studies on these possible causes, considering, for example, "the simultaneous drought over eastern Brazil and floods over the Amazon."

In the next chapter, we describe the dataset and present the main concepts and methods used, illustrated with one rain station located in the city of Pomerode, in the state of Santa Catarina. In Chapter 3, we present the main results. All analysis were done in the statistical software R [R Core Team, 2016]. To produce the return level maps, we used the **SpatialExtremes** package.

[1] Aerial rivers are an analogy to surface rivers as a pathway of moisture flow in the atmosphere.

Chapter 2

Dataset

The dataset used is publicly available on the Internet at the Hydrological Information System (HidroWeb), administered by Agência Nacional de Águas (ANA), a Brazilian regulatory agency. In its inventory, ANA has descriptive information about 20,164 rain stations spread throughout South America, but mainly contained within Brazil's borders. From these stations, 11,619 (57.62%) have daily rainfall values registered in the hydrological information system.

Each daily value was classified as being "blank" (missing value), "real" (accumulated rainfall in 24 hours), "estimated", "doubtful", or "accumulated" (when the observer does not make a measurement, rainfall accumulates until the measurement is done); the average proportion of these possible values were 7.72%, 92.06%, 0.03%, 0.07%, and 0.12%, respectively. The doubtful and accumulated values were removed. This reduced the number of stations to 11,368.

2.1 Selection of stations

Serinaldi e Kilsby [2014] selected from their database only time series covering the same periods. More specifically, they worked with two subsets of their original data: rainfall series spanning over the period 1970–2011, and series over 1900–2011, resulting in 1898 and 113 stations for each subset. As acknowledged by them, this is a very restrictive criterion, but they argue that, in this way, the series reflect the climate conditions over homogeneous time windows. Figure 1 shows the timeline of each station from ANA. For short time series, say a 40-year period, the homogeneous time windows criterion seems reasonable for some interval after 1960.

Like Serinaldi e Kilsby [2014], we are going to retain two subsets: stations with rainfall series spanning over the period 1972–2011, and stations with record length greater than 80 years. The distribution of record length, in years, for all stations is described in Table 1. This

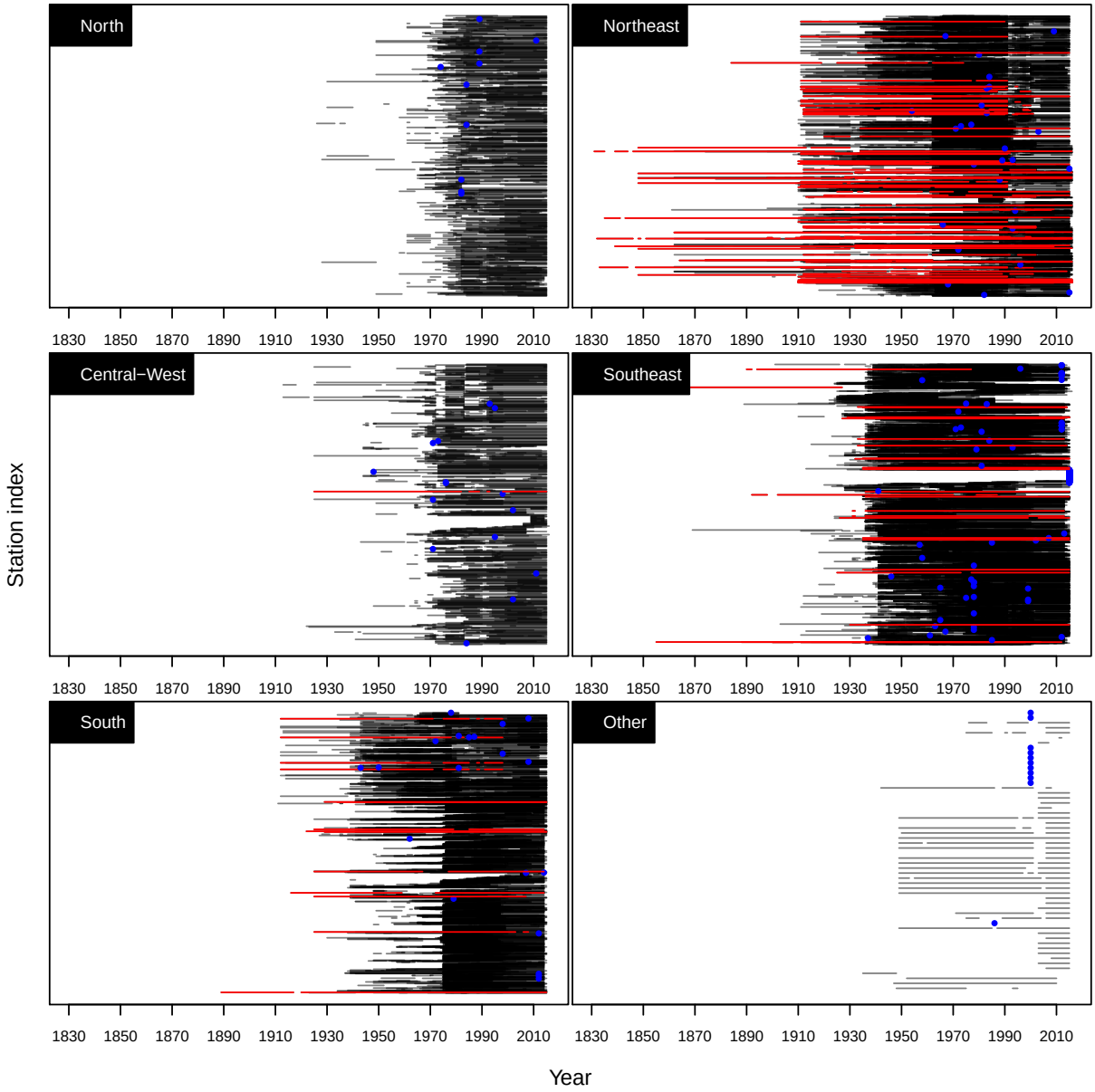

Figure 1 – Timelines of all stations. The first five plots corresponds to the five macroregions of Brazil, and the last plot, the region outside Brazil. Timelines longer than 80 years are shown in red, and timelines with a record smaller or equal to one year are shown as blue dots.

resulted in 1216 and 184 stations for the two subsets. The interval from 1972 to 2011 was chosen to maximize the number of observations in a 40-year span. Only the stations with less than 10% of missing values in this period were selected. The time windows for the 184 stations of the second subset are somewhat heterogeneous; see Figure 2. In order to obtain more homogeneous time windows, we kept the stations with record length greater than 80 years starting from 1909, reducing the number of stations to 164.

Table 1 – Distribution of the record lengths in years.

Interval	Number of stations	Cumulative frequency	Cumulative relative frequency (%)
149	1	1	0.01
$[110, 140)$	2	3	0.03
$[100, 110)$	33	36	0.32
$[90, 100)$	23	59	0.52
$[80, 90)$	125	184	1.62
$[70, 80)$	566	750	6.60
$[60, 70)$	493	1,243	10.93
$[50, 60)$	694	1,937	17.04
$[40, 50)$	1,416	3,353	29.50
$[30, 40)$	2,058	5,411	47.60
$[20, 30)$	1,797	7,208	63.41
$[10, 20)$	2,269	9,477	83.37
$[5, 10)$	998	10,475	92.14
$[1, 5)$	893	11,368	100.00

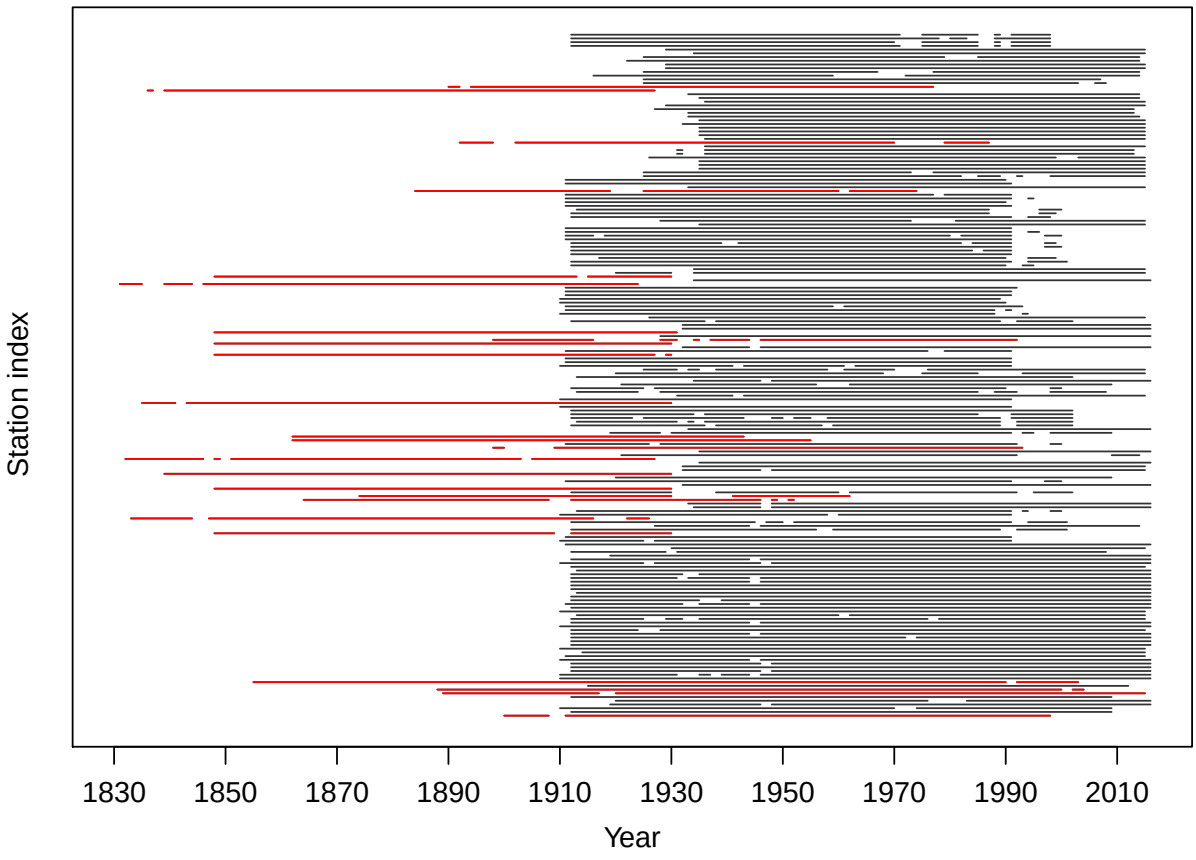

Figure 2 – Timelines of the 184 stations with record length over 80 years. Timelines beginning before 1900 are shown in red.

2.2 Quality control and homogenization

Before we attempt to do any kind of analysis, it is important that we check the quality of the data and trust the measurements made. Since 70 stations in the subset with the longest rainfall series are also in the subset containing the shortest series, we looked individually at 1310 unique stations (1216 + 164 - 70).

A rainfall series with no apparent causes of concern, may be useless if it stems from inappropriate measurement conditions. Metadata like the stations' history and photographs showing their location and measurement conditions is crucial to assess the measuring quality. As stated in the guidelines by the World Meteorological Organization on "climate observation networks and systems," an observation site should be representative of the climatic regime for which it is intended. Otherwise, the site becomes representative of local features only [PLUMMER et al., 2003]. The amount of rainfall measured is very sensitive to systematic wind-field deformation. Neither completely open exposure nor big objects close to the gauge[1]is desirable. The best sites for measuring precipitation are often found in places where objects act as an effective wind-break from all directions, for example, clearings within forests or orchards, among trees, in scrub or shrub forests. When there are no natural wind-breakers near the site, windshields are used. Figure 3 shows a typical station from ANA. Unfortunately, not all their stations have a corresponding photograph in the database, and we could not find photos for stations belonging to other entities. Moreover, a photo from just one direction is not enough, the photos have to show all possible obstacles around the rain gauge, and the land slope; see Jarraud [2008, p. 38].

For the rain stations run by ANA, daily measures are always made at seven in the morning. Stations from other entities may collect rainfall on smaller time intervals (three times a day, hourly or on intervals of 15 minutes). In these cases, the original values are transformed to accumulated daily values. None of the stations have measured negative rainfall values, but 15 stations have registered values greater than 500 mm. It is difficult to distinguish if these registered values were really observed or if they are outliers. Figure 4 exemplifies the problems we found when visualizing the 1,310 rainfall time series one by one:

- The first image shows a possible outlier. The station is in the Amazon basin, it has a measured daily rainfall of exactly 999.7 mm, the second largest value being 186.2 mm. In all regions, it is not uncommon to find stations displaying daily precipitation amounts greater than 300 mm or 400 mm. The volume capacity of some measuring devices might not be large enough. Standard German Hellmann rain gauges, for example, can collect only up to 200 mm, so they are not adequate for these regions;

- The following two plots (corresponding to stations 33 and 160) illustrate a very common

[1] It is recommended that objects do not be closer to the gauge than a distance twice their height above the gauge orifice.

Figure 3 – A typical station run by ANA.

problem: truncation of high precipitation events. It may happen that only events bellow a certain value are registered or a frequency peak at specific values are observed. These frequency peaks are also observed as jumps in the empirical distribution function (outliers may difficult visualizing truncation, so these must be dealt with first). Truncation happens when the observer does not properly understand how the measurement must be done, and it is not always easy to detect (it can appear in many different ways on the plot). Even though some of these stations seem to be registering extreme values, we decided to exclude time series showing even the slightest sign of truncation;

- The next plot (station 59) show a similar problem to truncation, high precipitation events seem to be hidden in a section of the time series, but no frequency peaks appears;
- The opposite behavior also happens (station 104), instead of seemingly unusual small values, there are sections of the time series with suspiciously high values;
- Low precipitation gaps are another frequent issue (stations 384). These gaps appear due to negligence of low precipitation measurements. With time, the scale and the water marks of the device become weathered and faded, especially the marks for low values, making the reading impossible, so the observer might erroneously interpret these unrecognizable marks as zero precipitation. But the most frequent source of this error comes from irregularly measures made by the observer, i.e., measures are made only after "substantial" rain events, and very low amounts are ignored;

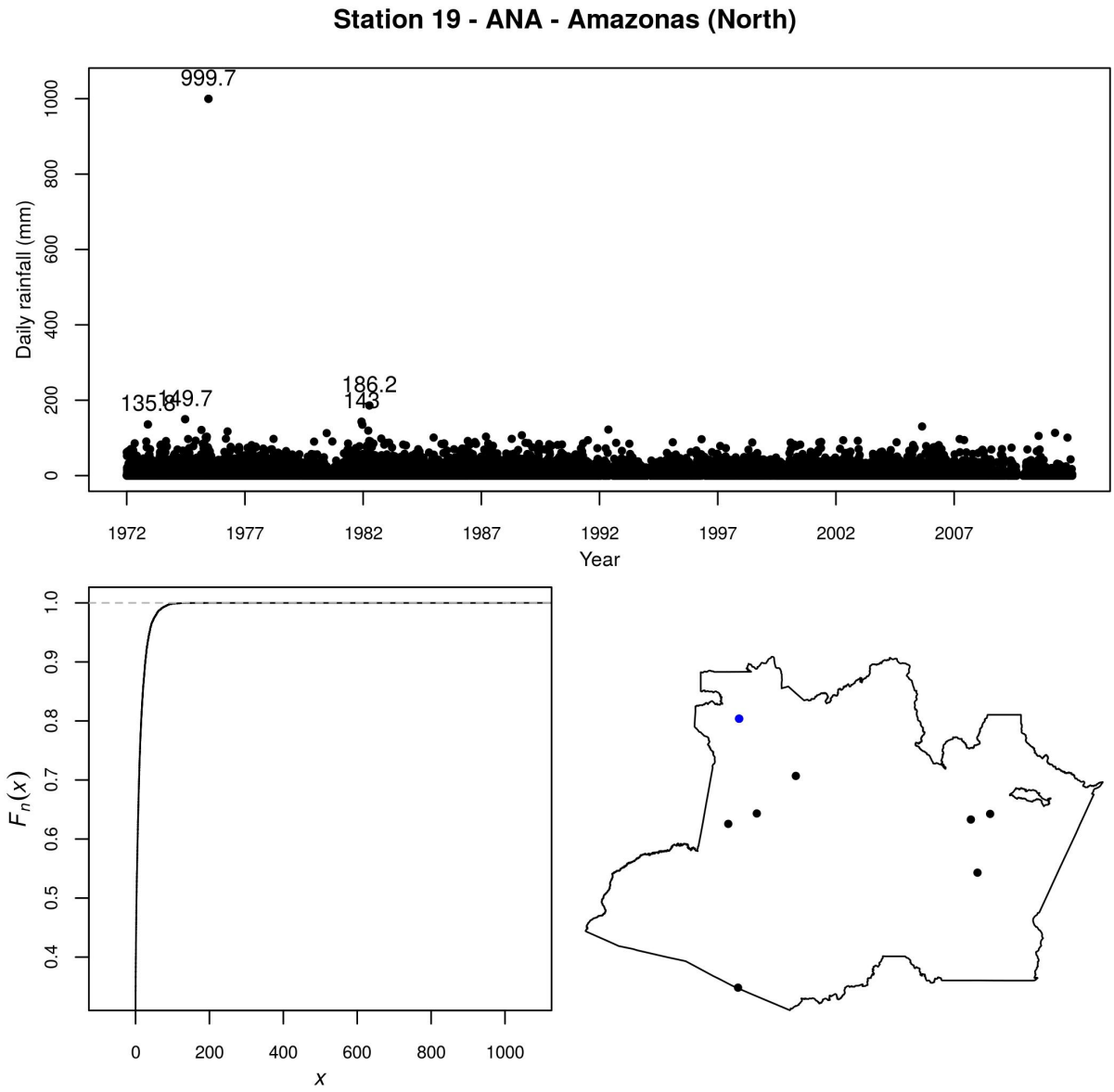

Figure 4 – Visual quality control applied for each of the 1310 selected rainfall time series. Top plot: daily rainfall values. Bottom line: empirical distribution function for the whole rainfall time series starting at the point mass zero (left), and map of the corresponding state with the displaying station in blue.

- Some stations had a large section of just zeros. Although this is probably just a classification problem, it is something that needs to be checked.

From the 1310 stations, we removed outliers from 60 stations and discarded 200 stations that had any of the other problems listed above.

When the observer does not make a measurement, rainfall accumulates inside the measuring device. So, if the observer misses one or more days, he must measure the accumulated rainfall and tag this measurement as accumulated. However, because this can happen due to negligence of the observer, accumulated values are often untagged [VINEY; BATES, 2004].

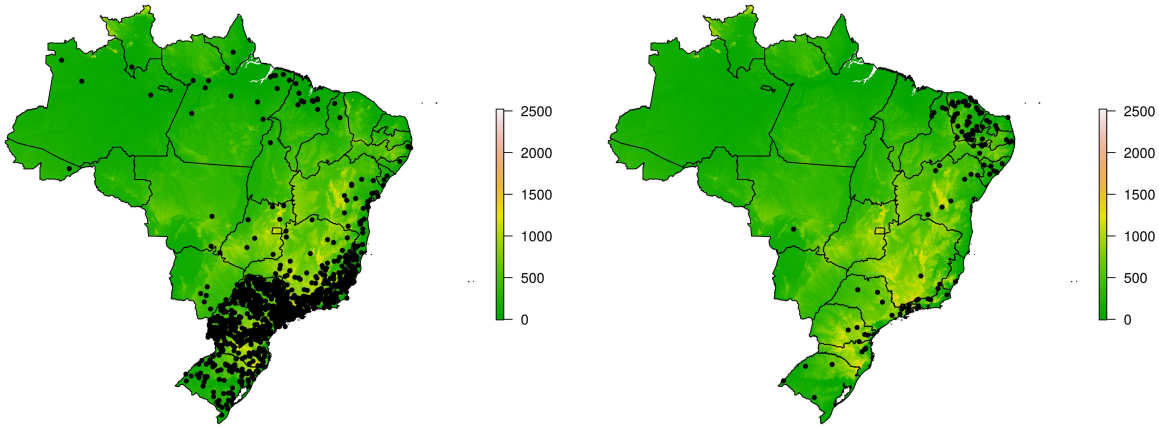

Figure 5 – Map of Brazil, with an insert showing the altitude, showing the stations in the subset with the shortest and longest rainfall series (left and right panels), selected from the database available in the Hydrological Information System (HidroWeb), administered by ANA.

And since daily routines differ on weekends from weekdays, the observer may fail to accomplish his task more frequently on weekends. In order to detect stations with untagged rainfall accumulation, we counted the number of dry and wet days for each day of the week, and performed a Pearson's chi-squared test for the homogeneity of the proportions. The null hypothesis was rejected for about 10% of the stations. Air pollution related to human activity might cause weekly cycles in rainfall time series, but its significance is dubious. Since accumulations of two or more days may significantly alter daily extremes, we preferred to also exclude these stations, remaining 893 and 104 stations for the subsets with the shortest and longest rainfall series. Figure 5 shows the location of these stations.

All stations have an observer making the measurements, even the automatic stations (so inconsistencies can be compared). The number of automatic stations in the "cleaned" dataset is 145 (about 15%). The institutions responsible for the "cleaned" stations can be seen in Table 2, almost all of them are government agencies or public companies. ANA classified the instruments used for measuring rainfall in three categories: pluviometer, rain gauge, and data logger. From the 997 stations, 990 used exclusively the pluviometer, the other 7 stations used the other instruments briefly. The daily rainfall values were also classified into "raw" and "consisted" (or "validated"). Consisted values, about 26% of the total, went through some kind of analysis to identify and correct erroneous values as well as to input missing data. Every year, agents of the Brazilian electrical sector are obliged to send reports to ANA on the consistency of the data collected the previous year and the data itself [FREITAS; NóBREGA, 2012].

Climate time series often contains artificial changepoints, i.e., shifts in the mean due to inevitable changes of the instrument used, changes in the instrument itself, changes of observers or procedures. The process of adjusting the time series in order to remove or diminish the effect of artificial shifts is called homogenization. Wang et al. [2010] describe a specific

Table 2 – Institutions responsible for the selected rain stations.

Institution	Number of stations	Proportion (%)
ANA[1]	464	46.77
DAEE-SP[2]	288	29.03
AGUASPARANÁ[3]	114	11.49
DNOCS[4]	26	2.62
CEEE[5]	25	2.52
FUNCEME[6]	24	2.42
INMET[7]	23	2.32
COPEL[8]	11	1.11
EMPARN[9]	6	0.60
Light[10]	4	0.40
Duke Energy[11]	3	0.30
CONS.CECS[12]	1	0.10
Itaipu dam[13]	1	0.10
SEMARH-AL[14]	1	0.10
Tractebel Energia[15]	1	0.10

[1] Agência Nacional de Águas, a federal regulatory agency.
[2] Departamento de Águas e Energia Elétrica, a state agency of São Paulo.
[3] Instituto das Águas do Paraná, a state agency of Paraná.
[4] Departamento Nacional de Obras Contras as Secas, a federal institution that acts in the semi-arid region of Brazil.
[5] Companhia Estadual de Geração e Transmissão de Energia Elétrica, a government-controlled company in the state of Rio Grande do Sul.
[6] Fundação Cearense de Meteorologia e Recursos Hídricos, a state agency of Ceará.
[7] Instituto Nacional de Meteorologia, a federal agency.
[8] Companhia Paranaense de Energia, a government-controlled company in the state of Paraná.
[9] Empresa de Pesquisa Agropecuária do Rio Grande do Norte, a public company formed by the state of Rio Grande do Norte and the federal government.
[10] A private company in the state of Rio de Janeiro.
[11] A private company acting in the state of São Paulo.
[12] Consórcio Energético Cruzeiro do Sul, a consortium between two public companies in the state of Paraná.
[13] A binational hydroelectric dam on the Paraná river.
[14] Secretaria do Meio Ambiente e dos Recursos Hídricos do Estado de Alagoas, a state agency of Alagoas.
[15] A private company, the correspondent station is in the state of Paraná.

procedure for homogenizing (nonzero) daily precipitation series. When changepoints are documented, this is just a matter of estimating the shift sizes. However, the stations' history are often absent or incomplete, which is the case for the ANA stations, and the changepoints themselves have to be estimated. WANG et al. claim that, by using their method, detected and documented changepoints are in agreement 70% of the time. Using the software they provide, we learned that the bulk of the data often appears to have some changepoints, but when looking only at observations above some high quantile, the few changepoints detected seem to be always false positives, from which we conclude that the possible artificial changepoints do not influence significantly on extreme values or even values near the median; see Figure 6.

Changes of observers, instrument replacements, or new observer instructions, could

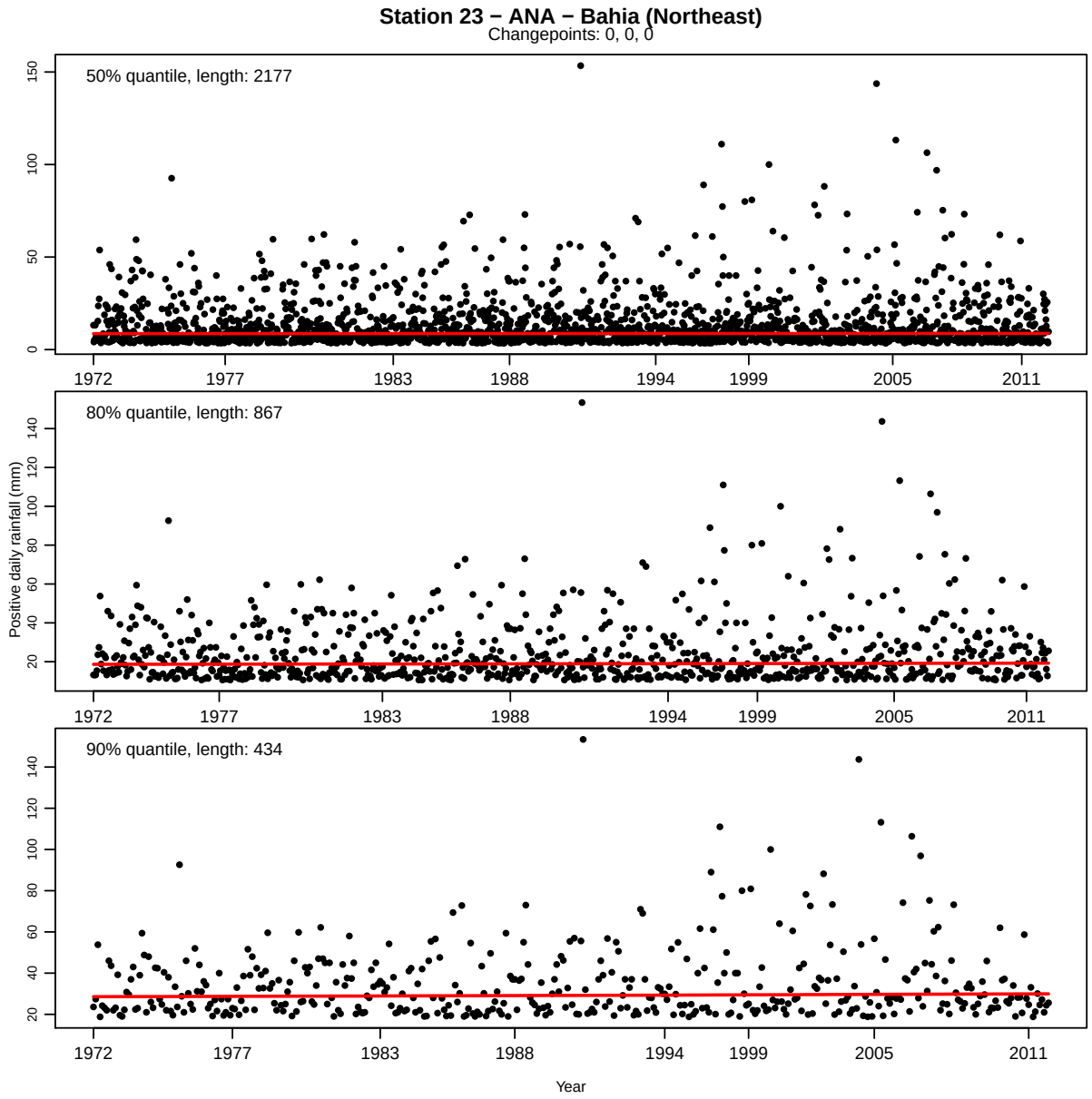

Figure 6 – Daily rainfall series above the median, the 80% quantile, and the 90% quantile (first, second, and third lines). The red line depicts possible mean shifts.

also cause the precision of measurements to change in time, although this is more apparent for temperature than rainfall series. In the ANA database, the precision is always of 0.1 mm for all stations and years.

2.3 El Niño Southern Oscillation

The El Niño Southern Oscillation (ENSO) is an irregularly periodical coupled ocean-atmosphere phenomenon that causes global climate variability. The warm and cold phases of the ENSO are known as El Niño and La Niña. The warm phase corresponds to warmer-than-average sea surface temperatures of the tropical eastern Pacific Ocean, accompanied by high air pressure in the western Pacific and low air pressure in the eastern Pacific, and conversely. Each phase typically lasts from three to four years.

Figure 7 summarizes the effects of the ENSO on weather in South America. Particularly in Brazil, El Niño causes wetter-than-normal conditions in the south mainly during the spring and early summer [GRIMM, 2003]. Drier and hotter weather occurs in north Amazon and in the northeast region [MARENGO, 1992; HASTENRATH; GREISCHAR, 1993]. During La Niña, the main effects are increased rainfall in the north-eastern region from December to February [MARENGO, 1992], and severe droughts in the south [GRIMM; BARROS; DOYLE, 2000]. Thus, it is natural to investigate the impact of ENSO events on the extremal behavior of the rainfall series.

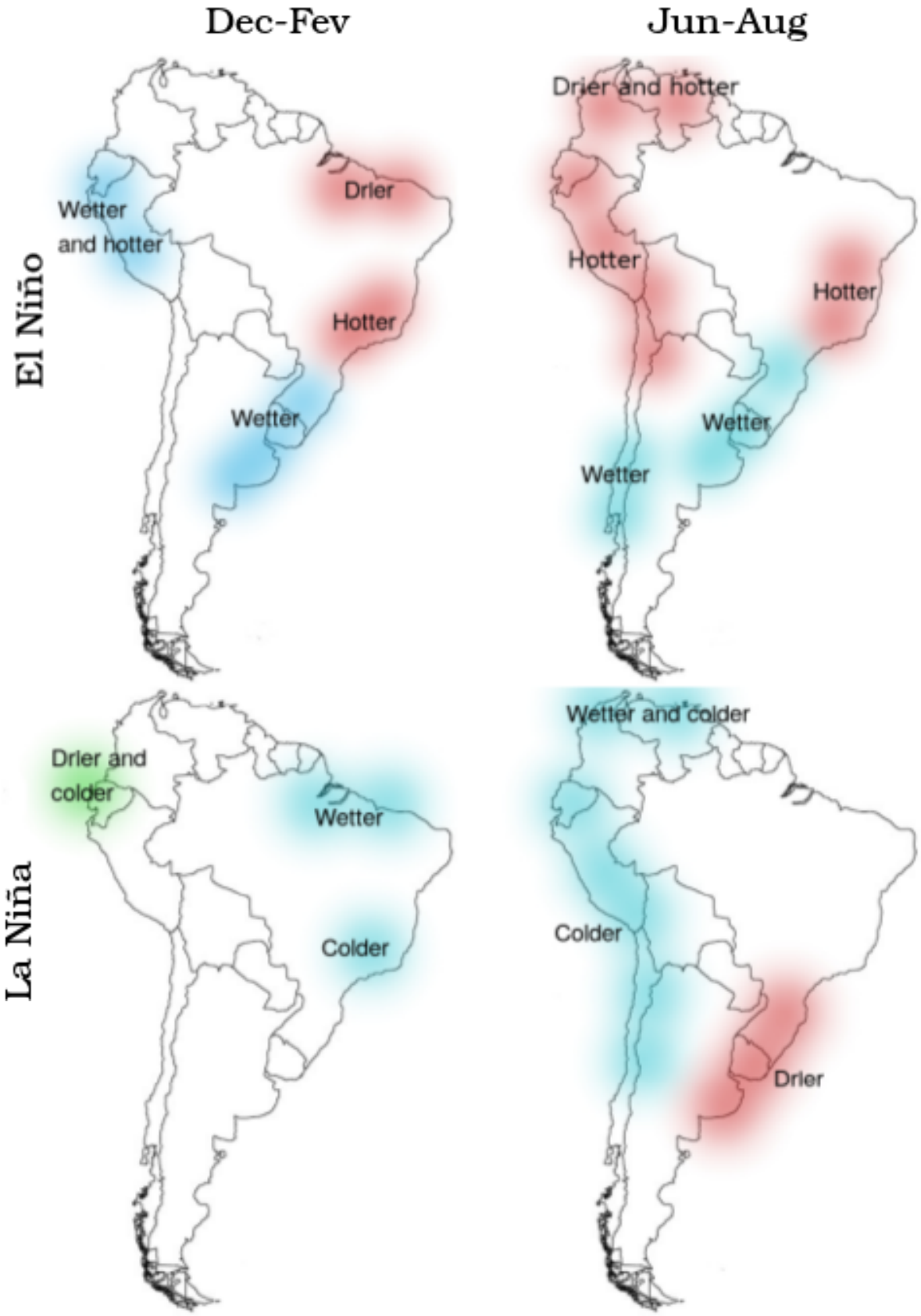

Figure 7 – Rough location of the ENSO effects on temperature and rainfall (when compared to normal conditions) in South America. Top row: El Niño. Bottom row: La Niña. Left column: months of December to February. Right column: June to August.

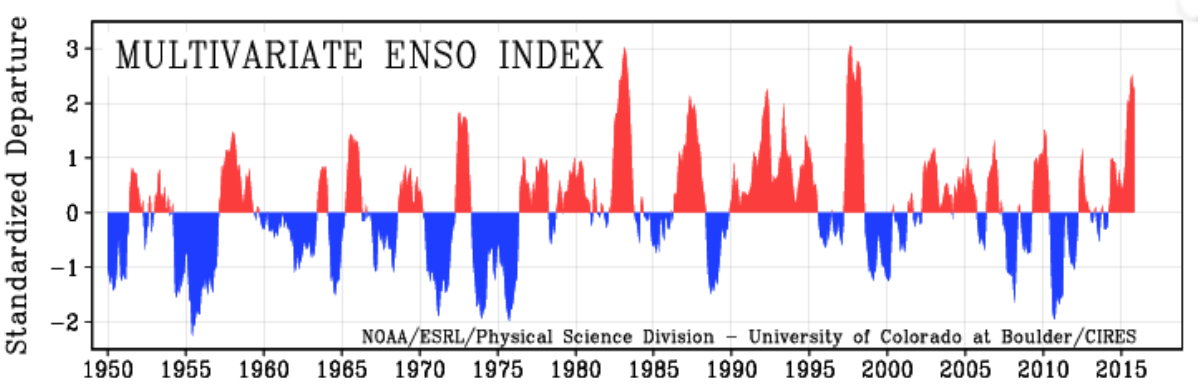

Figure 8 – The multivariate ENSO index (MEI). It is a scalar measure of the intensity of the ENSO phenomenon. The "multivariate" term comes from the way the index is calculated: as the first principal component of six meteorological and oceanographic variables. Source: National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Physical Sciences Division.

The multivariate ENSO index (MEI) is a monthly measure of the intensity of this phenomenon (see Figure 8). It is the first principal component of six meteorological and oceanographic variables: sea level pressure, zonal and meridional components of the surface wind, sea surface temperature, surface air temperature and cloudiness. Negative and positive values of the MEI represent the cold and warm ENSO phases, i.e., La Niña and El Niño [WOLTER, 2000]. Therefore we expect to observe a negative association between MEI and extreme rainfall in north-eastern Brazil, and a positive association in the southern region.

2.4 Deforestation

Since 1988, the National Institute for Space Research (INPE, in Portuguese) through the Amazon Deforestation Monitoring Project (PRODES, in Portuguese) has annually measured the deforestation rate for clearcutting of forested areas (see Figure 10) in Legal Amazon, using Landsat satellite imagery. The term "Legal Amazon" was created by the Brazilian government to designate an area of similar economical, political and social problems. This area covers 59% of Brazilian territory and is home to all the Amazon rainforest biome within Brazil, 37% of the Cerrado biome and 40% of the Pantanal biome. Figure 9 illustrates the size of the area covered by PRODES as well as the region covered by the Landsat images.

The annual deforestation rates are available by state and for the whole Legal Amazon in the PRODES website. The annual and accumulated deforestation rates in Legal Amazon are shown in Figure 11. As an example, in 2014, the total deforested area estimated by PRODES from August 2013 to July 2014 was 5012 km^2.

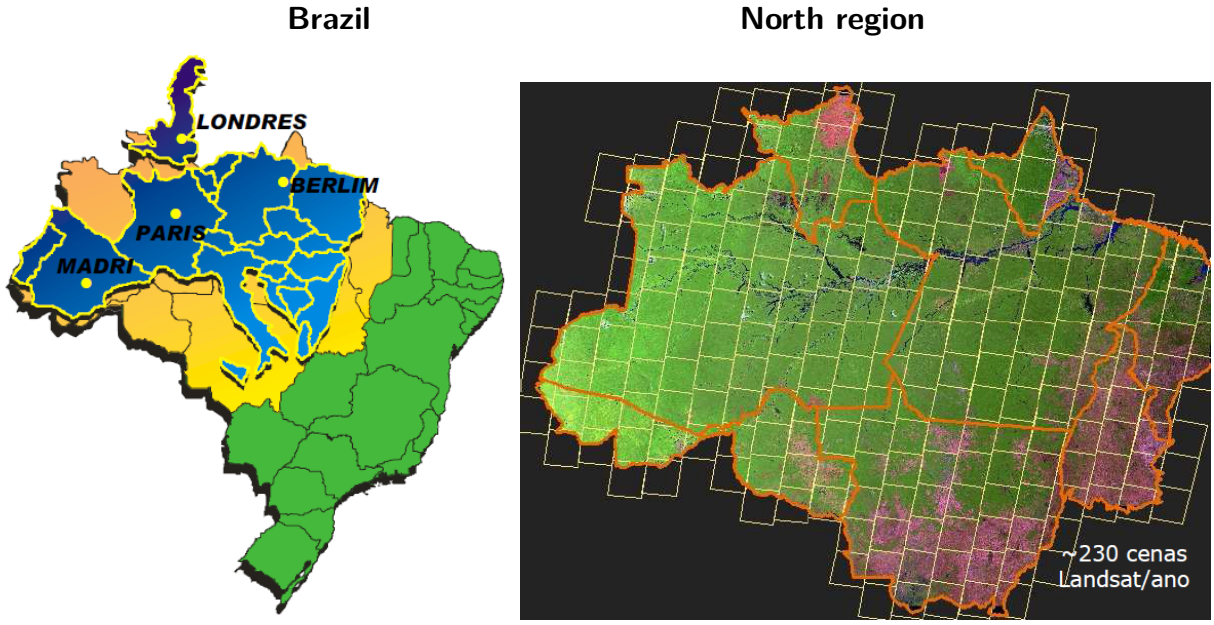

Figure 9 – The left and right panels illustrate the size of the North region (in yellow) of Brazil (in yellow and green) in comparison to a portion of Europe (in blue) and the region covered by the Landsat images (about 230 images per year). Source: National Institute for Space Research.

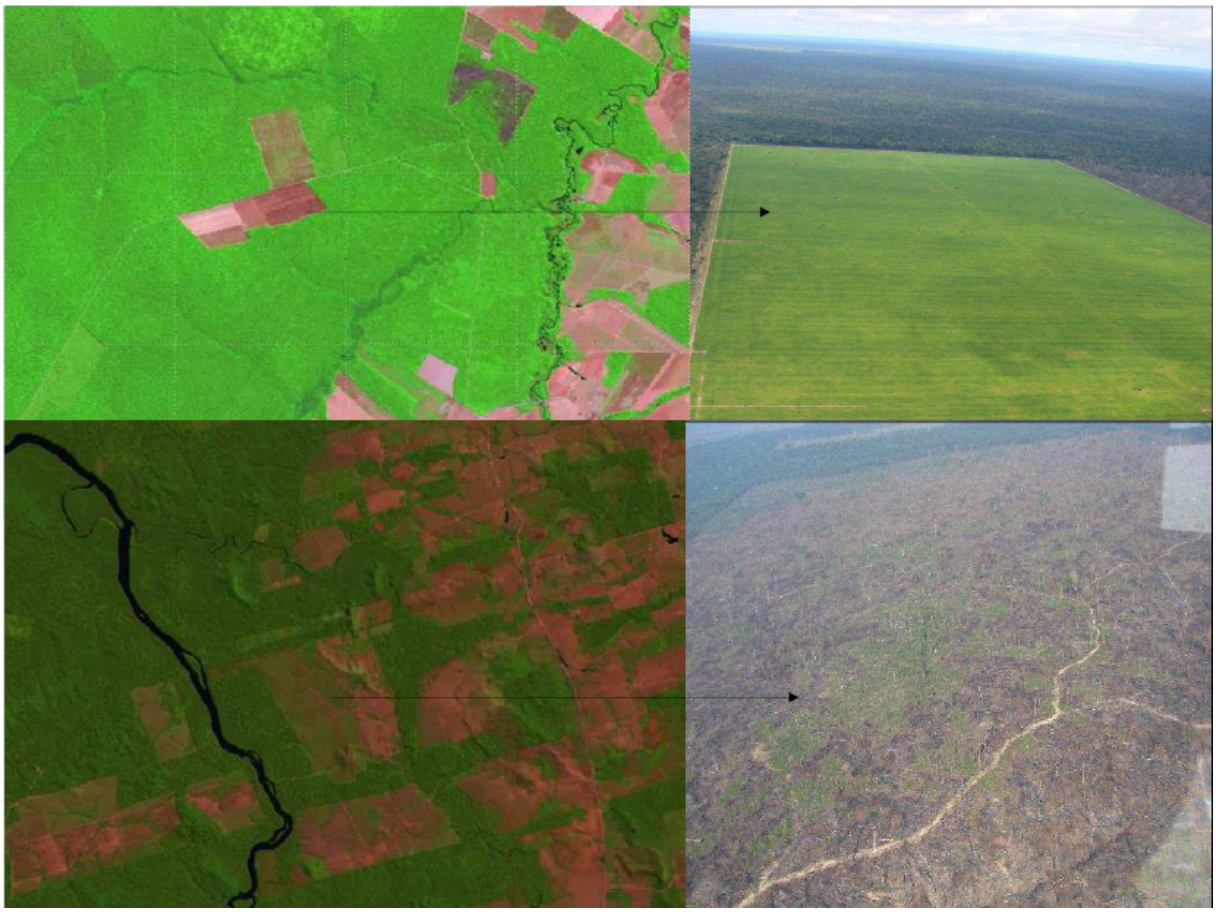

Figure 10 – Illustration of clear cut deforestation, i.e., total removal of forest cover in a short period of time. PRODES identifies clear cut areas larger than 6.25 hectares. Source: INPE's website.

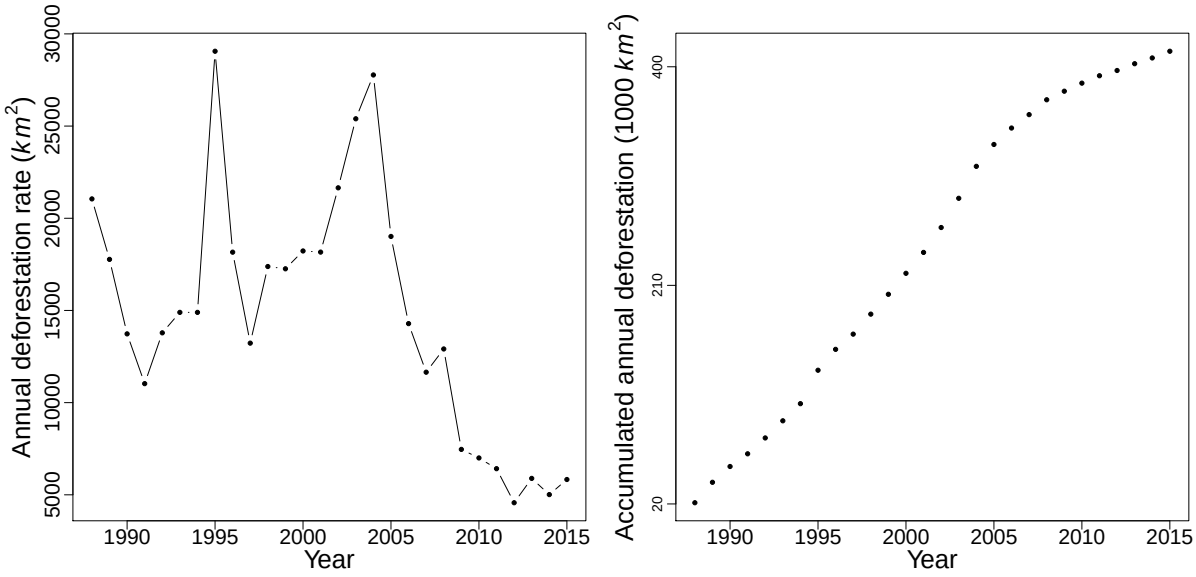

Figure 11 – Annual and accumulated deforestation rates in Legal Amazon assessed by PRODES.

Chapter 3

Methodology

Extreme value theory deals with the asymptotic distributional behavior of extremes, which can be defined as being the block maxima, the r largest order statistics, or exceedances over a high threshold. Block maxima are the maximum values extracted from blocks of observations, and the r largest order statistics are the r largest values within these blocks, whereas the exceedances refer to observations that exceed a given threshold.

In applications, the process is usually measured at regular intervals. For example, Figure 12 shows the (accumulated) daily rainfall registered in the city of Pomerode, Santa Catarina State. In this case, the blocks could correspond to the years, so the size of each block is the number of observations per year. Since the data go from 1929 to 2015, there are 87 years annual maxima. Using as a threshold the 98% quantile for all rainfall values, there are 624 exceedances, and for values greater than zero (wet days only), there are 294 exceedances. In either case, the samples are much larger than using the block maxima procedure. The exceedances approach is an attempt to include as much extremes as possible in the analysis. If a complete time series of observations is available, then better use of the data is made by adopting a threshold to characterize extreme values, avoiding blocking [COLES, 2001, p. 9, 74]. However, the choice of threshold involves a similar bias–variance trade-off when choosing block length. Extreme value theory provides a link between these two types of extremes and their underlying distributions.

3.1 Two types of extremes and their distributions

Consider a sequence of independent and identically distributed random variables, $X_1, \ldots, X_n$, with common distribution function F (referred to as the parent distribution). The distribution of the block maxima, $Z_n = \max\{X_1, \ldots, X_n\}$, is simply $G_n(z) = F^n(z)$, so one way to estimate the distribution of Z_n is to first estimate F, and then use F^n. Unfortunately, very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n [COLES,

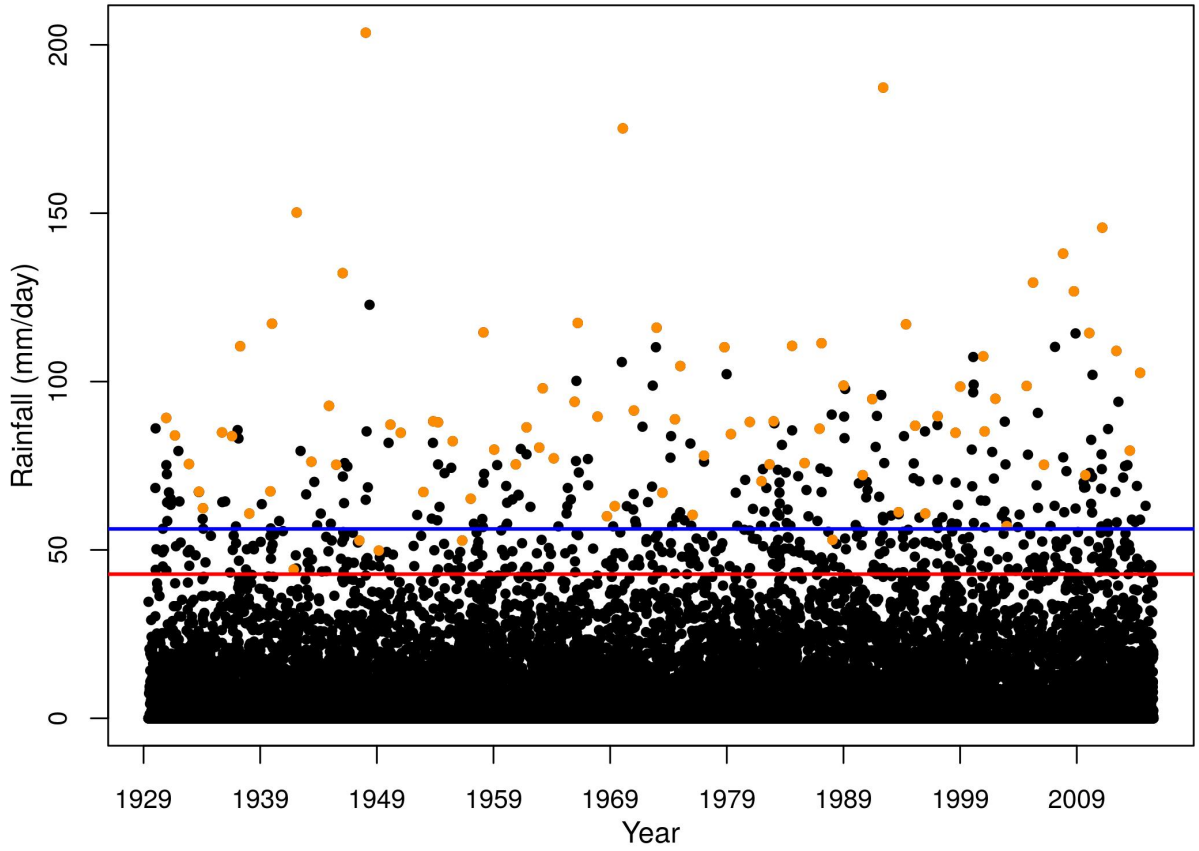

Figure 12 – Daily rainfall registered in the city of Pomerode, in the state of Santa Catarina. The red and blue lines represent the 98% quantile for all rainfall values and for values greater than zero (wet days only). The orange points are the annual maxima series.

2001, p. 46].

If n is not constant (for example, if we take only wet days), but rather can be regarded as a realization of a Poisson distributed random variable, N, with mean ν, then the distribution of Z_N is

$$G'_\nu(z) = \exp\{-\nu(1 - F(z))\}. \tag{3.1}$$

Since $\log u \doteq u - 1$ when $u \to 1$,

$$\log G_n(z) = n \log F(z) = n \log[1 - \{1 - F(z)\}] \doteq -n\{1 - F(z)\} = \log G'_n(z).$$

Even for small n, the difference between $G_n(z)$ and $G'_n(z)$ is very small. However, the evaluation of $G'_n(z)$ still requires the parent distribution to be known. An alternative is to consider possible limiting distributions for Z_n as $n \to \infty$, using an extreme value analog of the central limit theorem. But, just as the average $\bar{X}_n$ converges to the population mean μ with certainty, i.e., has a degenerate distribution (converges to a constant), the limiting distribution of Z_n is also degenerate, converging with probability 1 to the upper endpoint of F or diverging if the distribution has unbounded support. Therefore, a normalization is necessary. In the case of the

sample average, the function $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to the standard normal distribution.

A similar linear rescaling can be applied to the extreme order statistics, Z_n, to obtain a non-degenerate limit distribution. The extremal types theorem [FISHER; TIPPETT, 1928] shows that if the distribution of the linearly rescaled block maxima as the size of the blocks approaches infinity is non-degenerate, then this limiting distribution is one of three distributions (Gumbel, Fréchet, or reverse Weibull), which can all be described in terms of the generalized extreme-value distribution [JENKINSON, 1955],

$$G(z) = \begin{cases} \exp\left[-\left\{1 + \xi(z-\mu)/\sigma\right\}^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\left\{-(z-\mu)/\sigma\right\}\right], & \xi = 0, \end{cases} \tag{3.2}$$

whose domain is the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$ are the location, scale, and shape parameters. More precisely, if there are sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that $(Z_n - b_n)/a_n$ has a non-degenerate limit distribution as $n \to \infty$, then

$$\lim_{n\to\infty} F^n(a_n z + b_n) = G(z) \tag{3.3}$$

for every continuity point z of G. In practice, we assume that

$$P\{Z_n \leq z\} = P\{a_n^{-1}(Z_n - b_n) \leq a_n^{-1}(z - b_n)\} \doteq G\{a_n^{-1}(z - b_n)\}$$

for a finite block size n, i.e., the distribution of block maxima can be approximated by a generalized extreme-value distribution whose location and scale parameters depend on the block size, but its shape parameter does not [COLES, 2001, p. 48].

If $\xi = 0$, $\xi > 0$, or $\xi < 0$, the generalized extreme-value distribution is the Gumbel, Fréchet, or reverse Weibull distribution. The set of parent distributions F for which the block maxima have the same limit distribution is called a *max-domain of attraction*, denoted $\mathrm{MDA}(\xi)$. For example, the uniform distribution is an element of $\mathrm{MDA}(-1)$, while the t-distribution, the Cauchy, log-gamma, and Pareto distributions belong to the Fréchet max-domain of attraction, $\mathrm{MDA}(\xi)$ with $\xi > 0$. Koutsoyiannis [2004a] noted that, at least until the date of his writing, the Gumbel distribution was commonly used to model extreme rainfall, to the point where it was the only distribution mentioned in hydrological engineering textbooks. He explained that one of the reasons for this was theoretical: "most types of parent distribution functions that are used in hydrology, such as exponential, gamma, Weibull, normal and lognormal, belong to the domain of attraction of the Gumbel distribution."

The relationship between the block maxima distribution, $G'_\nu(x)$, and the tail of the parent distribution, i.e, the distribution of threshold exceedances of $u > 0$, $H(x) = P\{X \leq x \mid X > u\}$, can be seen by taking u such that the exceedance probability $1 - F(u)$ equals $1/\nu$, the reciprocal of the mean number of events in a block, and noting that

$$1 - H(x) = \frac{1 - F(x)}{1 - F(u)} = \nu\{1 - F(x)\}, \quad x \geq u.$$

Comparing this with equation 3.1, we have

$$H(x) = 1 + \log G'_\nu(x).$$

Thus, if the parent distribution F is in the domain of attraction of one of the extreme-value distributions, then the distribution of $Y = X - u$, conditional on $X > u$, converges to the generalized Pareto distribution as u increases [PICKANDS, 1975],

$$H(y) = \begin{cases} 1 - (1 + \xi y/\tau_u)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\tau_u), & \xi = 0, \end{cases} \tag{3.4}$$

whose domain is the set $\{y : y > 0, 1 + \xi y/\tau_u > 0\}$, where $\tau_u = \sigma + \xi(u - \mu) > 0$. The value of τ_u depends on the threshold except when $\xi = 0$. When $\xi = 0$ and $\xi = -1$, the generalized Pareto distribution is the exponential distribution with mean τ_u and the uniform distribution on $[0, \tau_u]$. Pareto distributions are obtained when $\xi > 0$. The probability density function is

$$h(y) = \begin{cases} \tau_u^{-1} (1 + \xi y/\tau_u)^{-(1+1/\xi)}, & \xi \neq 0, \\ \tau_u^{-1} \exp(-y/\tau_u), & \xi = 0, \end{cases} \tag{3.5}$$

where $y = x - u$ with $x \geq u$ for $\xi \geq 0$, and $u - \tau_u/\xi > x \geq u$ for $\xi < 0$. Thus, the support of the generalized Pareto distribution is always bounded below by u, bounded above by $u - \tau_u/\xi$ if $\xi < 0$ and unbounded if $\xi \geq 0$.

The mean and variance of the generalized Pareto distribution exist only if $\xi \leq 1$ and $\xi \leq 1/2$, respectively. In such cases, $\mathbb{E}(Y) = \tau_u/(1-\xi)$ and $\mathrm{Var}(Y) = \tau_u^2/\{(1-\xi)^2(1-2\xi)\}$. In general, the rth moment of (3.2) or (3.4) exists only if $\xi < 1/r$. For $\xi = 0$, the limiting distributions Gumbel and exponential have all moments finite.

In summary, if the block maxima have an approximate generalized extreme-value distribution, then the exceedances over a high threshold have a corresponding approximate distribution within the generalized Pareto family. Moreover, the parameters of the generalized Pareto distribution are uniquely determined by those of the associated generalized extreme-value distribution. In particular, the shape parameter has the same value in both distributions [COLES, 2001, p. 75].

The value of ξ determines the weight of the upper tail of the parent density, providing very different representations for the behavior of extreme values:

- if $\xi < 0$, the support has an upper bound;
- if $\xi > 0$, the support is not bounded above and the density function decays polynomially as $y \to \infty$, it is said to be heavy-tailed;
- if $\xi \to 0$, the generalized Pareto and generalized extreme-value distributions converge to exponential and Gumbel distributions (light-tailed densities that decay exponentially as $y \to \infty$).

Usually, $|\,\xi\,| < 1$ and estimates of ξ lie in the interval $(-1/2, 1/2)$, although this situation might be more common in environmental applications than in financial ones [DAVISON; HUSER, 2015].

3.2 The point process characterization

In the previous section, a Poisson point process was implicit. If the rescaled Z_n converges to the $G(z)$ given in equation (3.2), then, since $\log u \doteq u - 1$ when $u \to 1$,

$$\log F^n(z) = n \log F(z) \doteq n\{F(z) - 1\} \to \log G(z),$$

and after rearrangement, $1 - F(z) \doteq -n^{-1} \log G(z)$. For a threshold u, the number of exceedances follows a binomial distribution with parameters n and $p = -n^{-1} \log G(u)$. The Poisson limit for the binomial distribution implies that the rescaled variates $\{(X_i - b_n)/a_n : i = 1, \ldots, n\}$ converge to a Poisson process on $\mathbb{R}$ with mean measure $\Lambda_1\{[u, \infty)\} = -\log G(u)$. The two-dimensional point process $N_n = \{(i/(n+1), (X_i - b_n)/a_n) : i = 1, \ldots, n\}$, with mean measure in the time direction $\Lambda_2\{[t_1, t_2]\} = t_2 - t_1$, $0 \leq t_1 < t_2 \leq 1$, assuming homogeneity, converges to a Poisson process N on $[0, 1] \times \mathbb{R}$ with intensity measure $\Lambda\{[t_1, t_2] \times [u, \infty)\} = \Lambda_2\{[t_1, t_2]\} \times \Lambda_1\{[u, \infty)\}$, i.e,

$$\Lambda\{[t_1, t_2] \times [u, \infty)\} = \begin{cases} (t_2 - t_1)\{1 + \xi(u - \mu)/\sigma\}_+^{-1/\xi}, & \xi \neq 0, \\ (t_2 - t_1) \exp\{-(u - \mu)/\sigma\}, & \xi = 0. \end{cases} \tag{3.6}$$

So, times of exceedances of u occur according to a Poisson process of constant rate $\zeta_u = -\log G(u)$, and the limiting distribution of threshold exceedances belong to the generalized Pareto family, since

$$P\{X > u + y \mid X > u\} = \frac{\Lambda_1\{[u + y, \infty)\}}{\Lambda_1\{[u, \infty)\}} = \frac{-\log G(u + y)}{-\log G(u)} = H(y).$$

On the other hand, given the Poisson point process N, the event $\{(Z_n - b_n)/a_n \leq z\}$ is equivalent to the event of having no counts in the set $A_z = [0, 1] \times [z, \infty)$, i.e, $N_n(A_z) = 0$; hence

$$P\{(Z_n - b_n)/a_n \leq z\} = P\{N_n(A_z) = 0\} \to P\{N(A_z) = 0\} = \exp\{-\Lambda(A_z)\},$$

which is of the generalized extreme-value form. This representation of extremes has the advantage of accounting for the rate of exceedances ζ_u, and having the original parameters (μ, σ, ξ), which do not depend on the choice of u [COLES, 2001; DAVISON; HUSER, 2015].

3.3 Extracting the annual maxima

For the block maxima approach, the ideal would be to have complete time series. For example, for annual maxima taken over daily values, if there are many values missing, the

Table 3 – Number of stations and proportion of remaining daily values left after excluding from each station the years with more than a certain percentage of missing daily values.

Missing values (%)	Number of stations	Relative number of daily values (%)
0	940	70.84
5	940	73.58
10	947	89.91
15	947	89.97
20	950	95.03
25	950	95.76

maximum may not have been observed. However, rarely do data sets have complete records and we will simply have to settle for less, maybe for just three quarters of the time series' length [SVOBODA; HAYES; WOOD, 2012, p. 3]. Table 3 shows how many stations would remain if we removed from each station the observed years that have more than a certain percentage of missing values. To form the series of annual maxima, we adopt the same criteria used by Papalexiou e Koutsoyiannis [2013]: the maximum of each year is extracted irrespective of the year's missing-values percentage, then if the rank of a value in the series is smaller than or equal to $0.4 \times m$, where m is the number of years, and the missing-values percentage within the corresponding year is larger than or equal to $1/3$, we exclude this value, assuming it to be unknown.

3.4 Threshold selection

In the proof of the extremal types theorem, we see that a necessary condition for a limiting distribution for maxima is the *max-stability* property. A distribution function G is said to be max-stable if for any $n \in \mathbb{N}$ there are sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$G^n(a_n z + b_n) = G(z), \quad z \in \mathbb{R}.$$

In words, the max-stability property is satisfied when the distribution of a random variable remains identical, apart from a change of the location and scale parameters, after taking the maxima of n copies of this random variable. Only the generalized extreme-value family satisfies this property.

Similarly, the generalized Pareto distribution is also characterized by a *threshold stability* property [DAVISON; SMITH, 1990]: if X is a random variable having the generalized Pareto distribution with scale and shape parameters σ and ξ, $X \sim GPD(\sigma, \xi)$, then the conditional distribution of $X - u$ given $X > u$ is also generalized Pareto, $X - u \mid X > u \sim GPD(\sigma + \xi u, \xi)$. This property motivates a method for selecting a threshold by trying increasing values of u until

the distribution of the exceedances is judged to be appropriately described by the generalized Pareto distribution. If it is valid over a threshold u_0, then its validity must hold over thresholds $u > u_0$. In equation (3.4), this implies that parameter ξ must be constant for $u > u_0$, causing τ_u to vary linearly in u. Unfortunately, in practice this is not so clear-cut, and as mentioned before, the choice of u_0 involves a similar bias–variance trade-off when choosing block length: lower u_0 induces lower estimation variance but higher bias, and *vice versa*.

For the Poisson point process characterization, estimates of the parameters (μ, σ, ξ) should be stable across a range of thresholds. A plot of these estimates with pointwise confidence intervals is called a parameter stability plot (see Figure 13) and it is the main tool used in fixed threshold selection due to its simplicity. The threshold is selected as the lowest value above which the parameters are deemed to be constant. However, parameter stability plots suffer from a lack of interpretability, since the confidence intervals are strongly dependent.

As Wadsworth [2016] mentions, there is no panacea for this problem, but he proposes two complementary threshold diagnostic plots. Using his notation, if $\mathbf{X}$, of random length N, contains realizations from a Poisson point process, with intensity $\lambda_{\boldsymbol{\theta}}(x)$ and mean measure $\Lambda_{\boldsymbol{\theta}}(R)$ on a interval R, where $\boldsymbol{\theta}$ represents the parameters of the model, we need to maximize the likelihood

$$L_R(\boldsymbol{\theta}) = \left\{ \prod_{i=1}^{N} \lambda_{\boldsymbol{\theta}}(x_i) \right\} \exp\{-\Lambda_{\boldsymbol{\theta}}(R)\},$$

partitioning R into nested intervals $R_1, \ldots, R_k$, where

$$R_j = (u_j, \infty), \quad j = 1, \ldots, k,$$

with $u_1 < \cdots < u_k$. The problem is to obtain the joint distribution of the maximum likelihood estimators $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k$, which come from overlapping samples. Wadsworth [2016] describes the asymptotic distribution of these estimators, and the immediate consequence that the increments $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \boldsymbol{\theta}_2 - \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_k$ are independent. Isolating ξ as the parameter of interest, and denoting ξ_j^*, for $j = 1, \ldots, k-1$, as the difference $\hat{\xi}_j - \hat{\xi}_{j+1}$ standardized by its asymptotic variance, the sequence $\boldsymbol{\xi}^* = (\xi_1^*, \ldots, \xi_k^*)$ is asymptotically a white-noise process, i.e, is a sequence of independent standard normal variables. Wadsworth [2016] proposes a simple changepoint model for $\boldsymbol{\xi}^*$ in order to compare the likelihood for a threshold u_j and that for the lowest threshold u_1. Therefore, the suggested two additional diagnostic plots show the white noise process and a likelihood ratio statistic against the threshold; see Figure 13.

This likelihood-based approach for threshold selection can be automated by testing the significance of the largest likelihood ratio statistic. If this test is not significant, the lowest threshold is taken. For the Pomerode station, the selected threshold was the 63% quantile for wet days, or 8.4 mm, resulting in 5,418 exceedances, which is about 4 and 7 times larger than the 90% and 95% quantile.

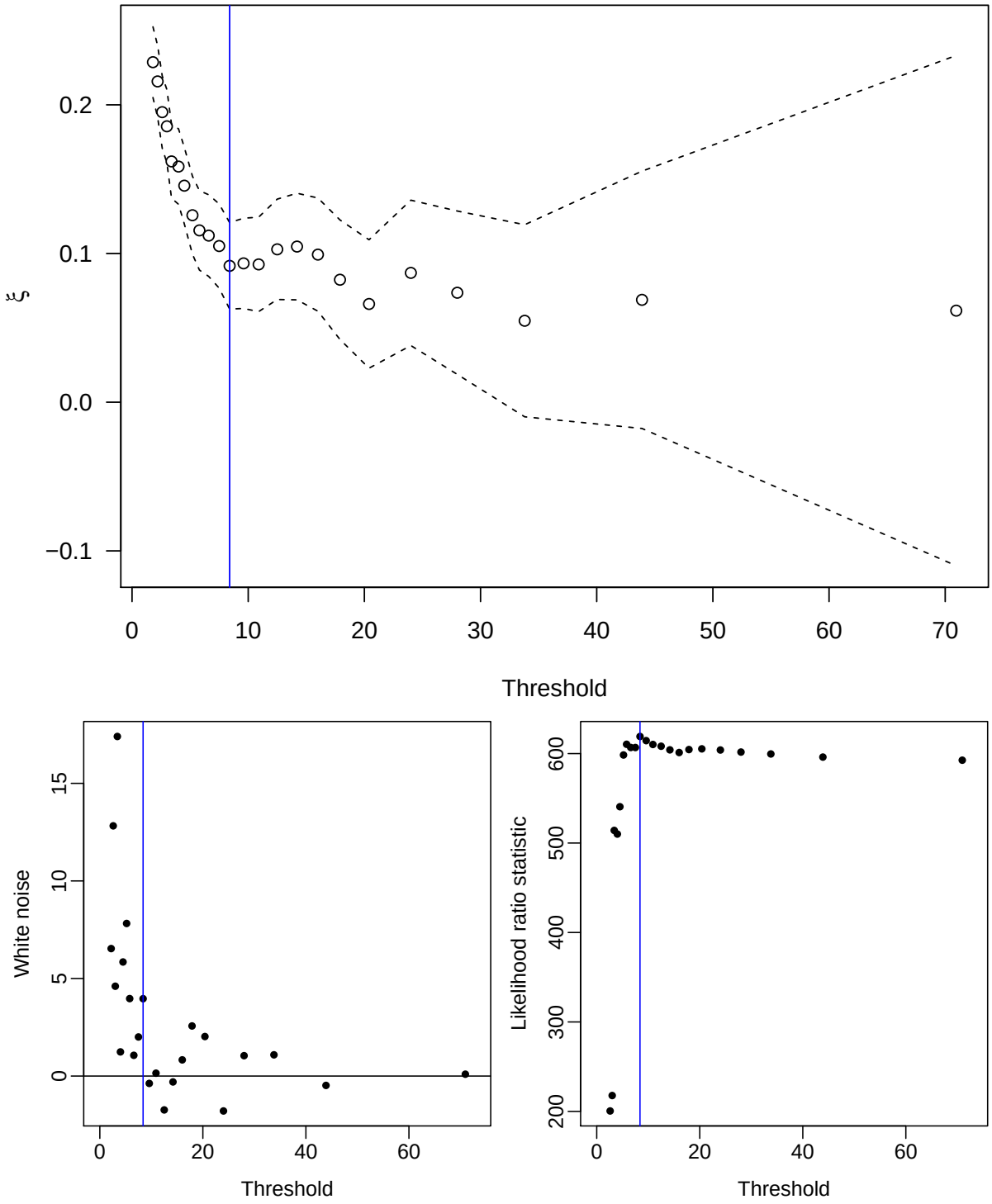

Figure 13 – Threshold diagnostic plots for ξ using the data from the Pomerode station. Top panel: parameter stability plot. Bottom panel: white noise process $\boldsymbol{\xi}^*$ (left) and likelihood ratio statistic (right). The blue line highlights the selected threshold.

3.5 Penultimate approximations

Convergence to the limiting distributions mentioned in Section 3.1 should not be taken for granted. Fisher e Tippett [1928] had already discussed this issue. They showed that convergence in distribution of the rescaled Z_n to the Gumbel limit for underlying normal variables is very slow (Davison e Huser [2015] illustrates this by means of an animation), and that the reverse Weibull distribution provides a better approximation for Z_n, n finite. In other words, even though the normal distribution belongs to the Gumbel max-domain of attraction, the Weibull form provides a better representation for finite samples. The Weibull is the penultimate approximation through which maxima from normal variables pass before reaching their ultimate destination. In most cases, the penultimate approximation has $\xi \neq 0$ even if the limiting distribution is of the Gumbel form [DAVISON; HUSER, 2015]. Fisher e Tippett [1928] and Smith [1987] proposed approximating the shape parameter for finite n, ξ_n, by $r'(b_n)$, where

$$r(x) = \frac{1 - F(x)}{f(x)}$$

is the reciprocal hazard function, and $b_n = F^{-1}(1 - 1/n)$.

Various distributions like the gamma, log-normal, Pareto, and mixed exponential have been used to model (positive) daily rainfall. Wilson e Toumi [2005] provided a physical justification, by interpretation of the water balance equation, for using the Weibull distribution for heavy daily rainfall, i.e,

$$F(x) = 1 - \exp\{-(x/\lambda)^k\},$$

for large x, with shape parameter $k = 2/3$ and scale parameter $\lambda > 0$. They expressed precipitation as the product of three independent random variables, each having approximately a normal distribution for daily totals (but not for shorter or longer timescales); they are the mass flux, the specific humidity, and the precipitation efficiency.

The probability of the product of k independent variables $X_1, \ldots, X_k$ is controlled by realizations where all terms in the product are of the same order, i.e, by the joint probability of all variables assuming common values, $X^{1/k}$ [FRISCH; SORNETTE, 1997]. Therefore, standardizing the three random variables that control rainfall, the probability density function of their product is, to a leading order, just the product of three standard normal densities evaluated at a common value $x^{1/3}$. The tail of the resulting distribution is $P\{X_1X_2X_3 > x\} \propto \exp\{-(x^{1/3})^2\}$, which is of the stretched exponential form; hence the Weibull distribution with shape parameter $k = 2/3$. According to Wilson e Toumi [2005], the shape of the tail should be "largely unaffected by climate change," and "the robustness of the shape parameter may now seem unsurprising given the physical basis of moisture conservation."

Figure 14 shows the probability distribution for daily rainfall in Pomerode. The Weibull distribution provides a great fit for all positive values and, indeed, the estimated shape parameter of 0.74 is close to the $2/3$ constant (the estimated scale parameter is $\hat{\lambda} = 8.7$). The

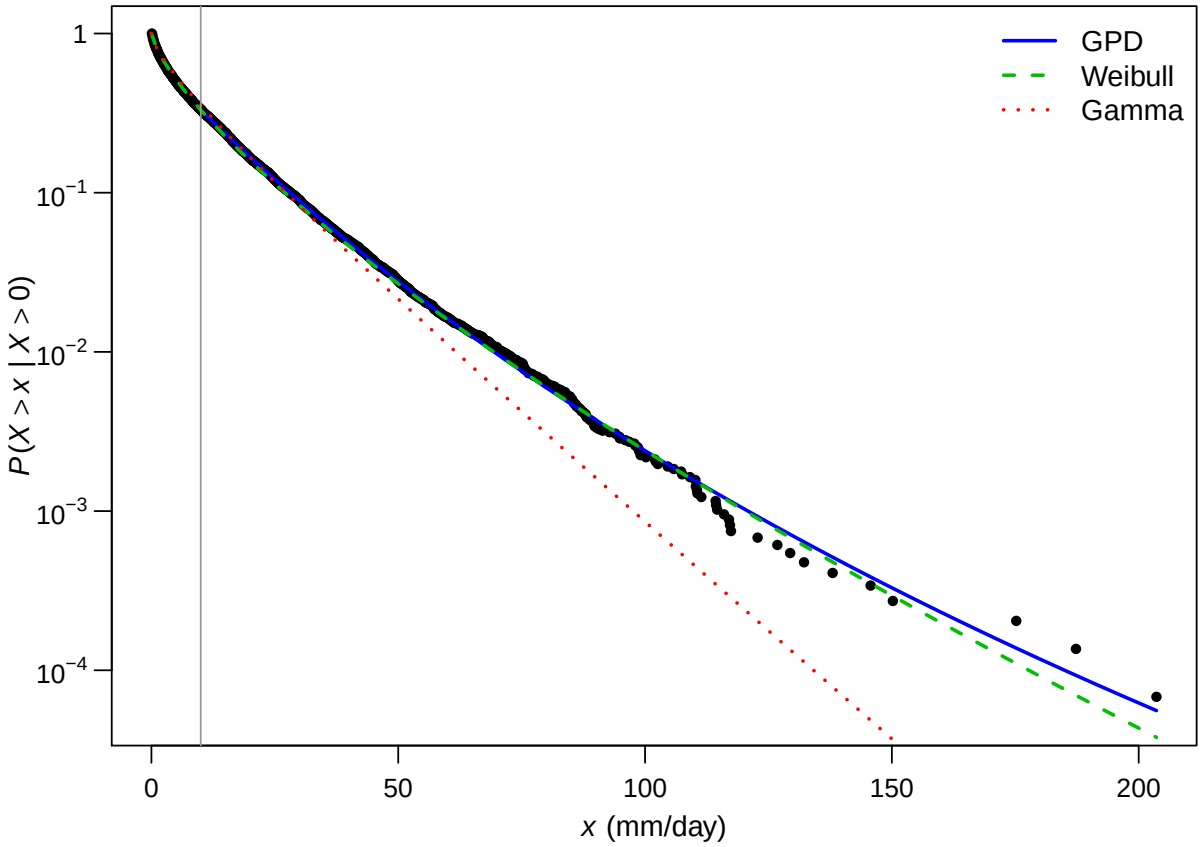

Figure 14 – The empirical probability $P\{X > x \mid X > 0\}$ according to x for the station in Pomerode, in the state of Santa Catarina. The gamma and Weibull distributions were fitted to all positive rainfall values, and the generalized Pareto distribution (GPD) was fitted for values greater than 10 mm. The three distributions were estimated by maximum likelihood.

generalized Pareto distribution is not so far off, with an estimated shape parameter of 0.10 (and $\hat{\tau}_{10} = 14.3$), reflecting the stretched exponential tail of the Weibull distribution.

The Weibull distribution is not threshold stable, so fitting it to different thresholds will provide different estimates for the shape parameter (in the example, $\hat{k} = 3.14, 4.51$ for values larger than 50 and 100). However, suppose F is Weibull. As mentioned before, the Weibull distribution belongs to the Gumbel max-domain of attraction, but its convergence is very slow. Koutsoyiannis [2004a] showed that even for $n = 10^6$ some departure is apparent between the Gumbel limit and the exact distribution function Z_n. Indeed, calculating $\xi_n = r'(b_n)$ for some specific values of n, Table 4, we see that either for 90 wet days per year (approximately one quarter of the year) or for 100% rainy days, the shape parameter will be close to 0.1, and only for very large block sizes, of the order 10^{16}, would ξ be close to zero. For $k = 1/2$, ξ_n is doubled, and for $k = 0$, ξ_n is always zero since the Weibull is reduced to the exponential distribution. Global analyses have showed that ξ is on average around 0.1 [KOUTSOYIANNIS, 2004b; WILSON; TOUMI, 2005; PAPALEXIOU; KOUTSOYIANNIS, 2013; SERINALDI; KILSBY, 2014], so if a Weibull distribution can be assumed for rainfall, its shape parameter should be

Table 4 – Values of $\xi_n = r'(b_n)$ for the Weibull distribution with different shape parameters k.

k \ n	90	365	3650	36500	365×10^{10}	365×10^{12}	365×10^{14}
$1/2$	0.22	0.17	0.12	0.10	0.03	0.03	0.00
$2/3$	0.11	0.08	0.06	0.05	0.02	0.01	0.00
$3/2$	-0.07	-0.06	-0.04	-0.03	-0.01	-0.01	0.00

close to $2/3$.

3.6 Checking the distributional assumptions

In order to diagnose and distinguish between distributions, we use the L-moment ratio diagram, introduced by Hosking [1990], a very useful graphical tool when a large number of samples are observed. This tool is especially useful for extreme values, since these are, by definition, scarce, and it is often difficult to detect a lack of fit using a statistical test that has enough power.

Hosking [1990] gathered and extended results about L-moments, providing a "unified approach to the use of order statistics for the statistical analysis of univariate probability distributions." L-moments are expectations of certain linear combinations of order statistics. Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be the order statistics of a random sample of size n drawn from the distribution F of a real-valued random variable X. The first four L-moments are

$$\begin{aligned}
\lambda_1 &= \mathbb{E}(X), \\
\lambda_2 &= \frac{1}{2}\mathbb{E}(X_{2:2} - X_{1:2}), \\
\lambda_3 &= \frac{1}{3}\mathbb{E}(X_{3:3} - 2X_{2:3} + X_{1:3}), \\
\lambda_4 &= \frac{1}{4}\mathbb{E}(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}).
\end{aligned}$$

These L-moments can be regarded as measures of location, scale, skewness and kurtosis. For example, the third L-moment, λ_3, is the central second difference of the median of a sample of size 3, i.e, the difference between the maximum and the median minus the difference between the minimum and the median. In general,

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \mathbb{E}(X_{r-k:r}), \quad r = 1, 2, \ldots. \tag{3.7}$$

A finite mean implies finite expectation of all order statistics. Thus, if X has finite mean, then all its L-moments exist. And, as Hosking [1990] shows, they fully characterize F. Moreover, the standardized L-moments of X, called L-moment ratios, the quantities $\tau_r = \lambda_r/\lambda_2$, $r = 3, 4, \ldots$, satisfy $|\tau_r| < 1$, $r \geq 3$, which makes them easier to interpret. In particular, τ_3 and τ_4 are named L-skewness and L-kurtosis.

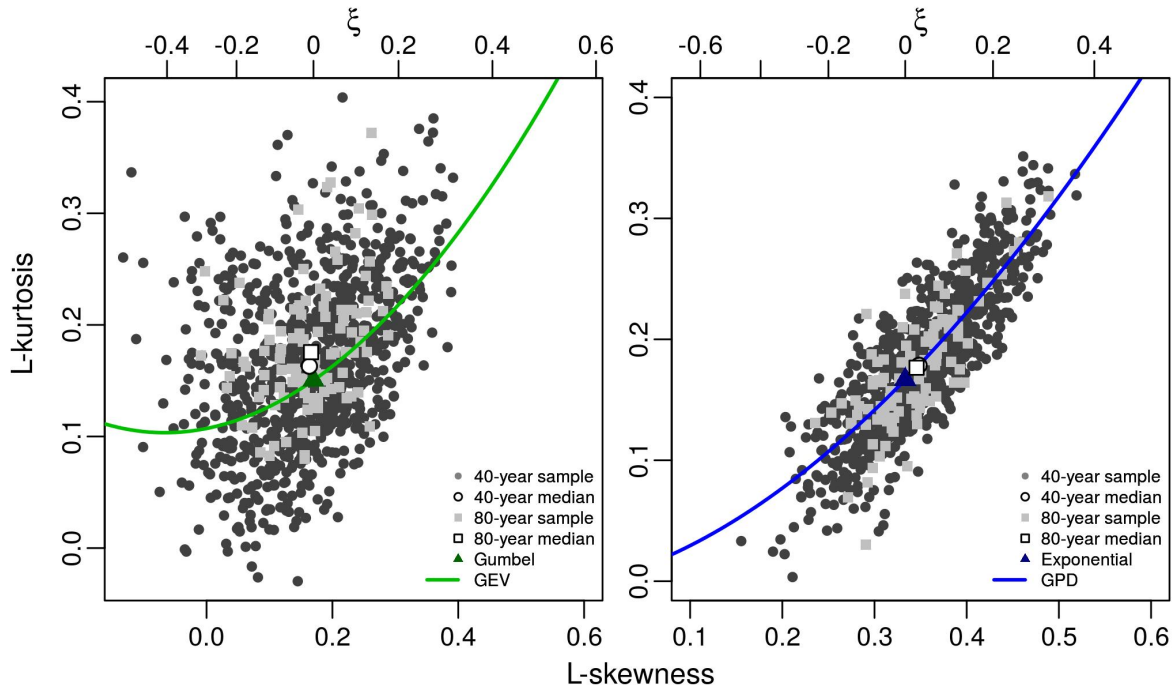

Figure 15 – Observed and theoretical L-kurtosis and L-skewness for all stations. Left panel: annual maxima. Right panel: exceedances above the 98% quantile for wet days. The two subsets, containing the shortest and longest series, are distinguished. The shape parameters of the generalized extreme-value and generalized Pareto distributions are shown at the tops of the plots.

Natural estimators of L-moments are U-statistics, i.e., the corresponding function of sample order statistics averaged over all subsamples of size r which can be constructed from the observed sample of size n. Let $x_1, x_2, \ldots, x_n$ be the sample and $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$ the ordered sample, the first four sample L-moments are

$$
\begin{aligned}
l_1 &= n^{-1} \sum_i x_i, \\
l_2 &= \frac{1}{2}\binom{n}{2} \sum_{i>j} \sum (x_{i:n} - x_{j:n}), \\
l_3 &= \frac{1}{3}\binom{n}{3} \sum_{i>j>k} \sum \sum (x_{i:n} - 2x_{j:n} + x_{k:n}), \\
l_4 &= \frac{1}{4}\binom{n}{4} \sum_{i>j>k>l} \sum \sum \sum (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}).
\end{aligned}
$$

The first four L-moments can be interpreted in the same way as the conventional moments, but they are guaranteed to exist if the first L-moment exist, the L-ratios are constrained to lie within the interval $(-1, 1)$, and their estimates are more robust to outliers. Ordinary sample moments of third and fourth order can be very unstable, especially when the sample consists of extreme values. These moments characterize the shape of a distribution, and indeed, the L-skewness and L-kurtosis for the generalized extreme-value distribution, or the generalized Pareto distribution, are simple functions of the shape parameter alone. For the generalized

Pareto distribution,

$$\tau_3 = (1+\xi)/(3+\xi),$$

$$\tau_4 = (1+\xi)(2+\xi)/\{(3+\xi)(4+\xi)\},$$

and for the generalized extreme-value distribution,

$$\tau_3 = 2(1-3^{\xi})/(1-2^{\xi}) - 3,$$

$$\tau_4 = \{5(1-4^{\xi}) - 10(1-3^{\xi}) + 6(1-2^{\xi})\}/(1-2^{\xi}).$$

Thus, for a large number of samples, we can compare the sample L-kurtosis and L-skewness of the block maxima, or the exceedances above a threshold, with the line given by the theoretical L-kurtosis and L-skewness; see Figure 15. We can also compare the resulting cloud of points with other distributions if there are simple expressions for their first two L-ratios (expressions which may define points, lines or regions in the L-moments ratio diagram), i.e., we can identify distributions without fitting them. For example, in Figure 15 we can visualize how many rainfall series are "close" to the Gumbel distribution, which is more informative than applying any of the more than 13 tests for the hypothesis that the shape parameter is zero in the generalized extreme-value distribution [HOSKING, 1984].

In Figure 15, there is not much difference between the two subsets, and the generalized Pareto distribution seems to fit better the exceedances than the generalized extreme-value distribution fits the annual maxima. Looking at the monthly maxima, Figure 16, there is a clear difference between the two subsets in some months, and sometimes the Weilbull distribution seems to provide a better description. Furthermore, for the last six months, the median point corresponding to the longest series clearly lies on the right of the Gumbel distribution.

3.7 Estimation methods

Hosking, Wallis e Wood [1985] and Hosking e Wallis [1987] used $\mathbb{E}\{ZG(Z)^r\}$ and $\mathbb{E}[Y\{1-H(Y)\}^r]$, $r = 0, 1$, as particular probability-weighted moments for the generalized extreme value and generalized Pareto distributions, respectively. The relationship between the parameters and these quantities is simpler than their relationship with conventional moments. Probability weighted moments are linear combinations of L-moments. As Hosking [1990] writes, "L-moments are more convenient, however, because they are more directly interpretable as measures of the scale and shape of probability distributions." In small samples, parameters estimates obtained from L-moments are less subject to bias, are closer to normality, and can be more accurate than maximum likelihood estimates. In a simulation study, Hosking, Wallis e Wood [1985] and Hosking e Wallis [1987] favored the probability-weighted moments when compared to likelihood estimation. Even though maximum likelihood estimators have the smallest possible asymptotic variances, the variance of the estimator by the probability-weighted moments method was smaller for small samples (with sizes $n = 15, 25$) and comparable for

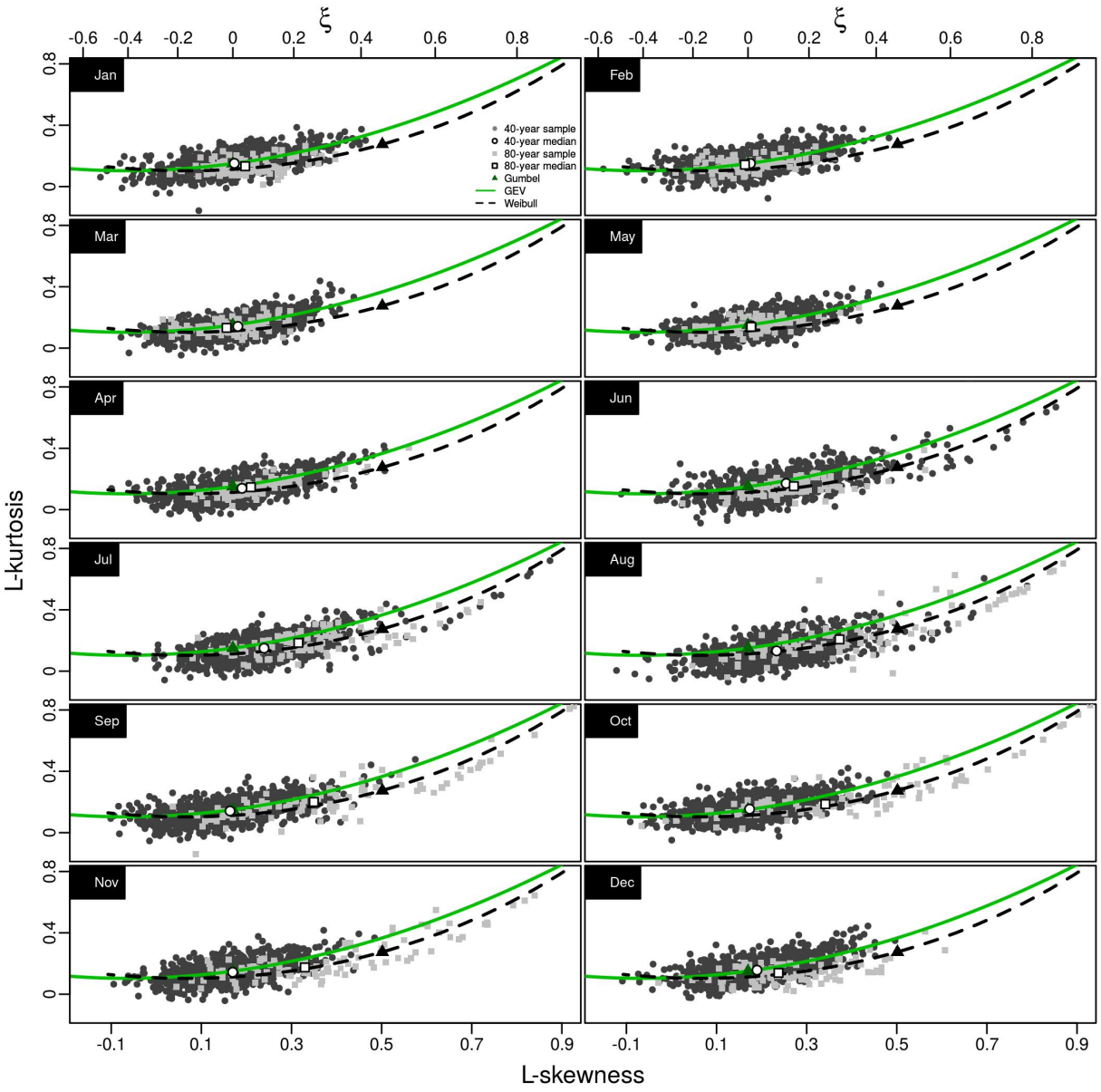

Figure 16 – Observed L-kurtosis and L-skewness for all stations and the theoretical lines of the generalized extreme-value (green line) and Weibull distributions (dashed line). The two subsets, containing the shortest and longest series, are distinguished. The shape parameters of the generalized extreme-value is shown at the tops of the plots.

moderate sample sizes ($n = 50, 100$). In both articles, they restricted their attention to the case $-1/2 < \xi < 1/2$. Smith [1985] showed that the maximum likelihood estimators obey the regularity conditions that are required for the usual asymptotic properties to be valid only if $\xi > -1/2$. When $\xi > 1/4$ and $\xi > 1/2$, Hosking e Wallis [1987] showed that the variances of the estimators by the method of moments and the probability-weighted moments are not of asymptotic order n^{-1}, respectively.

The likelihood framework provides a general approach to estimation and uncertainty assessment, but it is often based on first-order asymptotic theory. When the samples are not

Table 5 – Estimates with standard errors (in parentheses) for the parameters of the generalized extreme-value (GEV) and generalized Pareto (GPD) distributions using maximum likelihood (ML) and L-moments (LM). The threshold was set equal to 10 mm.

	GPD		GEV		
	τ_u	ξ	μ	σ	ξ
ML	14.29 (0.3)	0.096 (0.02)	78.1 (2.5)	21 (1.9)	0.05 (0.08)
LM	14.3	0.097	77.8	20.6	0.07

large, the maximum likelihood estimators may be unstable. And for the two extreme value distributions considered, a potential difficulty arises when estimating the parameters, because the end-points of the two extreme value distributions are functions of the parameter values.

For example, for a random sample $Y_1, \ldots, Y_n$ from the generalized Pareto distribution, since the support of the distribution is $y \geq 0$ for $\xi \geq 0$, and $\tau_u/\xi > y \geq 0$ for $\xi < 0$, the range of τ_u is $\tau_u > 0$ for $\xi \geq 0$, and $\tau_u > \xi Y_{(n)}$ for $\xi < 0$, where $Y_{(n)}$ is the extreme order statistic. When $\xi < -1$, the log-likelihood is not bounded above, so to obtain a finite maximum, the constraint $\xi \leq -1$ must be imposed. Therefore, maximum likelihood estimation is an optimization problem on the constrained space $\{\tau_u > 0, \xi > 0\} \cup \{-1 \leq \xi < 0, \tau_u > \xi Y_{(n)}\}$.

Differentiating the log-likelihood with respect to τ_u and ξ, the maximum with respect to ξ is achieved when

$$\xi = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \frac{\xi}{\tau_u} Y_i\right). \tag{3.8}$$

Reparametrizing the parameters (τ_u, ξ) to (θ, ξ), where $\theta = \xi/\tau_u$, and writing the profile likelihood for θ (substituting ξ with the expression in equation 3.8), yields

$$l(\theta) = -n - \sum_{i=1}^{n} \log(1 + \theta Y_i) - n \log\left\{ n^{-1} \sum_{i=1}^{n} \theta^{-1} \log(1 + \theta Y_i) \right\}.$$

Maximum likelihood estimation is reduced to a one-dimensional search on the space $\{\theta < 1/Y_{(n)}, \theta \neq 0\}$ [DAVISON; SMITH, 1990]. However, the optimization algorithm still has to be constrained, avoiding convergence to 0 and $1/Y_{(n)}$. And because of the behavior of the profile log-likelihood near $1/Y_{(n)}$, its first derivative will have multiple roots, each of which must be found. Grimshaw [1993] describes an efficient algorithm for obtaining the local maximum.

Table 5 shows the estimates for the parameters of the generalized extreme-value and generalized Pareto distributions using the methods discussed. The shape parameter estimates range approximately from about 0.05 to 0.1.

3.8 Bias corrections

In Section 3.5, we discussed bias due to penultimate approximation. Another source of bias comes from small samples. In Section 3.7, we mentioned that L-moments estimators are less subject to bias, but we did not consider improved maximum likelihood inference for the parameters of the extreme value distributions.

One question is how fast does the maximum likelihood estimator for ξ_n, given a finite n, converge to its mean. We simulated 1,000 rainfall series, each having 100-year length, from the Weibull distribution with $k = 2/3$ and $\lambda = 1$, considering three surreal cases: 90 and 10^4 wet days per year, and 100% rainy days, i.e, $n = 90, 365, 10^4$. We split the series in subsamples, taking the first $m = 10, 15, 20, \ldots, 100$ years, and fit by maximum likelihood the generalized extreme-value distribution. Then, for each m, we computed the average and the standard deviation of the estimates for the shape parameter. Plotting these empirical values against the sample size m, see Figure 17, they follow very closely the curve

$$\phi(m) = \alpha + \beta m^{-\gamma}, \quad \alpha, \beta \in \mathbb{R}, \gamma > 0, \tag{3.9}$$

where $\phi(m)$ can represent either the mean of the estimates, $\mu_\xi(m)$, or their standard deviation, $\sigma_\xi(m)$. Parameter α is the limit of $\phi(m)$ when $m \to \infty$, and parameter γ describes the rate at which $\phi(m)$ converges to α. The estimates for these parameters are shown in Table 6. The curves for the standard deviations are almost the same, irrespective of the block size. And, the smaller the block size, the faster is the convergence to the asymptotic limit α, to the point where the curve degenerates to a constant when the block size is 90.

Table 6 – Estimates with standard errors (in parentheses) for the parameters of equation (3.9) according to the block size n, i.e, the number of wet days from the Weibull distribution.

Statistic	Parameter	$n = 10^4$	$n = 365$	$n = 90$
$\mu_\xi(m)$	$\alpha_\mu \equiv \mu_\xi$	0.049 (0.002)	0.072 (0.001)	0.101 (0.003)
	β_μ	−2.6 (2.1)	−5.1 (5.8)	
	γ_μ	1.78 (0.34)	2.15 (0.48)	
$\sigma_\xi(m)$	$\alpha_\sigma \equiv \sigma_\xi$	0.031 (0.002)	0.04 (0.003)	0.035 (0.002)
	β_σ	2.2 (0)	2.3 (0.1)	2.1 (0.1)
	γ_σ	0.83 (0.01)	0.87 (0.02)	0.82 (0.02)

This idea of splitting the rainfall series in subsamples appears in Papalexiou e Koutsoyiannis [2013], where they propose a bias correction for an estimate of the shape parameter ξ based on a short rainfall series. Under the hypothesis that, for any finite sample size m, $\hat{\xi}(m)$ has a normal distribution with mean $\mu_\xi(m)$ and standard deviation $\sigma_\xi(m)$, i.e.,

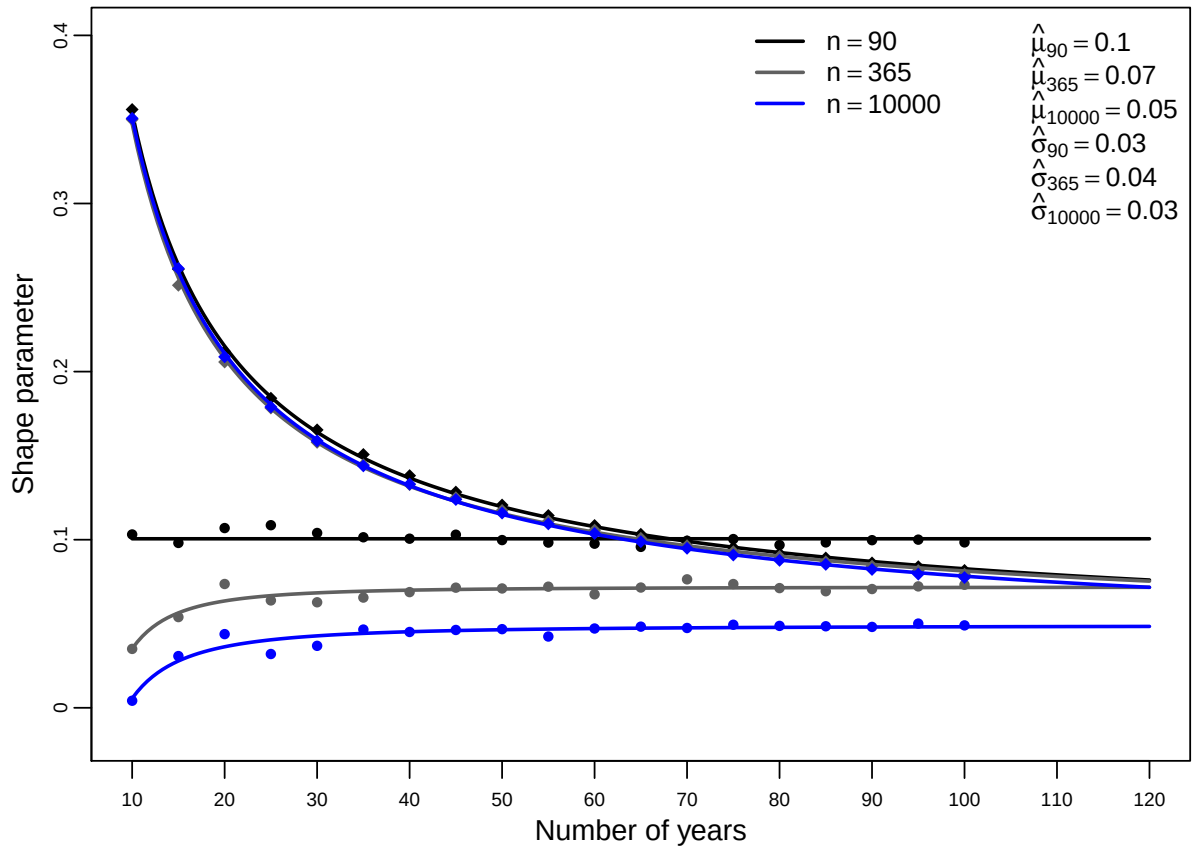

Figure 17 – Average and standard deviation of the maximum likelihood estimates for the shape parameter of the generalized extreme-value distribution according to subsamples of the 1,000 simulated rainfall series. The thick lines are fitted curves of the form (3.9).

$\hat{\xi}(m) \sim N\{\mu_\xi(m), \sigma_\xi(m)\}$, where $\mu_\xi(m) = \mu_\xi + \beta_\mu m^{-\gamma_\mu}$, $\sigma_\xi(m) = \sigma_\xi + \beta_\sigma m^{-\gamma_\sigma}$, and $\mu_\xi \equiv \alpha_\mu$ and $\sigma_\xi \equiv \alpha_\sigma$, the estimate

$$\tilde{\xi}(m) = \sigma_\xi \left\{ \frac{\hat{\xi}(m) - \mu_\xi(m)}{\sigma_\xi(m)} \right\} + \mu_\xi, \tag{3.10}$$

i.e., $\hat{\xi}(m)$ standardized by its true mean and standard deviation, and then transformed to a realization of the asymptotic limit distribution, $N(\mu_\xi, \sigma_\xi)$, is bias corrected.

In order to estimate the curves $\mu_\xi(m)$ and $\sigma_\xi(m)$, they used 15,137 rainfall series in several parts of the world, with lengths from 40 to 163 years. They found the normal distribution to be adequate, and the asymptotic parameters to be $\hat{\mu}_\xi = 0.114$ and $\hat{\sigma}_\xi = 0.045$.

Since equation 3.10 is a linear transformation, the ranks of the original estimates are preserved and also their spatial patterns (when present). After applying this transformation to their data, they found that, for all stations, the shape parameter of the generalized extreme value distribution was always positive, thus indicating that heavy-tailed distributions describe extreme daily rainfall values more accurately.

Serinaldi e Kilsby [2014] used a subset of Papalexiou and Koutsoyiannis' database to

investigate the impact of threshold selection and record length on the upper tail behavior of exceedances in the generalized Pareto model. After selecting 113 rainfall series spanning from 1900 to 2011 with less than 5% of values missing and dividing them into four seasons (summer, autumn, winter, and spring), they split the series in subsamples of length from 10 to 110 years by 5-year steps, and estimated via maximum likelihood the shape parameter of the generalized Pareto distribution for each record length and season.

However, Serinaldi e Kilsby [2014] arrived at different results: their correction preserved a large number of negative estimates for ξ, raising the question of an upper limit for precipitation in some geographic regions.

Aside from this empirical correction, a theoretical one can be obtained based on the work of Cox e Snell [1968]. For a random sample of size n from the generalized Pareto distribution with scale and shape parameters σ and ξ, Giles et al. [2011] and Previdelli e Davison [2011] showed that the biases of their maximum likelihood estimators are approximately

$$\mathbb{E}\left(\hat{\xi}-\xi\right) \doteq -\frac{(1+\xi)(3+\xi)}{n(1+3\xi)}, \quad \mathbb{E}\left(\hat{\sigma}-\sigma\right) \doteq \sigma\frac{(4\xi^2+5\xi+3)}{n(1+3\xi)}, \quad \xi > -1/3. \tag{3.11}$$

Since $\hat{\xi}$ is expected to have a negative bias, and $\hat{\sigma}$ a positive bias, the tail weight is typically underestimated. This correction is taking into account only the sample size bias, while the one proposed first by Papalexiou e Koutsoyiannis [2013] also takes into account the penultimate approximation bias. One can use the correction (3.10) or use the information gathered by this kind of study into a Bayesian framework.

3.9 Return levels

In practice, our main interest relies on the estimation of extreme quantiles or return levels. This is difficult since extreme values are, by definition, scarce. Moreover, we are often required to estimate events that are rarer than those already observed. Extreme value theory provides a class of models based on asymptotic arguments that allows extrapolation from observed to non-observed (future extreme) levels.

For the generalized Pareto distribution with scale and shape parameters σ and ξ, high quantiles, i.e., values satisfying

$$P\{Y \leq q_p\} = P\{X \leq u + q_p \mid X > u\} = 1 - p$$

for values of p close to 0, are obtained by solving the equation $p = 1 - H(q_p)$, giving

$$q_p = \begin{cases} \dfrac{\sigma}{\xi}\left\{\left(\dfrac{1}{1-p}\right)^{\xi} - 1\right\}, & \xi \neq 0, \\ -\sigma\log(1-p), & \xi = 0. \end{cases} \tag{3.12}$$

In common terminology, q_p is the return level associated with the return period $1/p$, since q_p is expected to be exceeded on average once every $1/p$ years. If $\xi < 0$, the upper endpoint of the support of Y is $q_1 = -\sigma/\xi$.

Another way of obtaining a return level is in terms of the mean time for some level to be to be exceeded. The (mean) level $x_m > u$ that is exceeded once every m observations is the solution of the equation relating frequency and period:

$$P\{X > x_m\} = \frac{1}{m}, \tag{3.13}$$

which can be solved by noting that

$$P\{X > x_m\} = P\{X > x_m, X > u\} = P\{X > u\}P\{X > x_m \mid X > u\},$$

where

$$P\{X > x_m \mid X > u\} = P\{Y > x_m - u\} = 1 - H(x_m - u) = p.$$

Denoting $\zeta_u = P\{X > u\}$, the solution is the same as in equation 3.12, but u is added and $1-p$ is replaced by $(m\zeta_u)^{-1}$. By construction, x_m is the m-observation return level. If there are n observations by year, the return level for N years is given by equation (3.12) with $m = N \times n$. As with the generalized extreme value model for block maxima, this quantity requires the estimation of three parameters, with σ and ξ determining the distribution of exceedances, and ζ_u the probability that the threshold u is exceeded. This is analogous to a semiparametric approximation to the parent distribution, using the empirical distribution function below the threshold u, and fitting the generalized Pareto distribution for the observations above u.

The quantile function can also be written in terms of the Box–Cox transformation as $q_p = -\sigma g(1-p; -\xi)$, where

$$g(z;\xi) = \begin{cases} \xi^{-1}(z^{\xi} - 1), & \xi \neq 0, \\ \log z, & \xi = 0. \end{cases} \tag{3.14}$$

For $\xi \neq 0$, the transformation $g(1-p;\xi)$ may be written, using the series expansions $e^x =$

$1 + x + (1/2)x^2 + \cdots$ and $\log(1+x) = x - (1/2)x^2 + \cdots$, as

$$\begin{aligned}
g(1-p;\xi) &= \frac{1}{\xi}\left\{e^{\xi\log(1-p)} - 1\right\} \\
&= \frac{1}{\xi}\left\{e^{\xi\left(-p-\frac{1}{2}p^2+\cdots\right)} - 1\right\} \\
&= \frac{1}{\xi}\left\{1 + \xi\left(-p - \frac{1}{2}p^2 + \cdots\right) + \frac{1}{2}\xi^2\left(-p - \frac{1}{2}p^2 + \cdots\right)^2 + \cdots - 1\right\} \\
&= -p - \frac{1}{2}p^2 + \frac{1}{2}\xi\left(-p - \frac{1}{2}p^2 + \cdots\right)^2 + \cdots \\
&= -p - \frac{1}{2}p^2 + \frac{1}{2}\xi\left\{-p\left(1 + \frac{1}{2}p + \cdots\right)\right\}^2 + \cdots \\
&= -p - \frac{1}{2}p^2 + \frac{1}{2}\xi p^2\left(1 + \frac{1}{2}p + \cdots\right)^2 + \cdots \\
&= -p\left\{1 + \frac{1}{2}(1+\xi)p + O(p^2)\right\}, \quad p \to 0.
\end{aligned}$$

Thus, for small p, the accuracy of $\hat{q}_p$ is largely determined by the accuracy of $\hat{\sigma}$.

For the generalized extreme-value distribution, high quantiles are obtained by solving the equation $p = 1 - G(q_p)$, giving

$$q_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0, \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{3.15}$$

The (mean) level x_N that is exceeded once every N years is given by equation 3.15 with $p = 1/N$.

For the city of Pomerode, the maximum likelihood estimate for the 25-year return level is $\hat{x}_m = \hat{x}_{365\times25} = 182$ mm using the generalized Pareto distribution with a threshold at 10 mm. Thus, once every 25 years, we might expect a daily rainfall in Pomerode to exceed about 182 mm.

Engineers are often conservative, designing their structures according to the upper bound of confidence intervals for the return level. A 95% Wald confidence interval for x_m is approximately $(162 \text{ mm}, 202 \text{ mm})$. Figure 18 displays the profile log-likelihoods for the shape parameter and the 25-year return level. The confidence interval obtained via the likelihood ratio statistic, $(164 \text{ mm}, 205 \text{ mm})$, is similar to the Wald interval. However, the profile likelihood is usually highly asymmetric, reflecting greater uncertainty for extreme quantiles, and should be preferred [COLES, 2001].

Taking annual maxima and using the generalized extreme-value distribution gives $\hat{x}_N = \hat{x}_{25} = 151$ mm and 95% confidence interval $(135 \text{ mm}, 180 \text{ mm})$, obtained via the likelihood ratio statistic. In the next sections, we explore these differences in the estimates.

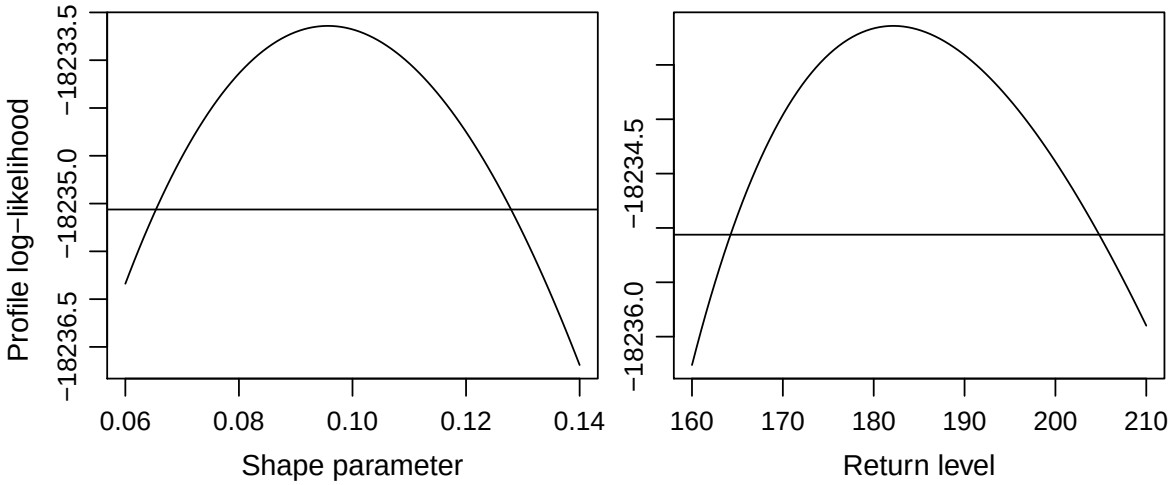

Figure 18 – Profile likelihood for ξ and the 25-year return level (left and right panels) using the generalized Pareto distribution with a threshold at 10 mm.

3.10 Extremal dependence

Knowledge that it rained heavily today might influence the probability of extreme rainfall in one or two days' time. Therefore, the independence assumption might not be reasonable for daily or hourly observations. Moreover, if we define the blocks for the maxima as being the weeks instead of the years, we may not even assume that we have an independent sample of block maxima or that the generalized extreme-value distribution is adequate.

Fortunately, the convergence of block maxima to the generalized extreme-value distribution still holds if long-range dependence at extreme levels is sufficiently weak [COLES, 2001]. If $X_1, \ldots, X_n$ is a stationary sequence of random variables, then, under suitable conditions, the distribution of Z_n is approximately $F^{n\theta}$, where $\theta \in (0, 1]$ is the extremal index, which quantifies the extent of extremal dependence: $\theta = 1$ for a independent process; $\theta \to 0$ for increasing levels of (extremal) dependence. The asymptotic distribution of Z_n is simply G^θ, which is still a generalized extreme-value distribution due to the max-stability property. However, when dependence is present, convergence of block maxima to the limit distribution will be slower, effectively reducing the block size n.

The extremal index of a stationary process summarizes the degree of clustering of its extremes [DAVISON; HUSER, 2015], and its value can be thought of as the reciprocal of the expected number of exceedances in a block of small length. To estimate θ, the *intervals estimator* of Ferro e Segers [2003], based on an asymptotic result for the times between threshold exceedances, can be used. For the station in Pomerode, the estimate is 0.77 with 95% confidence interval $(0.72, 0.82)$, so the average cluster size is a little larger than 1 day. Figure 19 illustrates how temporal dependence is weak for this rainfall series in Pomerode.

The modeling approach for dependent extremes is unchanged when using block max-

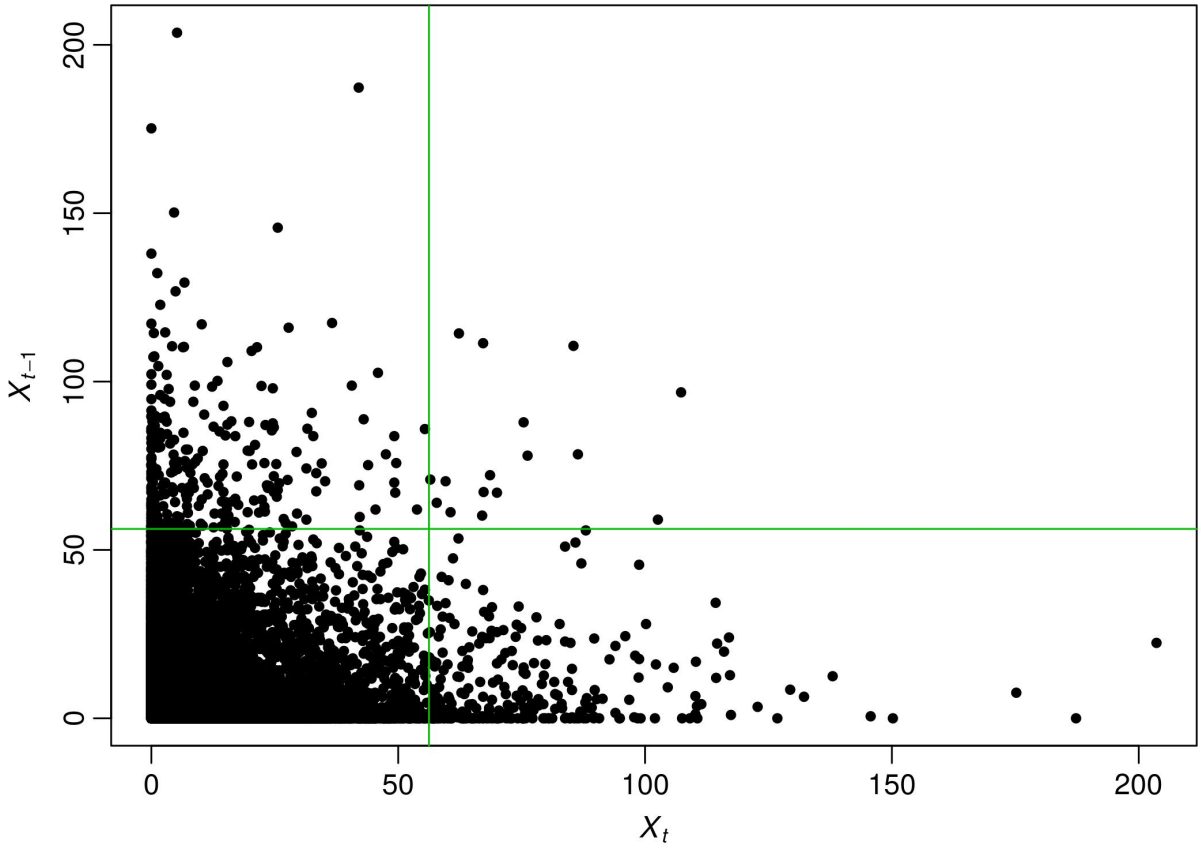

Figure 19 – Rainfall series of the station in Pomerode plotted against itself at lag 1. The green lines represents the 98% quantile for positive rainfall.

ima. For the threshold exceedance procedure, a popular approach is to filter out an approximately independent set of excesses, i.e, identify independent clusters and extract only their maximum. One way to identify clusters is through runs declustering, which assumes that exceedances belong to the same cluster if they are separated by fewer than a certain number, the run length, of values below the threshold. So, this approach, often referred to as the peaks over threshold approach [DAVISON; SMITH, 1990], works by:

1. choosing a run length κ (a "declustering parameter");
2. identifying clusters through the entire series. A cluster of threshold excesses is deemed to have terminated when at least κ consecutive observations fall below the threshold;
3. extracting the maximum observation from each cluster, and fitting the generalized Pareto distribution to these "peaks".

Although this is a very pragmatic method for dealing with clustered extremes, this and other declustering schemes are very wasteful of data and can introduce serious bias when estimating return levels [FAWCETT; WALSHAW, 2012]. When dependence is present, the return level

$x_N > u$ that is exceeded once every N years is the solution of the equation

$$P\{X > x_N\}^\theta = \frac{1}{N}, \tag{3.16}$$

since the distribution of Z_n at x_N is approximately $F^{n\theta}(x_N) = P\{X \leq x_N\}^\theta$, which is asymptotically $[1 - \zeta_u\{1 - H(x_N - u)\}]^\theta$. Solving for x_N gives

$$x_N = \begin{cases} u + \dfrac{\sigma}{\xi}\left[\zeta_u^{-1}\left\{1 - (1 - 1/N)^{1/\theta}\right\}\right]^{-\xi} - 1, & \xi \neq 0, \\ u - \sigma \log\left\{1 - \zeta_u^{-1}(1 - 1/N)^{1/\theta}\right\}, & \xi = 0. \end{cases} \tag{3.17}$$

Fawcett e Walshaw [2012] recommends the use of this equation, using all threshold excesses to estimate (ζ_u, σ, ξ), and the use of bias-corrected, accelerated bootstrap confidence intervals for x_N. However, as we show in the next chapter, there is only a slight tendency for rainfall extreme values to cluster, so the effect on inference is minimal.

3.11 Extremogram

The autocorrelation function determines the distribution of a stationary Gaussian sequence, but it does not capture the dependence structure of sequences whose finite-dimensional distribution have power-law tails. Davis e Mikosch [2009] defined an analog of the autocorrelation function, the extremogram, for the extremes of strictly stationary sequences of random vectors with finite-dimensional distributions regularly varying according to a positive tail index. The tail dependence coefficient of a one-dimensional strictly stationary sequence X_t,

$$\rho(h) = \lim_{u \to \infty} P\{X_{t+h} > u \mid X_t > u\} \quad h = 0, 1, 2, \ldots,$$

is a special case of the extremogram, and it can be interpreted as a particular autocorrelation function, providing quantitative descriptions of the persistence of a shock (an extreme event) at future instants of time, i.e, it allows us to visually check the size of clusters of extreme values. This quantity is a conditional probability of rare events, and its non-parametric estimation, the joint exceedances at lag h above some threshold u, requires large samples, which is why we use the extremogram for the subset with the longest series.

The cross-extremogram, introduced by Davis, Mikosch e Cribben [2012], is a measure of extremal serial dependence between two or more time series. For bivariate time series (X_t, Y_t), one particular choice for the cross-extremogram is

$$\rho(h) = \lim_{u \to \infty} P\{Y_{t+h} > u \mid X_t > u\} \quad h = 0, 1, 2, \ldots.$$

When calculating the cross-extremograms between two stations, only the days for which there are observations on both stations are used, so the sample sizes can be greatly reduced.

3.12 Extremal coefficient

In order to check for spatial dependence among sites, we use the F-madogram, a summary dependence measure similar to the semi-variogram. It was proposed by Cooley, Naveau e Poncet [2006], and it is defined as

$$\nu_F(s) = \frac{1}{2}\mathbb{E}\left[|F\{Z(x)\} - F\{Z(x+s)\}|\right], \quad x, s \in \mathbb{R}^2, \tag{3.18}$$

where F denotes the cumulative distribution function of a max-stable process $Z(x)$ over a random (precipitation) field in $\mathbb{R}^2$. The F-madogram is related to the extremal coefficient function as follows:

$$\theta(s) = \frac{1 + 2\nu_F(s)}{1 - 2\nu_F(s)}, \quad s \in \mathbb{R}^2. \tag{3.19}$$

When $\theta(s) = 2$, the process is an independent one, and when $\theta(s) \to 1$, the process exhibits increasing levels of spatial extremal dependence.

The F-madogram, and therefore the extremal coefficient function, is easily estimated by its empirical counterpart. To reduce sample variability, a binned version of this estimate is usually used, i.e., $\hat{\nu}_F$ is averaged over suitable classes of pairwise distances. See Ribatet, Dombry e Oesting [2016] for details.

3.13 Nonstationary extremes

Environmental processes typically present temporal nonstationarity. Different seasons may have different climate patterns, and long term trends might also be observable due to climate change. Furthermore, as discussed in Section 2.3, a rainfall series may be associated with other variables such as the multivariate ENSO index.

Davison e Smith [1990] discuss regression in the parameters of the generalized Pareto distribution using least squares or maximum likelihood methods of estimation, while keeping the threshold constant. Given a process X_t with associated covariates $\mathbf{V}_t$, they model the exceedance rate,

$$\zeta_u(\mathbf{v}_t) = P\{X_t > u \mid \mathbf{V}_t = \mathbf{v}_t\},$$

and the distribution of excesses Y_t of a threshold u by a generalized Pareto distribution with shape and scale parameters depending on the observed covariates $\mathbf{v}_t$. Since the scale parameter, $\sigma(\mathbf{v}_t)$, must be positive, it is natural to take its logarithm as a link function to the linear predictor $\mathbf{v}_t^T\boldsymbol{\sigma}$. For an independent sample of size n, this gives the likelihood

$$\prod_{t=1}^{n}\{1-\zeta_u(\mathbf{v}_t)\}^{1-I(x_t>u)}\left[\zeta_u(\mathbf{v}_t)\sigma(\mathbf{v}_t)^{-1}\left\{1+\xi(\mathbf{v}_t)\frac{x_t-u}{\sigma(\mathbf{v}_t)}\right\}_+^{-1/\xi(\mathbf{v}_t)-1}\right]^{I(x_t>u)}. \tag{3.20}$$

The *conditional return level* $q_{p,t}$, defined as the solution of

$$P\{X_t > q_{p,t} \mid \mathbf{V}_t = \mathbf{v}_t\} = p,$$

is, for $q_{p,t} > u$,

$$q_{p,t} = u + \frac{\sigma(\mathbf{v}_t)}{\xi(\mathbf{v}_t)}\left[\left\{\frac{\zeta_u(\mathbf{v}_t)}{p}\right\}^{\xi(\mathbf{v}_t)} - 1\right]. \tag{3.21}$$

To obtain a *(marginal) return level* q_p, we can assume a model for the joint density $f_{\mathbf{V}_t}$ of the covariates $\mathbf{V}_t$ and integrate them out,

$$P\{X_t > q_p\} = \int_{\mathbf{v}_t} P\{X_t > q_p \mid \mathbf{V}_t = \mathbf{v}_t\} f_{\mathbf{V}_t}(\mathbf{v}_t) d\mathbf{v}_t.$$

Under the assumption that the observerd covariates form a representative sample from this joint distribution in some specified period of interest, we can estimate it empirically,

$$P\{X_t > q_p\} = \frac{1}{n}\sum_{t=1}^{n} P\{X_t > q_p \mid \mathbf{V}_t = \mathbf{v}_t\}. \tag{3.22}$$

For $q_p > u$,

$$\begin{aligned} P\{X_t > q_p\} &= P\{X_t > q_p, X_t > u\} \\ &= \frac{1}{n}\sum_{t=1}^{n} P\{X_t > u \mid \mathbf{V}_t = \mathbf{v}_t\} P\{X_t > q_p \mid X_t > u, \mathbf{V}_t = \mathbf{v}_t\} \\ &= \frac{1}{n}\sum_{t=1}^{n} \zeta_u(\mathbf{v}_t)\left\{1+\xi(\mathbf{v}_t)\frac{q_p-u}{\sigma(\mathbf{v}_t)}\right\}_+^{-1/\xi(\mathbf{v}_t)} = p. \end{aligned} \tag{3.23}$$

After maximizing the likelihood (3.20), the maximum likelihood estimate $\hat{q}_p$ can be found by replacing the parameters in equation (3.23) by their estimates, and solving it numerically.

Due to the threshold stability property (see Section 3.4), the same covariates used for σ must be used for ξ so that this property remains valid. This is not desirable since the shape parameter can usually be considered as fairly constant. Moreover, $\sigma(\mathbf{v}_t)$ can only retain the same functional form if the link function is linear. These disadvantages were first noted by Eastoe e Tawn [2009], who suggested preprocessing the data before carrying out the extreme value analysis. Their strategy is first to try modeling nonstationarity in the bulk of the data, and then to apply a extreme value model to its residuals. Since the extremes might have a different form of nonstationarity than in the central portion of the data, or even after preprocessing, there may still have some nonstationarity leftover, they still include covariates in the generalized Pareto model, hoping the lack of threshold stability will at least be a minor problem.

As mentioned in the previous chapter, we are unsure about the quality of the data in its body, and, as pointed out by Davison e Smith [1990], preprocessing seems best confined to cases where the physical origin of the nonstationarity is well understood. So, we pursue another strategy; we use block maxima with linear covariate models in the parameters, taking the logarithm as a link for the scale parameter. The series might be considered approximately stationary during the period in which maxima occur, although the resulting effective block size may then be much reduced [DAVISON; HUSER, 2015]. Taking annual maxima would completely avoid the need to model seasonality, but it would also result in a loss of information (as seen in the beginning of this chapter, the number of annual maxima is several times smaller than the number of exceedances of some high threshold).

We divide the year into separate seasonal units and assume that extremes within each unit are stationary. There may be no natural or obvious partition of the year. For example, Coles e Pericchi [2003] used in their initial analyses a three-season structure (November-February, March-June, July-October), and they ended up using a two-season pattern, using the Bayesian paradigm to make inference on the changepoint between the seasons. They found evidence for "a seasonal breakdown that constitutes mid-November to April as the winter period, and the remaining months as the summer period."

Reboita et al. [2010] identified eight rainfall patterns in South America. They based their classification on the shape of several monthly rainfall series spread across South America, showed in Figure 20, which is very illustrative of the typical rainfall regimes present in South America. In the central parts (most of Peru, Bolivia, Paraguay, Brazil, and the north of Argentina), the monthly rainfall series have a typical "U" shape.

Brazil, in particular, has various rainfall regimes, but for the most part, the rainy season occurs during the austral summer (November through March). Exceptions are: the northern parts of the Amazon basin, where the wet season happens in the boreal summer (mostly from

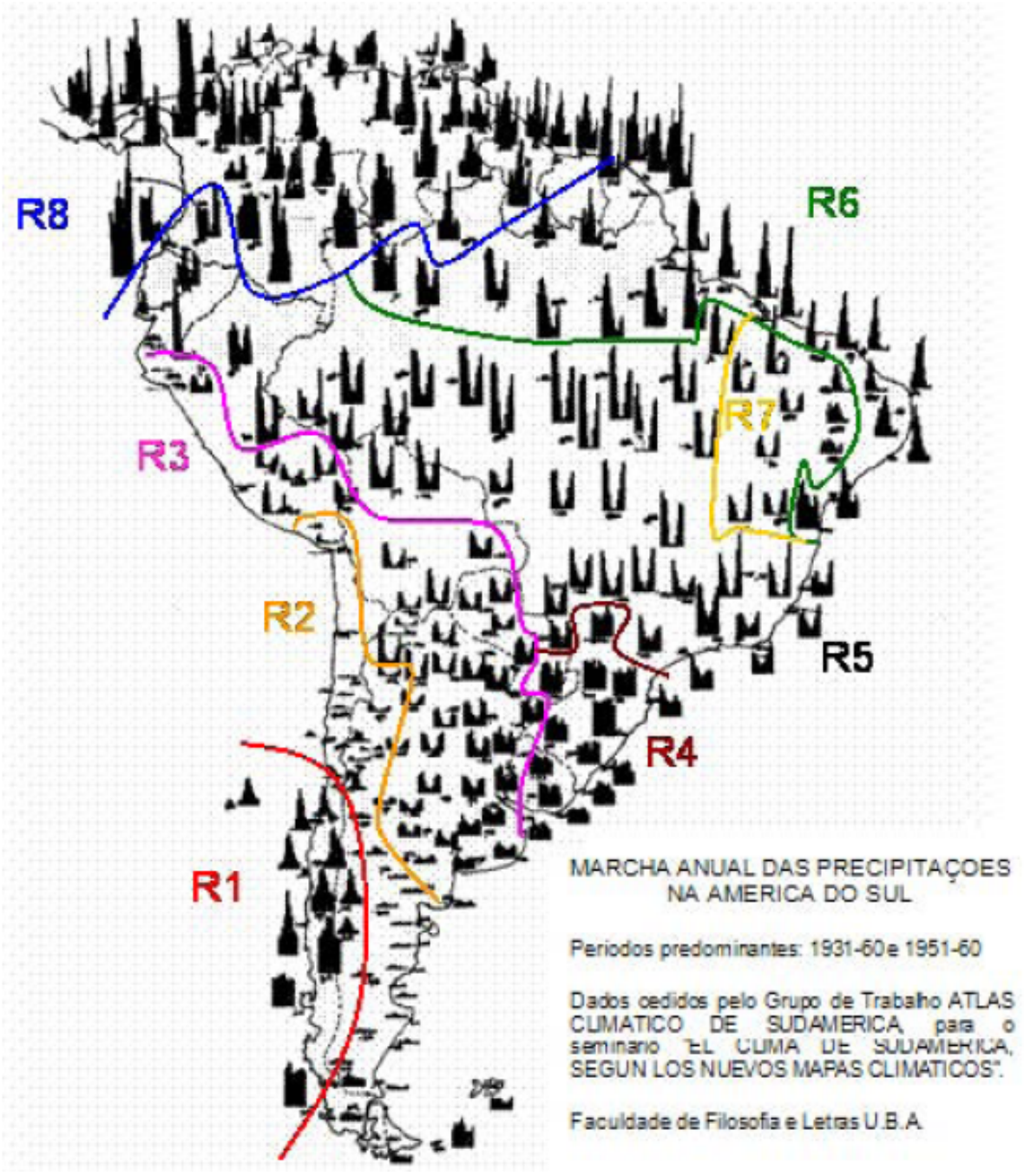

Figure 20 – Monthly rainfall series in South America and rough boundaries of the different rainfall patterns. The figure was elaborated by Universidad de Buenos Aires, and the boundaries were added by Reboita et al. [2010]. The graphs for some of the stations correspond to the period 1931–1960, and for others, the period 1951–1960.

June to August); the south region, which does not have a clear rainy and dry season (rainfall is relatively uniform across time); and the northeast of Brazil, where most of the rainfall occurs in a span of three months (this 3 month interval is quite heterogeneous for the whole region). A recent and detailed description of the rainfall patterns in Brazil can be found in Rao et al. [2016]. They mention that the central parts of the country have the main characteristic

of a monsoon region: "6 months of rain during the austral summer followed by 6 months of scanty rainfall in austral winter." Figure 2 in the aforementioned article shows the six rainiest consecutive months across Brazil. We use their figure to divide the year into two seasons of six months, one rainy and the other dry.

Retaining only the data from the rainy season, we use four approaches to rainfall extremes. For the longest series, we fit the generalized extreme-value distribution with time-varying parameters to monthly maxima. To account for seasonality and possible trend, we suppose the linear predictor for the location and scale parameters have the form

$$p_4(t) + \beta_1 \sin(2\pi t/365) + \beta_2 \cos(2\pi t/365), \tag{3.24}$$

where t is indexing, on a daily scale, when the maxima occurred, β_2 and β_3 are the amplitudes of the harmonic terms, and p_4 is an orthogonal polynomial of degree 4. We are taking nonstationarity into account only to properly estimate the shape parameter ξ, and borrow this information into a Bayesian setting to estimate return levels and covariate effects, using the shortest series, because the longest ones are too sparse, and the covariates have a short record. The monthly mean carbon dioxide and the deforestation rates start in 1980 and 1988, respectively. As for the multivariate ENSO index, the National Oceanic and Atmospheric Administration provides two data sets, one from 1950 to 2016, and the other data from 1871 to 2005. Only the former dataset is updated regularly.

So, we estimate 25-year return levels using a Bayesian hierarchical model, described in the next section, to extreme areal rainfall. This model could theoretically include temporal covariates, but its current implementation in the package **SpatialExtremes** only allows spatial covariates. So, to investigate the effect of ENSO, CO_2 levels, and deforestation rates, we take an approach similar to the one used for the longest series, keeping the seasonal components in (3.24), and replacing $p_4(t)$ by a linear trend plus the covariates of interest. We note that there is no need to consider the lag of the ENSO effect on precipitation because of the way the ENSO index is built.

As to the dependence measures mentioned in Sections 3.11 and 3.12, we need to have stationary series, and so, some kind of preprocessing of the data is necessary. Eastoe e Tawn [2009] use a Box–Cox location–scale model of the form

$$\frac{X_t^{\nu(\mathbf{v}_t)} - 1}{\nu(\mathbf{v}_t)} = \mu(\mathbf{v}_t) + \sigma(\mathbf{v}_t) Z_t,$$

where μ, $\log \sigma$, and ν are linear functions of the covariates. The residuals, Z_t, should be approximately stationary. Instead, we use a generalized additive model for location, scale and shape (GAMLSS), assuming a Box–Cox t distribution, a generalization of the Box–Cox normal distribution, to X_t, i.e,

$$Z_t = \frac{1}{\sigma(\mathbf{v}_t)\nu} \left\{ \left(\frac{X_t}{\mu_t(\mathbf{v}_t)} \right)^{\nu} - 1 \right\}, \quad \mu_t, \sigma_t > 0, \, \nu \neq 0. \tag{3.25}$$

Table 7 – Number of parameters p, negative maximized log-likelihood l, and p-value of the likelihood ratio test for the successive nested models.

Model	p	l	p-value
1. Time homogeneous	3	744	
2. As 1. plus the ENSO effect in μ and $\log\sigma$	5	743	0.49
3. As 2. plus the CO_2 effect in μ and $\log\sigma$	7	743	1.00
4. As 3. plus linear trend in μ and $\log\sigma$	9	743	0.99
5. As 3. plus the deforestation effect in μ and $\log\sigma$	9	743	0.96
6. As 3. but periodic in μ	9	742	0.00
7. As 6. but periodic in $\log\sigma$	11	742	0.46

Table 8 – Estimates with standard errors (in parentheses) for the parameters of the selected model.

	Intercept	ENSO	CO_2	Sine	Cosine
μ	40.6 (33)	-0.7 (1.7)	0 (0.1)	2.2 (2)	1 (4.8)
$\log\sigma$	2.79	-0.07 (0.08)	0		
ξ	0.06 (0.06)				

The residuals Z_t should follow a truncated t distribution with $\tau > 0$ degrees of freedom. X_t is said to have a Box–Cox t distribution with parameters μ, σ, ν, and τ. This distribution and its extension to regression are described in Rigby e Stasinopoulos [2006]. The covariates in μ and $\log\sigma$ were the seasonal components in (3.24) plus a B-spline with three degrees of freedom. We could include covariates in ν, the parameter controlling skewness, but we found out, by using the generalized Akaike information criterion with different penalties for the parameters, that this was not necessary for any of the rainfall series.

Table 7 shows some possible models fitted to the data in Pomerode. Based on the likelihood ratio statistics, we selected model number 5, which does not include a linear trend in $\log\sigma$. Table 8 shows the parameters estimates for the selected model. The effect of the ENSO phenomenon is expected to be positive (see Figure 7), but it is not significant (at least not for this particular station and time period). There is also no evidence for a carbon dioxide effect. Figure 21 shows the conditional return levels for some return periods. These return levels are much smaller than those found in Section 3.9. Marginal return levels are obtained using equation (3.22) and solving for q_p, with p small,

$$p = P\{Z_t > q_p\} = \frac{1}{n}\sum_{t=1}^{n}\left\{1 - \exp\left(-\left[1 + \xi\left\{q_p - \mu(t)\right\}/\sigma(t)\right]^{-1/\xi}\right)\right\}, \tag{3.26}$$

where Z_t is the monthly maxima at time t, and n is the sample size. Solving equation (3.26), the estimate for the marginal 25-year return level is $\hat{q}_{1/25} = 104$.

If conditional on the fitted parameter values, the maxima at month k, Z_k, follows an

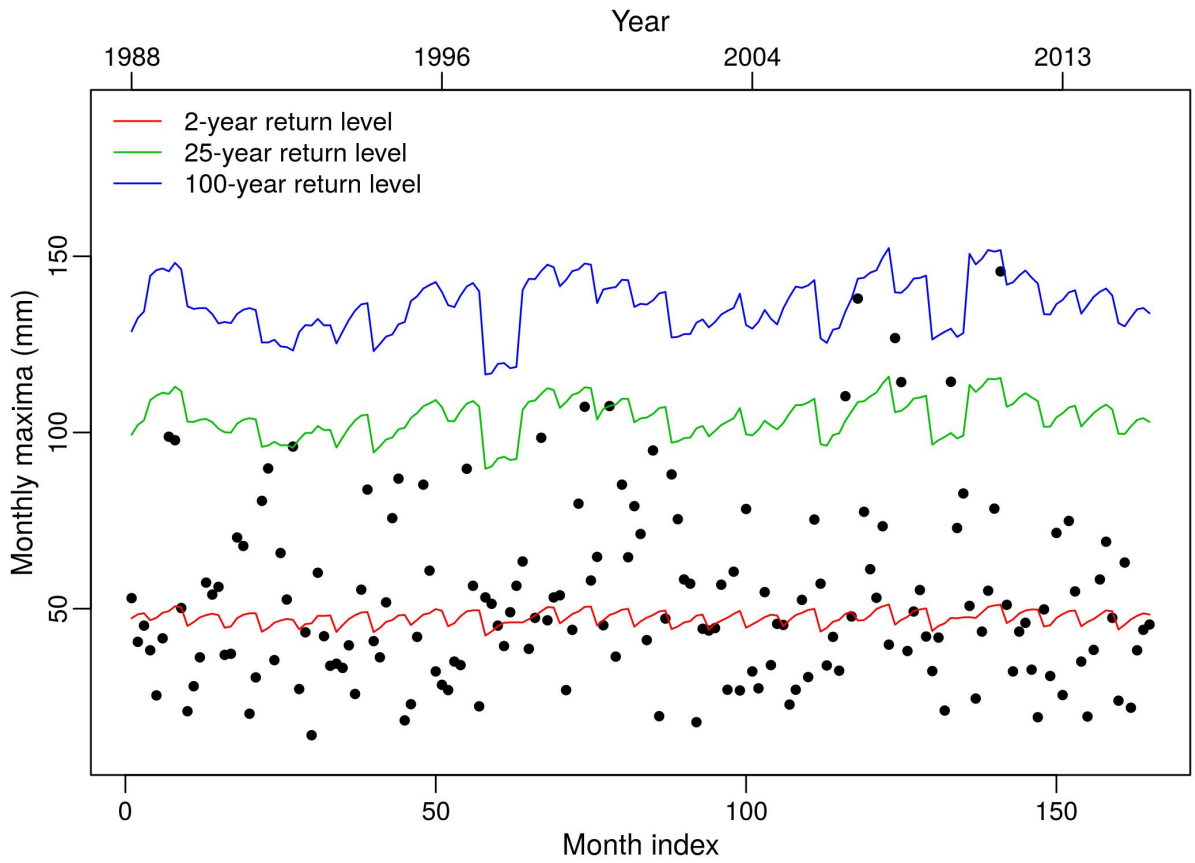

Figure 21 – Conditional return levels and monthly maxima at the station of Pomerode.

extreme-value distribution with parameters $\{\hat{\mu}(k), \hat{\sigma}(k), \hat{\xi}\}$, then the standardized variable

$$\begin{aligned}\tilde{Z}_k &= -\log\left[-\log\left\{\exp\left(-\left[1+\hat{\xi}\left\{Z_k-\hat{\mu}(k)\right\}/\hat{\sigma}(k)\right]^{-1/\hat{\xi}}\right)\right\}\right] \\ &= -\log\left(-\left[1+\hat{\xi}\left\{Z_k-\hat{\mu}(k)\right\}/\hat{\sigma}(k)\right]^{-1/\hat{\xi}}\right) \\ &= \hat{\xi}^{-1}\log\left[1+\hat{\xi}\left\{Z_k-\hat{\mu}(k)\right\}/\hat{\sigma}(k)\right]\end{aligned}$$

has a standard Gumbel distribution, and we can make probability and quantile plots of the observed $\tilde{z}_k$ with reference to this distribution [COLES, 2001]; see Figure 22. On both scales, the linearity of the points indicates a good fit of the model.

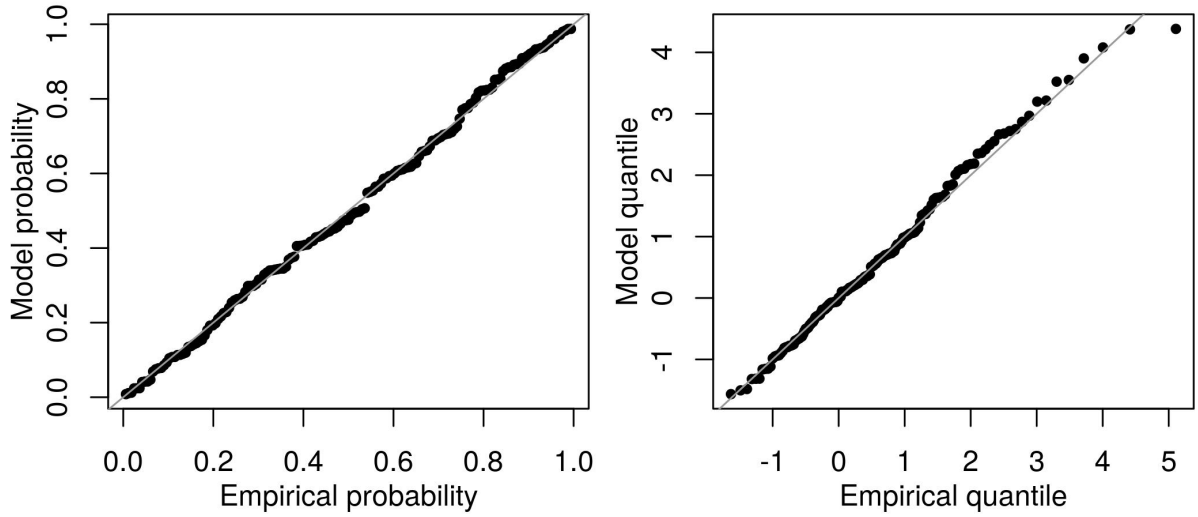

Figure 22 – Residual probability and quantile plots (left and right panels) of a nonstationary model using the generalized extreme-value distribution for monthly maxima at the station of Pomerode.

3.14 Bayesian hierarchical model

We take rainfall annual maxima (in order to avoid modeling seasonality), and we use the same latent variable model described in Davison, Padoan e Ribatet [2012], assuming that the generalized extreme-value parameters in each station with coordinates $s \in \mathbb{R}^2$, $\{\eta(s), \tau(s), \xi(s)\}$, vary linearly according to three independent Gaussian processes, i.e,

$$\phi(s) = f_\phi(s; \boldsymbol{\beta}_\phi) + S(s; \alpha_\phi, \lambda_\phi), \quad \phi = \eta, \tau, \xi,$$

where f_ϕ is a deterministic function depending on regression parameters $\boldsymbol{\beta}_\phi$, and S_ϕ is a zero mean, stationary Gaussian process with exponential covariance function $\alpha_\phi \exp(-\|h\|/\lambda_\phi)$, where α_ϕ is the sill parameter and $\exp(-\|h\|/\lambda_\phi)$ is the Matérn correlation function with shape parameter $\kappa = 0.5$. For the location and scale parameters of the generalized extreme-value distribution, we allow their means to depend on spatial covariates (longitude, latitude, and mean annual precipitation – MAP),

$$\eta(s) = \beta_{0,\eta} + \beta_{1,\eta}\,\mathrm{lon}(s) + \beta_{2,\eta}\,\mathrm{lat}(s) + \beta_{3,\eta}(s)\mathrm{MAP}(s), \tag{3.27}$$

$$\tau(s) = \beta_{0,\tau} + \beta_{1,\tau}\,\mathrm{lon}(s) + \beta_{2,\tau}\,\mathrm{lat}(s) + \beta_{3,\eta}(s)\mathrm{MAP}(s), \tag{3.28}$$

$$\xi(s) = \beta_{0,\xi}. \tag{3.29}$$

Thus, conditional on the values of the Gaussian processes, the maxima for n years observed at D stations are assumed to be independent with

$$\begin{aligned} &Y_t(x_d) \mid \{\eta(x_d), \tau(x_d), \xi(x_d)\} \\ &\qquad \sim \mathrm{GEV}\{\eta(x_d), \tau(x_d), \xi(x_d)\}, \quad t = 1, \ldots, n,\ d = 1, \ldots, D. \end{aligned}$$

A joint prior density must be defined for the mean parameters, $\boldsymbol{\beta}_\eta$, $\boldsymbol{\beta}_\tau$, and $\beta_{0,\xi}$, and for those of the covariance function, $\boldsymbol{\alpha} = (\alpha_\eta, \alpha_\tau, \alpha_\xi)^T$, $\boldsymbol{\lambda} = (\lambda_\eta, \lambda_\tau, \lambda_\xi)^T$. So, we first did an exploratory analysis of the marginal distributions for the rainfall annual maxima. After fitting of the generalized extreme-value distribution and obtaining estimates $\{\hat{\eta}(x), \hat{\tau}(x), \hat{\xi}(x)\}$, we compute their empirical variograms and fit the exponential covariance function by maximum likelihood.

Since the inverse gamma and the multivariate normal distributions are conjugate priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, they are used in order to reduce the computational burden. We attribute independent priors with large variances, but with means similar to the maximum likelihood estimates:

- Normal priors for the regression parameters with means $\boldsymbol{\mu}_\eta^* = (80, 0, 0)^T$, $\boldsymbol{\mu}_\tau^* = (20, 0, 0)^T$, and $\mu_\xi^* = 0.06$, and covariance matrices $\Sigma_\eta^* = \Sigma_\tau^* = \mathrm{diag}(400, 100, 100)$, and $\Sigma_\xi^* = 50$;
- Inverse gamma distributions as priors for $\boldsymbol{\alpha}$, with shape parameters $\boldsymbol{\kappa}_\alpha^* = (\kappa_{\alpha_\eta}^*, \kappa_{\alpha_\tau}^*, \kappa_{\alpha_\xi}^*)^T = (1/2, 1/2, 1/2)^T$ and scale parameters $\boldsymbol{\theta}_\alpha^* = (\theta_{\alpha_\eta}^*, \theta_{\alpha_\tau}^*, \theta_{\alpha_\xi}^*)^T = (180/2, 30/2, 0.03/2)^T$;

- Gamma priors for $\boldsymbol{\lambda}$ with shapes $\boldsymbol{\kappa}^*_\lambda = (\kappa^*_{\lambda_\eta}, \kappa^*_{\lambda_\tau}, \kappa^*_{\lambda_\xi})^T = (10, 10, 5)^T$ and scales $\boldsymbol{\theta}^*_\lambda = (\theta^*_{\lambda_\eta}, \theta^*_{\lambda_\tau}, \theta^*_{\lambda_\xi})^T = (7, 7, 3)^T$.

The joint density for the data, $\mathbf{y}$, and all these parameters, $\boldsymbol{\Theta}$, is

$$\begin{aligned}
\pi(\mathbf{y}, \boldsymbol{\Theta}) &= \pi(\mathbf{y}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\beta}_\eta, \boldsymbol{\beta}_\tau, \beta_{0,\xi}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\kappa}^*_\alpha, \boldsymbol{\theta}^*_\alpha, \boldsymbol{\kappa}^*_\lambda, \boldsymbol{\theta}^*_\lambda, \boldsymbol{\mu}^*_\eta, \boldsymbol{\mu}^*_\tau, \mu^*_\xi, \Sigma^*_\eta, \Sigma^*_\tau, \Sigma^*_\xi) \\
&= \pi(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\xi}) \\
&\times \pi(\boldsymbol{\eta} \mid \boldsymbol{\beta}_\eta, \alpha_\eta, \lambda_\eta)\pi(\alpha_\eta \mid \kappa^*_{\alpha_\eta}, \theta^*_{\alpha_\eta})\pi(\lambda_\eta \mid \kappa^*_{\lambda_\eta}, \theta^*_{\lambda_\eta})\pi(\boldsymbol{\beta}_\eta \mid \boldsymbol{\mu}^*_\eta, \Sigma^*_\eta) \\
&\times \pi(\boldsymbol{\tau} \mid \boldsymbol{\beta}_\tau, \alpha_\tau, \lambda_\tau)\pi(\alpha_\tau \mid \kappa^*_{\alpha_\tau}, \theta^*_{\alpha_\tau})\pi(\lambda_\tau \mid \kappa^*_{\lambda_\tau}, \theta^*_{\lambda_\tau})\pi(\boldsymbol{\beta}_\tau \mid \boldsymbol{\mu}^*_\tau, \Sigma^*_\tau) \\
&\times \pi(\boldsymbol{\xi} \mid \beta_{0,\xi}, \alpha_\xi, \lambda_\xi)\pi(\alpha_\xi \mid \kappa^*_{\alpha_\xi}, \theta^*_{\alpha_\xi})\pi(\lambda_\xi \mid \kappa^*_{\lambda_\xi}, \theta^*_{\lambda_\xi})\pi(\beta_{0,\xi} \mid \mu^*_\xi, \Sigma^*_\xi). \qquad (3.30)
\end{aligned}$$

The posterior distribution is approximated using a Gibbs sampler. The full conditional distribution of a parameter ψ needed for this Markov chain Monte Carlo computation is obtained, up to a normalizing constant, by dropping the terms of the joint distribution in equation (3.30) which do not depend on ψ. For example, for the location parameter,

$$\begin{aligned}
\pi(\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\Theta}_{-\eta}) &\propto \pi(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\xi})\pi(\boldsymbol{\eta} \mid \boldsymbol{\beta}_\eta, \alpha_\eta, \lambda_\eta), \\
\pi(\boldsymbol{\beta}_\eta \mid \mathbf{y}, \boldsymbol{\Theta}_{-\boldsymbol{\beta}_\eta}) &\propto \pi(\boldsymbol{\eta} \mid \boldsymbol{\beta}_\eta, \alpha_\eta, \lambda_\eta)\pi(\boldsymbol{\beta}_\eta \mid \boldsymbol{\mu}^*_\eta, \Sigma^*_\eta), \\
\pi(\alpha_\eta \mid \mathbf{y}, \boldsymbol{\Theta}_{-\alpha_\eta}) &\propto \pi(\boldsymbol{\eta} \mid \boldsymbol{\beta}_\eta, \alpha_\eta, \lambda_\eta)\pi(\alpha_\eta \mid \kappa^*_{\alpha_\eta}, \theta^*_{\alpha_\eta}), \\
\pi(\lambda_\eta \mid \mathbf{y}, \boldsymbol{\Theta}_{-\lambda_\eta}) &\propto \pi(\boldsymbol{\eta} \mid \boldsymbol{\beta}_\eta, \alpha_\eta, \lambda_\eta)\pi(\lambda_\eta \mid \kappa^*_{\lambda_\eta}, \theta^*_{\lambda_\eta}),
\end{aligned}$$

where $\boldsymbol{\Theta}_{-\psi}$ contains all the model's parameters except ψ. Given a value of the Markov chain at iteration i, the next state of the chain is obtained as follows. Each component of $\boldsymbol{\eta}_i = \{\eta_i(x_1), \ldots, \eta_i(x_D)\}$ is updated separately. For station d, a proposal $\eta_p(x_d)$ is generated from a symmetric random walk and the acceptance probability (a likelihood ratio times a ratio of multivariate normal distributions),

$$\alpha\{\eta_i(x_d), \eta_p(x_d)\} = \min\left[1, \frac{\pi\{\mathbf{y}_d \mid \eta_p(x_d), \tau_i(x_d), \xi_i(x_d)\}}{\pi\{\mathbf{y}_d \mid \eta_i(x_d), \tau_i(x_d), \xi_i(x_d)\}} \frac{\pi(\eta_p \mid \beta_\eta, \alpha_\eta, \lambda_\eta)}{\pi(\eta_i \mid \beta_\eta, \alpha_\eta, \lambda_\eta)}\right],$$

is computed. With probability $\alpha\{\eta_i(x_d), \eta_p(x_d)\}$, the component $\eta_i(x_d)$ is updated to $\eta_p(x_d)$, otherwise it remains the same. For the regression parameters, $\boldsymbol{\beta}_\eta$ is drawn directly from a multivariate normal distribution having covariance matrix and mean vector

$$\{(\Sigma^*_\eta)^{-1} + \mathbf{X}^T_\eta \Sigma^{-1}_\eta \mathbf{X}_\eta\}^{-1},\ \{(\Sigma^*_\eta)^{-1} + \mathbf{X}^T_\eta \Sigma^{-1}_\eta \mathbf{X}_\eta\}^{-1}\{(\Sigma^*_\eta)^{-1}\mu^*_\eta + \mathbf{X}^T_\eta \Sigma^{-1}_\eta \boldsymbol{\eta}\},$$

where Σ_η is a $D \times D$ matrix determined by the covariance function, and $\mathbf{X}_\eta$ is the design matrix related to the regression coefficients $\boldsymbol{\beta}_\eta$. Also due to the use of a conjugate prior, α_η is drawn directly from an inverse gamma distribution. For the range parameter λ_η, the same procedure for the components in $\boldsymbol{\eta}_i$ is used. The parameters of the other two Gaussian processes are updated similarly (the proposal distributions in the Metropolis–Hastings steps are log-normal for the scale parameter).

Chapter 4

Results and discussion

In the first section, we consider the impact of record length on the estimate of the shape parameter, taking only the subset of the longest series. In the other sections, we use the subset of the shortest series to plot estimates of return level maps and the effects of some covariates, and to estimate dependence measures in extreme levels.

4.1 Impact of record length

For each station, we fitted by maximum likelihood the nonstationary model (3.24) to the monthly maxima in the rainy season, and proceeded as in Section 3.8, splitting the series in subsamples, taking the first $m = 10, 15, 20, \ldots, 95$ years, and refitting the model. Then, for each subsample, we computed the average and the standard deviation of the estimates for the shape parameter. Figure 23 shows the plot of these empirical values against the sample size m. We excluded 9 stations for having very discrepant estimates for ξ, leaving 95 stations.

Interestingly, a logistic model for $\mu_\xi(m)$ was more appropriate than model (3.9), i.e,

$$\mu_\xi(m) = \frac{\alpha}{1 + \exp\{(\nu - m)/\delta\}}, \quad \alpha \in \mathbb{R}, \nu, \delta > 0, \tag{4.1}$$

where α is the horizontal asymptote as $m \to \infty$, ν is the inflection point of the curve, and δ represents the distance between ν and the point where the curve is approximately $3\alpha/4$. The parameters estimates are given in Table 9. The estimate for the asymptotic value of μ_ξ is 0.06, lower than the value of more or less 0.1 found by previous studies [KOUTSOYIANNIS, 2004b; WILSON; TOUMI, 2005; PAPALEXIOU; KOUTSOYIANNIS, 2013; SERINALDI; KILSBY, 2014]. Our sample is most representative of northeast Brazil, about 65% of the 95 stations, followed by the southeast and the south (about 21% and 13% of the stations).

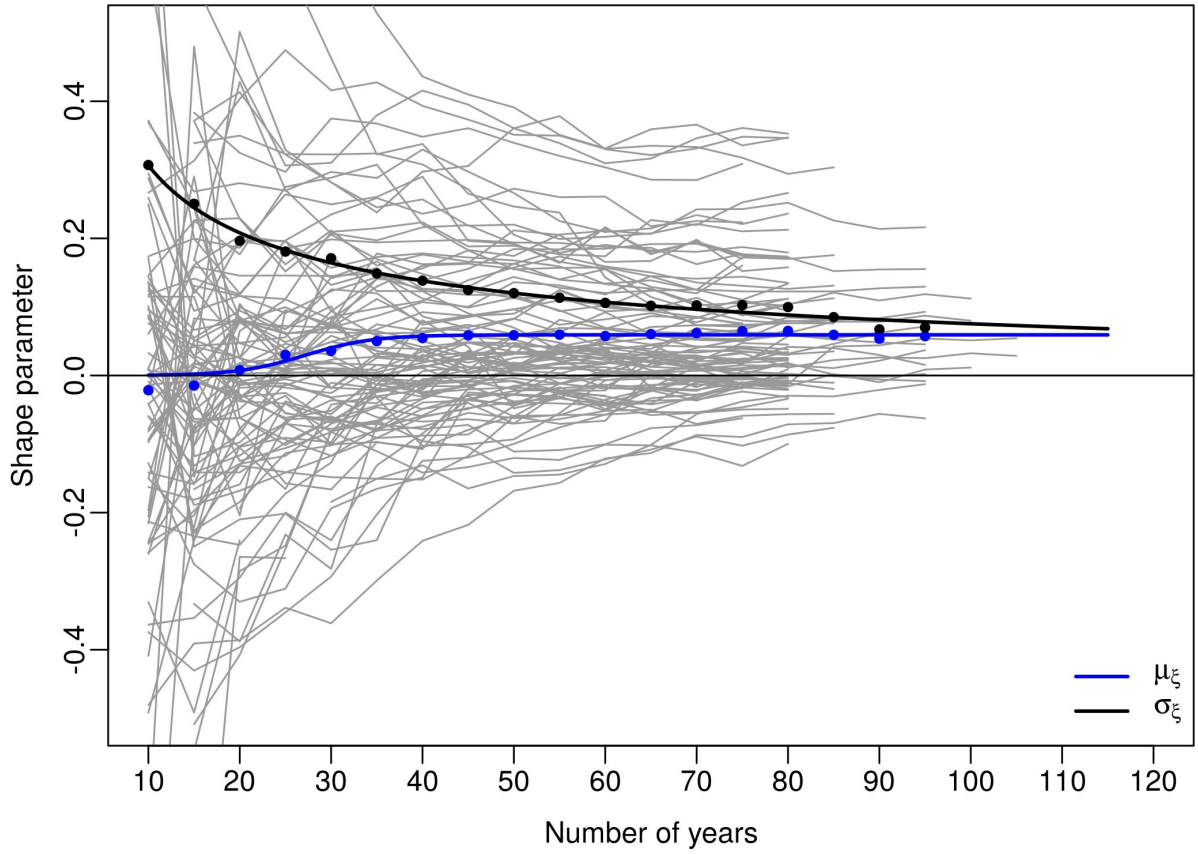

Figure 23 – Maximum likelihood estimates for the shape parameter of the nonstationary generalized extreme-value model according to subsamples of the longest rainfall series. Equations (3.9) and (4.1) were fitted to the average and the standard deviation of the estimates (blue and black lines).

Table 9 – Estimates with standard errors (in parentheses) for the parameters of equations (4.1) and (3.9), corresponding to μ_ξ and σ_ξ, respectively.

Parameter	μ_ξ	Parameter	σ_ξ
α	0.06 (0.002)	α	-0.03 (0.03)
ν	27 (1.5)	β	1.05 (0.12)
δ	4 (1.3)	γ	0.49 (0.08)

Table 10 – Porcentage of rejection, at the 5% significance level, of the likelihood ratio test for successive nested models.

Model	Rejection rate (%)
1. Time homogeneous	
2. As 1. plus the ENSO effect in μ and $\log\sigma$	12
3. As 2. plus the CO_2 effect in μ and $\log\sigma$	8
4. As 3. plus linear trend in μ and $\log\sigma$	2
5. As 3. plus the deforestation effect in μ and $\log\sigma$	4
6. As 3. but periodic in μ	88
7. As 6. but periodic in $\log\sigma$	17

4.2 Covariate effects

Since we are interested in plotting the covariate effects, we need to select the same model for each station. So, we follow the same approach as in the previous section, but we substitute $p_4(t)$ in (3.24) for the covariates. Based on Table 10, we opted not to include neither a linear trend nor a deforestation effect, keeping only an intercept plus the effects for the ENSO phenomenon and the carbon dioxide levels.

Then, we estimated the selected model by a Bayesian approach, using a informative prior for the shape parameter, a normal distribution with mean 0.06 (the asymptotic value found in Section 4.1) and standard deviation 0.02, and vague priors for the other parameters.

The next figures show, for the north and south regions and for the state of São Paulo, the posterior mean estimate of each covariate, as well as ordinary kriging based on maximum likelihood estimation for the underlying Gaussian random fields. As expected, we have a strong positive effect of the ENSO phenomenon in the south region (but not in São Paulo), and a negative effect in the north, while the effects due to the carbon dioxide levels are almost negligible. Surprisingly, there is also a strong seasonal effect in the south region despite the monthly mean being almost constant across the year.

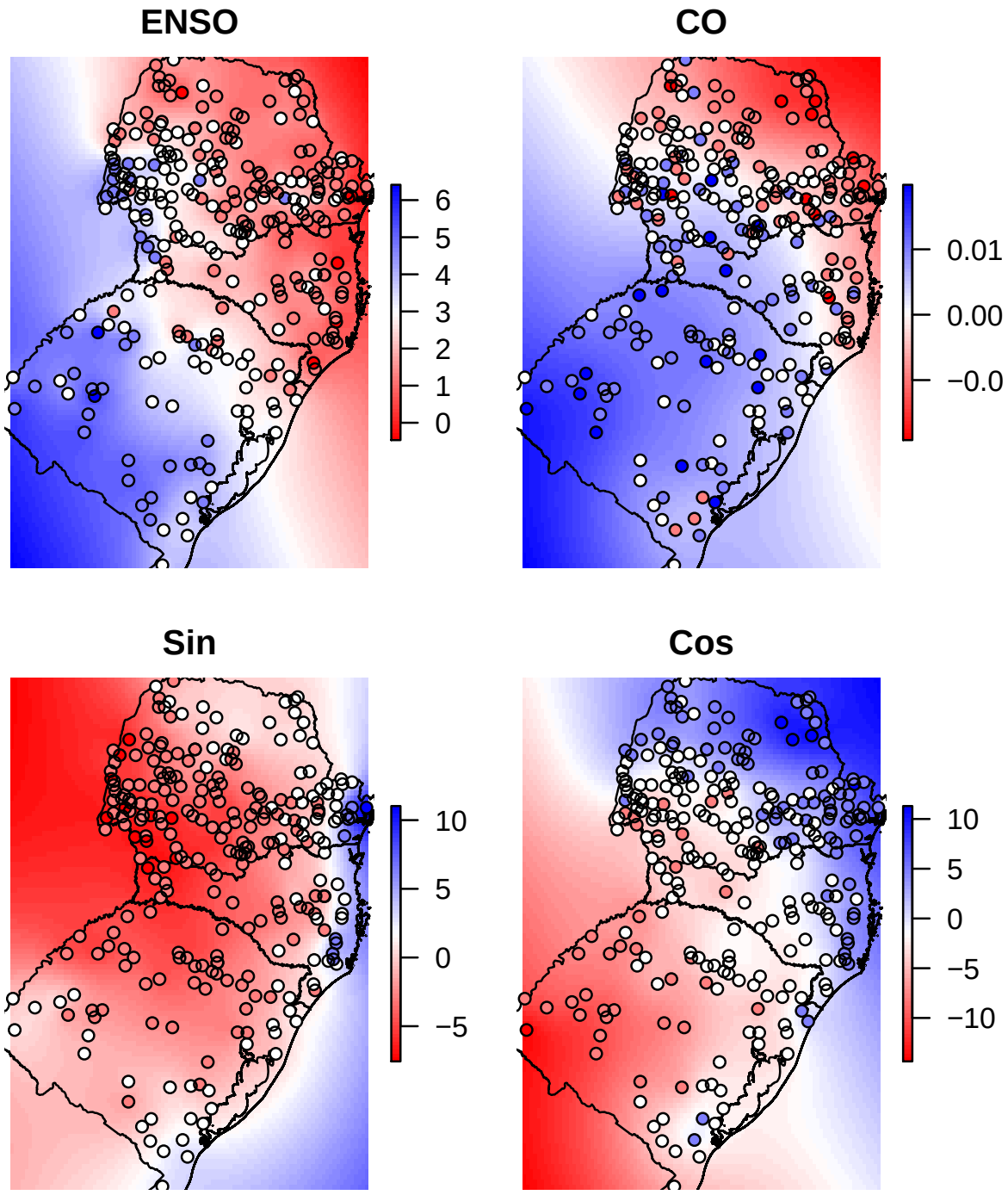

Figure 24 – Map of the south region and ordinary kriging for the coefficient estimates corresponding to each covariate in the location parameter of the generalized extreme-value distribution.

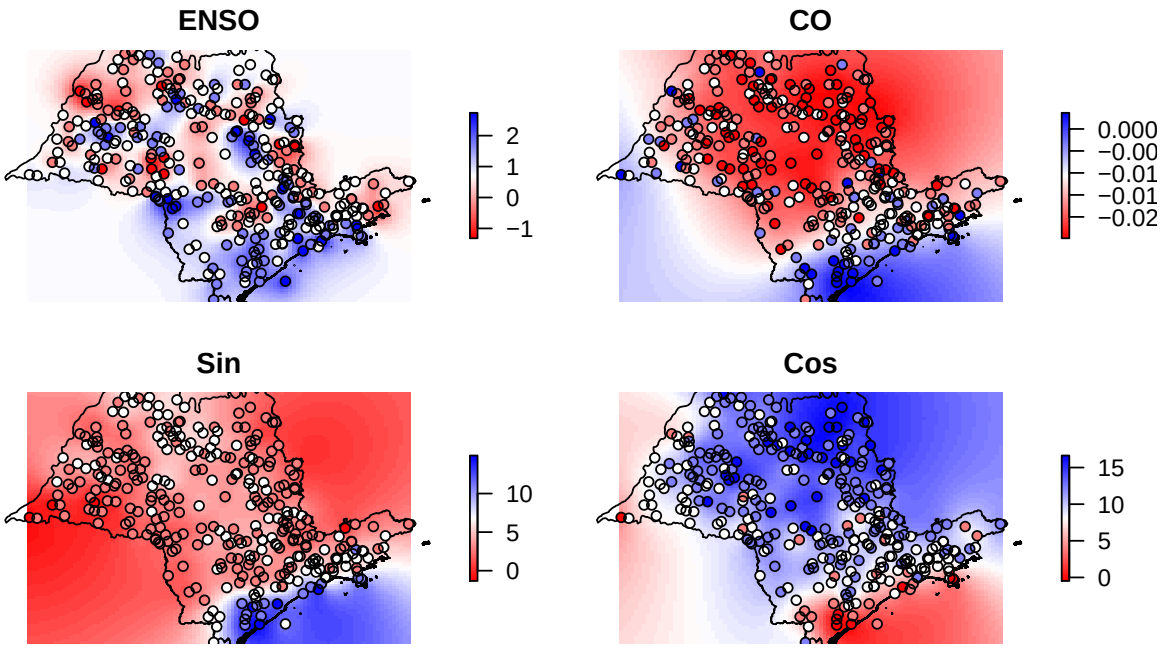

Figure 25 – Map of São Paulo and ordinary kriging for the coefficient estimates corresponding to each covariate in the location parameter of the generalized extreme-value distribution.

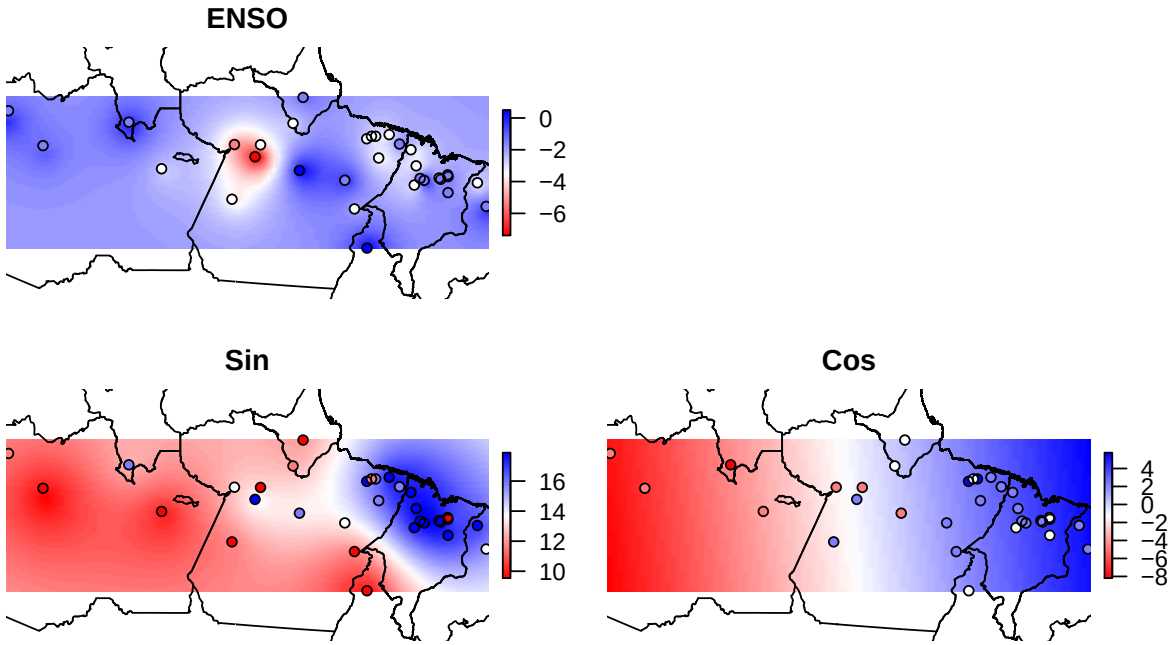

Figure 26 – Map of the north region and ordinary kriging for the coefficient estimates corresponding to each covariate in the location parameter of the generalized extreme-value distribution.

Table 11 – Total and fitting number of stations per region.

Region	Abbreviation	Total	Fitting stations
North*	North	28	19
Bahia, Sergipe	BA, SE	32	21
Rio Grande do Sul	RS	71	47
Paraná	PR	165	110
Rio de Janeiro	RJ	52	35
Espírito Santo	ES	52	35
Minas Gerais	MG	103	69
São Paulo	SP	322	107

* The "north" region includes the states of Tocantins, Amapá, Pará, Piauí, and Maranhão.

4.3 Return level maps

Due to the intense computational burden of the Bayesian hierarchal model, we fit it to separate regions of Brazil, as indicated in Table 11. For each region, we used two thirds of the stations to fit the model, and left the remaining to validation, except for the state of São Paulo. The mean of the prior distribution for ξ was set to 0.06 (the asymptotic value found in Section 4.1), and the standard deviation to 0.02. The priors used were the same for all regions, and are described in Section 3.14. We set the mean of the prior densities similar to the maximum likelihood estimates of each parameter, and the variance to some arbitrary large value. When estimating the parameters via maximum likelihood, we verified the adequacy of the exponential covariance function, and we noted that spatial correlation for the shape parameter is almost non-existent.

Our climate space consisted of longitude, latitude, and mean annual precipitation, but we also fitted the latent model without this last covariate, and compared this nested model using the deviance information criterion; see Table 12. For the states of Paraná, Rio de Janeiro, Minas Gerais, and São Paulo, the mean precipitation seems to be significant, so we used it to make the return level maps (using ordinary kriging like in the previous section). The maps can be seen in Figures 27 and 28, and a summary of the posterior for the covariance parameters is given in Table 13. These results were obtained after 20,000 iterations of the Markov chain, thinned by a factor of 30, preceded by a burn-in of 5,000 iterations.

Table 12 – Deviance information criterion (according to region).

Region	Model 1	Model 2
North	6,884	6,884
BA, SE	8,058	8,112
RS	17,691	17,688
PR	38,717	38,666
RJ	13,012	13,001
ES	12,704	12,699
MG	24,839	24,820
SP	39,224	39,107

Table 13 – Posterior means and associated 95% credible intervals for the covariance parameters of the latent processes (according to region) in each generalized extreme-value parameter.

	Region	Sill	Range	Region	Sill	Range
η	North	67 (27, 148)	76 (36, 130)	ES	43 (18, 93)	76 (37, 128)
	BA, SE	109 (48, 226)	78 (40, 127)	RJ	372 (172, 738)	69 (35, 117)
	RS	22 (11, 41)	85 (44, 139)	MG	26 (14, 46)	77 (41, 128)
	PR	38 (21, 64)	84 (47, 135)	SP	173 (100, 283)	86 (50, 136)
τ	North	16 (5, 38)	72 (35, 122)	ES	8 (3, 18)	79 (40, 132)
	BA, SE	7 (2, 16)	76 (37, 129)	RJ	39 (15, 84)	72 (36, 121)
	RS	5 (2, 11)	78 (37, 131)	MG	6 (3, 12)	73 (35, 126)
	PR	3 (2, 6)	93 (49, 149)	SP	38 (22, 62)	91 (54, 141)
ξ	North	0.011 (0.003, 0.03)	15 (5, 31)	ES	0.011 (0.004, 0.02)	19 (6, 38)
	BA, SE	0.011 (0.003, 0.03)	15 (5, 31)	RJ	0.012 (0.004, 0.03)	16 (6, 33)
	RS	0.006 (0.002, 0.01)	15 (5, 31)	MG	0.009 (0.004, 0.02)	15 (5, 31)
	PR	0.004 (0.002, 0.01)	14 (5, 29)	SP	0.008 (0.003, 0.01)	15 (5, 30)

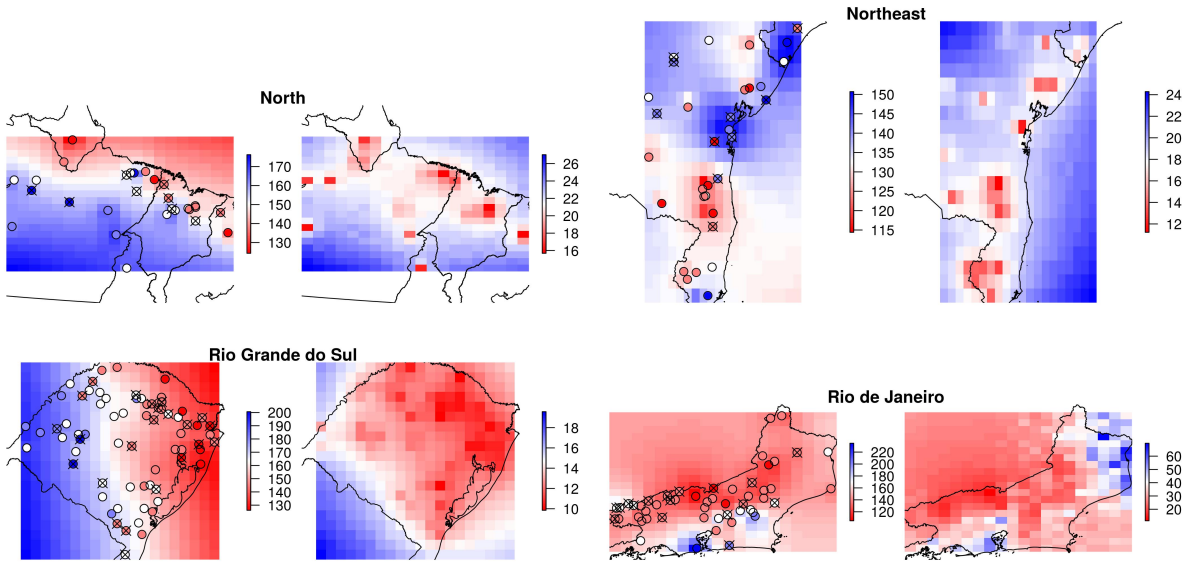

Figure 27 – Maps with predictive pointwise posterior mean estimates for the 25-year return level. The points represent the marginal maximum likelihood estimates. Points with a cross correspond to validation stations.

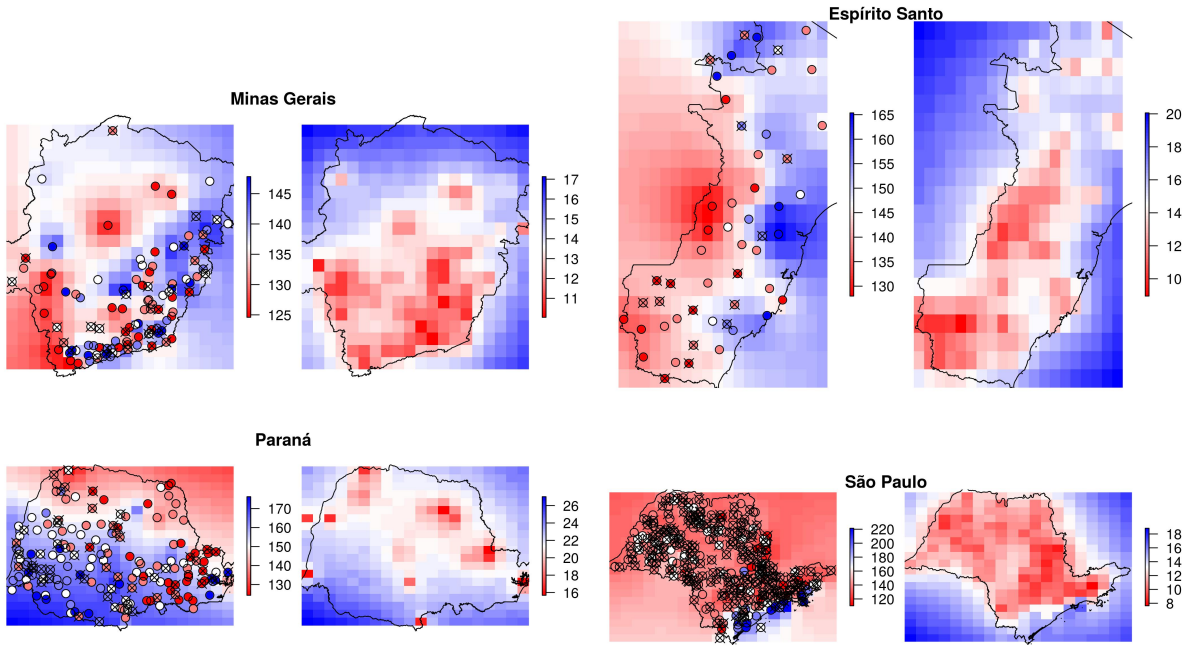

Figure 28 – Continuation of Figure 27.

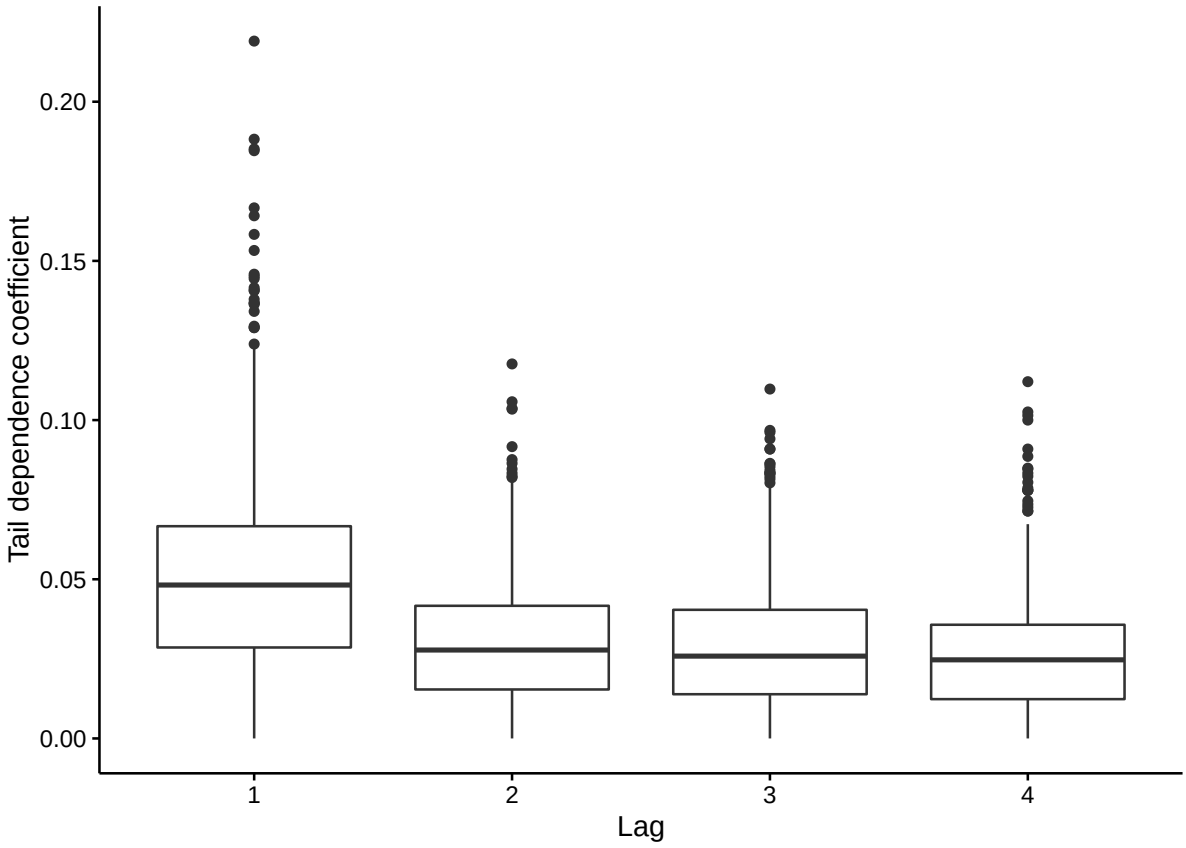

Figure 29 – Boxplot of the extremogram for the daily rainfall time series. For each series, the 98% quantile for wet days was used as the threshold.

4.4 Dependence measures

We assume that the quality in the body of the data is reasonable, and we apply model (3.25) to positive rainfall in order to obtain stationary residuals, and then proceed to calculate the various dependence measures.

Figure 29 shows the tail dependence coefficient until lag 4. The dependence at lag 1 is small, the largest value being 0.22, so there is only a slight tendency for rainfall extreme values to cluster. However, the upper 97.5% confidence limit for independent data, obtained by random permutation of the data, is exceeded around 39%, 14%, 11%, and 8% of the stations for lags 1, 2, 3, and 4, respectively.

Figure 30 displays the sample cross-extremogram for temporal lags $h = 0, 1$, and for pairs of stations classified in terms of the distance from the stations in each pair. When calculating the cross-extremograms between two stations, only the days for which there are observations on both stations are used, so the sample sizes can be greatly reduced. We kept only the pairs of stations that had more than 10 years of recording period in common. The cross-extremograms depicts a strong spatial pattern and significant dependence at lag 0 (i.e., extreme events happening in the same day) for stations distancing as far as 100 km. The empirical estimates for the pairwise extremal coefficient also shows this weak, but long range

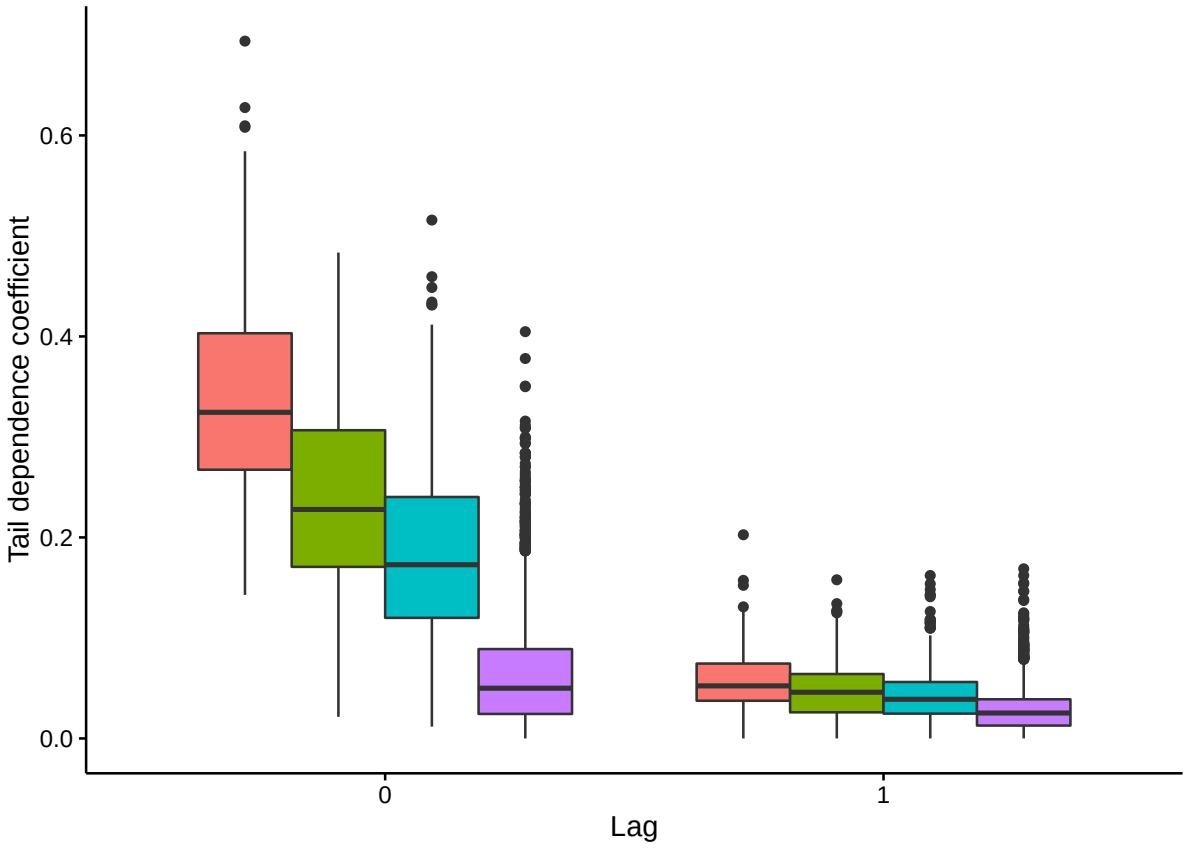

Figure 30 – Boxplot of the sample cross-extremogram for temporal lags $h = 0, 1$, and all pairwise combinations of stations distancing less than 25 km (red), between 25 km and 50 km (green), between 50 km and 100 km (blue), and greater than 100 km (purple). For each series, the 98% quantile for wet days was used as the threshold.

dependence; see Figure 31.

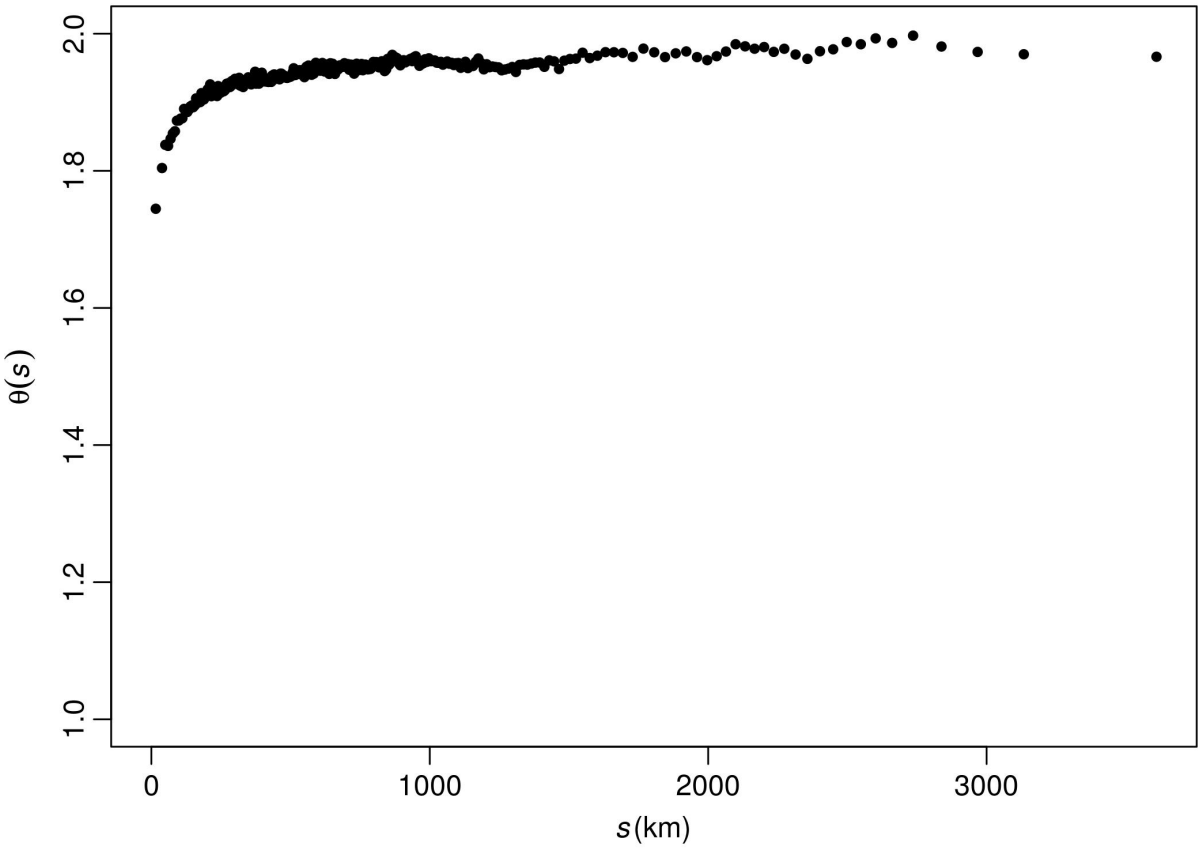

Figure 31 – Empirical estimates of the binned extremal coefficients (obtained from 300 bins).

Chapter 5

Final considerations

The ideal would be to have a large number of stations with long and stationary time series that were truly representative of the climatic regime of their sites, and that had good quality measurements. We did our best to select reliable rain stations in Brazil, and we noticed several obstacles in doing so, mainly the lack of important metadata like the stations' history and photographs showing their location and measurement conditions.

However, just the gathering of rainfall series made by the Agência Nacional de Águas (ANA) is already very valuable. This dataset, containing 11,619 stations, was never analyzed before, at least not with the scope we adopted. Despite the large number of stations, only 104 with record length greater than 80 years satisfied our selection criteria. Since our criteria required a detailed visual inspection of each rainfall series, we only retained the stations which had a record length greater than 80 years, and which had less than 10% of missing values in the period from 1972 to 2011. This second subset was necessary because the covariates we were interested in including in the analysis only had observations in limited time intervals, and the first subset did not have enough stations to produce return levels maps.

This study continues the work of Papalexiou e Koutsoyiannis [2013] and Serinaldi e Kilsby [2014], and complements their findings using another large dataset. Besides the return levels maps in Brazil, we make a deeper discussion about quality control, biases, and the application of extreme value theory to rainfall processes. We also investigate the effects of the ENSO phenomenon, CO_2 levels, and deforestation on extreme rainfall, as well as spatial and temporal dependence in extreme levels. Future developments in producing return level maps should allow temporal covariates in the mean of the latent variable model.

Bibliography

ARRAUT, J. M. et al. Aerial rivers and lakes: looking at large-scale moisture transport and its relation to Amazonia and to subtropical rainfall in South America. *Journal of Climate*, v. 25, n. 2, p. 543–556, 2012.

COLES, S. G. *An Introduction to Statistical Modeling of Extreme Values*. New York: Springer, 2001.

COLES, S. G.; PERICCHI, L. Anticipating catastrophes through extreme value modelling. *Applied Statistics*, Wiley Online Library, v. 52, n. 4, p. 405–416, 2003.

COOLEY, D.; NAVEAU, P.; PONCET, P. Variograms for spatial max-stable random fields. In: BERTAIL, P.; DOUKHAN, P.; SOULIER, P. (Ed.). *Dependence in Probability and Statistics*. New York: Springer, 2006, (Lecture Notes in Statistics). p. 373–390.

COX, D. R.; SNELL, E. J. A general definition of residuals (with Discussion). *Journal of the Royal Statistical Society. Series B*, JSTOR, v. 30, p. 248–275, 1968.

DAVIS, R. A.; MIKOSCH, T. The extremogram: A correlogram for extreme events. *Bernoulli*, Bernoulli Society for Mathematical Statistics and Probability, v. 15, n. 4, p. 977–1009, 2009.

DAVIS, R. A.; MIKOSCH, T.; CRIBBEN, I. Towards estimating extremal serial dependence via the bootstrapped extremogram. *Journal of Econometrics*, Elsevier, v. 170, n. 1, p. 142–152, 2012.

DAVISON, A. C.; HUSER, R. Statistics of extremes. *Annual Review of Statistics and its Application*, Annual Reviews, v. 2, p. 203–235, 2015.

DAVISON, A. C.; PADOAN, S. A.; RIBATET, M. Statistical modeling of spatial extremes. *Statistical Science*, Institute of Mathematical Statistics, v. 27, n. 2, p. 161–186, 2012.

DAVISON, A. C.; SMITH, R. L. Models for exceedances over high thresholds (with Discussion). *Journal of the Royal Statistical Society. Series B*, JSTOR, v. 52, p. 393–442, 1990.

DOURADO, F.; ARRAES, T. C.; SILVA, M. F. The "mega-disaster" in the mountainous region of the state of Rio de Janeiro: causes, mechanisms of mass movements and spatial allocation of investments for reconstruction post-disaster. *Anuário do Instituto de Geociências*, scieloppegeo, v. 35, p. 43–54, 2012.

EASTOE, E. F.; TAWN, J. A. Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society. Series C*, Wiley Online Library, v. 58, n. 1, p. 25–45, 2009.

ESCOBAR, H. Drought triggers alarms in Brazil's biggest metropolis. *Science*, American Association for the Advancement of Science, v. 347, n. 6224, p. 812–812, 2015.

FAWCETT, L.; WALSHAW, D. Estimating return levels from serially dependent extremes. *Environmetrics*, Wiley Online Library, v. 23, n. 3, p. 272–283, 2012.

FERRO, C. A. T.; SEGERS, J. Inference for clusters of extreme values. *Journal of the Royal Statistical Society. Series B*, Wiley Online Library, v. 65, n. 2, p. 545–556, 2003.

FISHER, R. A.; TIPPETT, L. H. C. Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, v. 24, n. 2, p. 180–190, 1928.

FREITAS, M. A. S.; NóBREGA, M. T. *Orientações para consistência de dados pluviométricos*. [S.l.], 2012.

FRISCH, U.; SORNETTE, D. Extreme deviations and applications. *Journal de Physique I*, EDP Sciences, v. 7, n. 9, p. 1155–1171, 1997.

GETIRANA, A. C. V. Extreme water deficit in Brazil detected from space. *Journal of Hydrometeorology*, v. 17, p. 591–599, 2016.

GILES, D. E. et al. *Bias-corrected maximum likelihood estimation of the parameters of the generalized Pareto distribution*. [S.l.], 2011.

GRIMM, A. M. The El Niño impact on the summer monsoon in Brazil: regional processes versus remote influences. *Journal of Climate*, v. 16, n. 2, p. 263–280, 2003.

GRIMM, A. M.; BARROS, V. R.; DOYLE, M. E. Climate variability in southern South America associated with El Niño and La Niña events. *Journal of Climate*, v. 13, n. 1, p. 35–58, 2000.

GRIMSHAW, S. D. Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics*, Taylor & Francis, v. 35, n. 2, p. 185–191, 1993.

HASTENRATH, S.; GREISCHAR, L. Further work on the prediction of northeast Brazil rainfall anomalies. *Journal of Climate*, v. 6, n. 4, p. 743–758, 1993.

HOSKING, J. R. M. Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika*, Biometrika Trust, v. 71, n. 2, p. 367–374, 1984.

HOSKING, J. R. M. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B*, Wiley, v. 52, n. 1, p. 105–124, 1990.

HOSKING, J. R. M.; WALLIS, J. R. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, Taylor & Francis, v. 29, n. 3, p. 339–349, 1987.

HOSKING, J. R. M.; WALLIS, J. R.; WOOD, E. F. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, Taylor & Francis Group, v. 27, n. 3, p. 251–261, 1985.

JARRAUD, M. *Guide to Meteorological Instruments and Methods of Observation*. [S.l.]: World Meteorological Organisation, 2008.

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

KOUTSOYIANNIS, D. Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, Taylor & Francis, v. 49, n. 4, p. 575–590, 2004.

KOUTSOYIANNIS, D. Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, Taylor & Francis, v. 49, n. 4, p. 591–610, 2004.

LIBERTO, T. D. Flooding in Chile's Atacama Desert after years' worth of rain in one day. *National Oceanic and Atmospheric Administration Climate.gov*, 2015.

MARENGO, J. A. Interannual variability of surface climate in the Amazon basin. *International Journal of Climatology*, Wiley Online Library, v. 12, n. 8, p. 853–863, 1992.

MARENGO, J. A. et al. Two contrasting severe seasonal extremes in tropical South America in 2012: flood in Amazonia and drought in northeast Brazil. *Journal of Climate*, v. 26, n. 22, p. 9137–9154, 2013.

MULLER, A. et al. Uncertainties of extreme rainfall quantiles estimated by a stochastic rainfall model and by a generalized Pareto distribution. *Hydrological Sciences Journal*, Taylor & Francis, v. 54, n. 3, p. 417–429, 2009.

NAZARENO, A. G.; LAURANCE, W. F. Brazil's drought: beware deforestation. *Science*, v. 347, n. 6229, p. 1427, 2015.

PAPALEXIOU, S. M.; KOUTSOYIANNIS, D. Battle of extreme value distributions: a global survey on extreme daily rainfall. *Water Resources Research*, Wiley Online Library, v. 49, n. 1, p. 187–201, 2013.

PARMESAN, C.; ROOT, T. L.; WILLIG, M. R. Impacts of extreme weather and climate on terrestrial biota. *Bulletin of the American Meteorological Society*, v. 81, n. 3, p. 443–450, 2000.

PICKANDS, J. Statistical inference using extreme order statistics. *The Annals of Statistics*, JSTOR, v. 3, n. 1, p. 119–131, 1975.

PLUMMER, N. et al. *Guidelines on Climate Observation: Networks and Systems*. [S.l.]: World Meteorological Organization, 2003.

PREVIDELLI, I. T. S.; DAVISON, A. C. *Improved likelihood estimation for rare event parameters*. 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <http://www.R-project.org/>.

RAO, V. B. et al. An update on the rainfall characteristics of Brazil: seasonal variations and trends in 1979–2011. *International Journal of Climatology*, Wiley Online Library, v. 36, n. 1, p. 291–302, 2016.

REBOITA, M. S. et al. Regimes de precipitação na América do Sul: uma revisão bibliográfica. *Revista Brasileira de Meteorologia*, SciELO Brasil, v. 25, n. 2, p. 185–204, 2010.

RIBATET, M.; DOMBRY, C.; OESTING, M. Spatial Extremes and Max-Stable Processes. In: *Extreme Value Modeling and Risk Analysis: Methods and Applications*. [S.l.]: Chapman and Hall/CRC, 2016. cap. 9.

RIGBY, R. A.; STASINOPOULOS, D. M. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, SAGE Publications, v. 6, n. 3, p. 209–229, 2006.

SERINALDI, F.; KILSBY, C. G. Rainfall extremes: toward reconciliation after the battle of distributions. *Water Resources Research*, Wiley Online Library, v. 50, n. 1, p. 336–352, 2014.

SMITH, R. L. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, Biometrika Trust, v. 72, n. 1, p. 67–92, 1985.

SMITH, R. L. *Approximations in extreme value theory*. [S.l.], 1987.

SÜVEGES, M.; DAVISON, A. C. A case study of a "Dragon-King": The 1999 Venezuelan catastrophe. *The European Physical Journal Special Topics*, Springer, v. 205, n. 1, p. 131–146, 2012.

SVOBODA, M.; HAYES, M.; WOOD, D. Standardized precipitation index user guide. *World Meteorological Organization Geneva, Switzerland*, 2012.

VINEY, N. R.; BATES, B. C. It never rains on Sunday: the prevalence and implications of untagged multi-day rainfall accumulations in the Australian high quality data set. *International Journal of Climatology*, Wiley Online Library, v. 24, n. 9, p. 1171–1192, 2004.

WADSWORTH, J. L. Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, Taylor & Francis, v. 58, n. 1, p. 116–126, 2016.

WANG, X. L. et al. New techniques for the detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*, v. 49, n. 12, p. 2416–2436, 2010.

WILSON, P. S.; TOUMI, R. A fundamental probability distribution for heavy rainfall. *Geophysical Research Letters*, Wiley Online Library, v. 32, n. 14, 2005.

WOLTER, K. *Multivariate ENSO Index (MEI)*. 2000. Climate Diagnostics Center Report at <www.cdc.noaa.gov>.